Discern Depth Under Foul Weather: Estimate 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 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 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 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 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 PM$_{2.5}$ value through the atmospheric scattering model. Experimental results demonstrate that our method achieves accurate 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, PM$_{2.5}$ estimation, depth estimation, SVR.

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. PM$_{2.5}$) need be quickly and accurately estimated in daily life. Existing methods measure PM$_{2.5}$ values with special devices, e.g. Hanvon M1. 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 PM$_{2.5}$ values, the detection stations are limited and the provided PM$_{2.5}$ value may not accurate for the location of user. Therefore, it is very important and urgently needed to accurately measure 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 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 proposed PMnet, together with dark channel maps. A depth estimation method is also proposed by learning the relationship between PM$_{2.5}$ and atmospheric attenuation coefficient $\beta$. We also collect a large dataset containing 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 PM$_{2.5}$ value is 180.198 $\mu g/m^3$.

The main contributions are summarized as follows:

- An image-based 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 PM$_{2.5}$ value.
- A hybrid CNN for learning distance-aware haze-relevant

![Image](image_url)

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.
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 $\beta$ 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$_{2.5}$ values. This method can work in both clean and foul weather, and pollution-related factors. Despite accuracy, these 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 an 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.

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].
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 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 PM$_{2.5}$ values, we capture 1575 images of different scenes using an Apple 5s mobile phone, and simultaneously measure the corresponding PM$_{2.5}$ values of the current scenes with Hanvon M1 which measures 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 PM$_{2.5}$ values are between 0-300 µg/m$^3$, which are categorized into three classes: Good (PM$_{2.5}$ < 75), Moderate (PM$_{2.5}$ ∈ [75, 150]), and Severe (PM$_{2.5}$ ≥ 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$_3$), 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 PM$_{2.5}$ values for learning.

IV. PM$_{2.5}$ Estimation

In order to obtain accurate 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 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, 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.

We extract the haze-relevant features by:

$$F_{Haze} = \omega_1 F_{Sky} \oplus \omega_2 F_{Building} \oplus F_{Class},$$  \hspace{1cm} (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 PM$_{2.5}$ estimation. It is well-known that the 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 $3 \times 3$ 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:

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

\hspace{1cm} (2)

²http://www.moji.com
where \( z_j \) is the feature of the \( j^{th} \) image and \( \gamma \) 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 \times 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 PM\(_{2.5}\) values and the features extracted from our PMnets by SVR:

\[
\begin{aligned}
\min_{\omega,b} \frac{1}{2} ||\omega||^2 + c \sum_{i=1}^{m} \ell(c) (f(x_i) - y_i)
\end{aligned}
\]

\[
\begin{aligned}
f(x_i) = \omega^T x_i + b
\end{aligned}
\]

\[
\ell(c) = \begin{cases} 
0, & \text{if } |c| \leq \varepsilon \\
|c| - \varepsilon, & \text{otherwise,}
\end{cases}
\]

where \( \omega \) 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 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]:

\[
I(x) = J(x)t(x) + A[1 - t(x)],
\]

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 \( d(x) \):

\[
t(x) = e^{-\beta d(x)},
\]

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

\[
d(x) = -\frac{1}{\beta} \ln t(x),
\]

Therefore, we need estimate transmission map \( t(x) \) and \( \beta \) to compute the depth map of scene.

**A. Transmission Estimation**

Some methods [4], [24], [25] recovered the clean image \( I(x) \) from \( I(x) \) by estimating the transmission \( t(x) \). 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 \( \hat{t}(x) \) [4]:

\[
\hat{t}(x) = 1 - w \min_{c} \left( \min_{y \in I(x)} \frac{I^c(y)}{A^c} \right),
\]

where \( w \) is an environmental factor set to be 0.95, \( \Omega(x) \) is a local patch centered at \( x \), \( A^c \) 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).

\[
\mathcal{E}_x = \sum_x \frac{(t(x) - \hat{t}(x))^2}{\sigma^2(x)} + \lambda \sum_x \sum_{y \in N(x)} \sqrt{\alpha_{x,y}} ||t(x) - t(y)||_1,
\]

where \( \sigma(x) \) is the standard deviation of \( \hat{t}(x) \), \( N(x) \) represents the neighborhood set of pixel \( x \), \( || \cdot ||_1 \) represents the \( \ell_1 \) norm, \( \lambda \) is a penalization parameter, and \( \alpha_{x,y} \) is a pairwise weight calculated by

\[
\alpha_{x,y} = \exp \left( \frac{||B_x \circ (P_x - P_y)||_2^2}{\sigma_1^2} \right),
\]

where \( P_x, P_y \) is an operator extracting a \( w \times w \) patch centered at \( x \) (y) on the hazy image \( I \), \( \circ \) represents the element-wise multiplication, \( \theta_1 \) determines the decay rate of exponential function, and \( B_x \) is a bilateral filter kernel defined as

\[
B_x(x,y) = \exp \left( -\frac{||x-y||_2^2}{\sigma_2^2} \right) \exp \left( -\frac{\sum_{i \in c} (I_{x,i} - I_{y,i})^2}{\sigma_3^2} \right),
\]

where \( \sigma_2 \) and \( \sigma_3 \) are constant parameters adjusting the spatial range and the intensity range, respectively. As shown in Fig.
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 x and every pixel y in the neighborhood of x are evaluated. We use non-local prior weighted by a bilateral kernel on a larger neighborhood to fully exploit structural correlation, and adopt $\ell_1$ 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):

$$T = [t(1,1), t(1,2), \cdots, t(w, h)],$$

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

$$W = \text{diag}\left(\frac{1}{\sigma(x_1)}, \frac{1}{\sigma(x_2)}, \cdots, \frac{1}{\sigma(x_n)}\right),$$

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

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

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

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

where $\lambda$ is penalty coefficient, $\left\|\cdot\right\|_2$ denotes the $l_2$ norm, $\mu$ 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 $\mu$ 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:

$$A^{(k+1)} = \text{arg}_{A} \min \lambda \left\|A\right\|_1 + \langle Y^{(k)}, A - QT^{(k)} \rangle + \frac{\mu^{(k)}}{2} \left\|A - QT^{(k)}\right\|_2^2,$$

$$T^{(k+1)} = \text{arg}_{T} \min \left\|W(T - \tilde{T})\right\|_2^2 + \langle Y^{(k)}, A^{(k+1)} - QT \rangle + \frac{\mu^{(k)}}{2} \left\|A^{(k+1)} - QT\right\|_2^2,$$

$$Y^{(k)} = Y^{(k)} + A^{(k+1)} - QT^{(k+1)},$$

$$\mu^{(k+1)} = \rho \mu^{(k)}, \rho > 0.$$  

(17)

**B. Atmospheric Attenuation Coefficient $\beta$ 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 PM$_{2.5}$ values will have different atmospheric attenuation coefficient $\beta$.

We learn the relationship between PM$_{2.5}$ and $\beta$ 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 synthetic hazy images from the same set of PM$_{2.5}$ values which corresponds to the average values of all the PM$_{2.5}$ values which corresponds to the same $\beta$ and fit the relationship between $\beta$ and PM$_{2.5}$, and fit the relationship between $\beta$ and PM$_{2.5}$, and find the relationship between PM$_{2.5}$ and $\beta$ through a lot of statistical experiments as shown in Fig. 5. Considering the existence of errors, we take the average values of all the PM$_{2.5}$ values which corresponds to the same $\beta$ and fit the relationship between $\beta$ and PM$_{2.5}$. Fitted by least squares method, the relationship between PM$_{2.5}$ and $\beta$ is

$$\beta = ax^b,$$

(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 PM$_{2.5}$ value by the proposed deep learning method in Sec. IV, and then compute the $\beta$ 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).

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{pm25.png}
\caption{PM$_{2.5}$ estimation result using different values of $\omega_1$ (a) and $\omega_2$ (b).}
\end{figure}

VI. EXPERIMENTAL RESULTS

In this section, we first evaluate the proposed 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 PM$_{2.5}$ values. The penalty coefficient $c$ is learnt by cross entropy. In addition, the parameters for transmission optimization are set as follows: $\lambda = 0.1$, $\vartheta_1 = 3.05$, $\vartheta_2 = 1000$, and $\vartheta_3 = 0.2$. 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 PM$_{2.5}$ estimation.

We also evaluate the influence of weights $\omega_1$ and $\omega_2$ 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 PM$_{2.5}$ values.

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): $\frac{1}{T} \sum_p \frac{|t^{\text{est}}_p - t^{\text{gt}}_p|}{t^{\text{gt}}_p}$;
- Root mean squared error (RMSE): $\sqrt{\frac{1}{T} \sum_p (t^{\text{gt}}_p - t^{\text{est}}_p)^2}$;
- log$_{10}$ error (log10): $\frac{1}{T} \sum_p |\log_{10} t^{\text{gt}}_p - \log_{10} t^{\text{est}}_p|$;
- PSNR: $10 \times \log_{10} \frac{255^2}{\text{RMSE}^2}$;
- SSIM: \[ \frac{(2\mu_{p} \sigma_{p} + \alpha_1 \sigma_{p}^2 + \alpha_2 \mu_p^2 + \alpha_2 \sigma^2_p)}{(\mu^2_p + \sigma^2_p + \alpha_1 \sigma^2_p + \alpha_2 \sigma^2_p)(\sigma^2_p + \sigma^2_p + \alpha_2 \sigma^2_p)} \]

where $t^{\text{gt}}_p$ is the ground-truth transmission at pixel $p$, $t^{\text{est}}_p$ is the corresponding estimated transmission, and $T$ is the number of image pixels. For SSIM, $\mu_p$ and $\mu_{\text{est}}$ represent the mean values of the ground-truth transmission and the estimated transmission, $\sigma_p$ and $\sigma_{\text{est}}$ represent the standard deviations of the two images, and $\sigma_{p,\text{est}}$ represents the covariance of the ground-truth transmission and the estimated transmission.

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].

<table>
<thead>
<tr>
<th>Feature Combinations</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sky only</td>
<td>45.93</td>
</tr>
<tr>
<td>Building only</td>
<td>50.89</td>
</tr>
<tr>
<td>Sky + Building</td>
<td>46.10</td>
</tr>
<tr>
<td>Sky concat Building</td>
<td>45.26</td>
</tr>
<tr>
<td>Sky concat Building</td>
<td>24.35</td>
</tr>
<tr>
<td>our method</td>
<td>21.95</td>
</tr>
</tbody>
</table>
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.

\[ \text{PM}_{2.5} \]

\[ \text{log10} \]

\[ \text{PSNR} \]

\[ \text{RMSE} \]

\[ \text{SSIM} \]

Fig. 6. PM\textsubscript{2.5} estimation results (PM\textsubscript{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 α is predicted value.

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): \( \frac{1}{T} \sum_p \frac{|d_{\text{gt}} - d_{\text{est}}|}{d_{\text{gt}}} \);
- Root mean squared error (RMSE):
  \[ \sqrt{\frac{1}{T} \sum_p (d_{\text{gt}}^2 - d_{\text{est}}^2)^2} ; \]
- \( \log_{10} \) error (\( \log_{10} \)):
  \[ \frac{1}{T} \sum_p \left| \log_{10} d_{\text{gt}} - \log_{10} d_{\text{est}} \right| ; \]

where \( d_{\text{gt}} \) is the ground-truth depth at pixel \( p \), \( d_{\text{est}} \) 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\textsubscript{2.5} value estimated by our method is 189.128 µg/m\textsuperscript{3}. Our method achieved the best results, which suggests that combining PM\textsubscript{2.5} to estimate depth is very effective.

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\textsubscript{2.5} value estimated by our method is 43.573 µg/m\textsuperscript{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 in 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 traditional depth estimation methods [2] cannot preserve the outline information of objects.

\[ T \]

\[ Q \]

\[ L \]

\[ U \]

\[ A \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]

\[ I \]

\[ T \]
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.

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].

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.

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.
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 compiled 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>
<thead>
<tr>
<th>Procedure</th>
<th>Dataset</th>
<th>Make3D (343 × 458)</th>
<th>NYU (561 × 427)</th>
<th>Internet (1019 × 624)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM_{2.5} Est - proposed net</td>
<td>0.129s</td>
<td>0.105s</td>
<td>0.225s</td>
<td></td>
</tr>
<tr>
<td>PM_{2.5} Est - SVR</td>
<td>0.705s</td>
<td>0.708s</td>
<td>0.702s</td>
<td></td>
</tr>
<tr>
<td>Transmission Estimation</td>
<td>7.404s</td>
<td>10.678s</td>
<td>28.361s</td>
<td></td>
</tr>
<tr>
<td>Depth Estimation</td>
<td>0.004s</td>
<td>0.004s</td>
<td>0.009s</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8.242s</td>
<td>11.495s</td>
<td>29.297s</td>
<td></td>
</tr>
</tbody>
</table>

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


