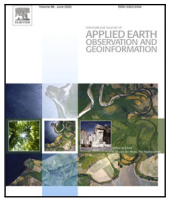
International Journal of Applied Earth Observation and Geoinformation 128 (2024) 103753

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Self-supervised multi-task learning framework for safety and health-oriented road environment surveillance based on connected vehicle visual perception

Shaocheng Jia [a,c], Wei Yao [a,b,*]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[b] The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
[c] Department of Civil Engineering, The University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Bidirectional process of image synthesis and decomposition (BPISD)
Self-supervised learning
Depth estimation
Visibility estimation
Airlight estimation
$PM_{2.5}$ mass concentration estimation

ABSTRACT

Cutting-edge connected vehicle (CV) technologies have drawn much attention in recent years. The real-time traffic data captured by a CV can be shared with other CVs and data centers so as to open new possibilities for solving diverse transportation problems. The trajectory data of CVs have been well-studied and widely used. However, image data captured by onboard cameras in a connected environment, as being a kind of fundamental data source, are not sufficiently investigated, especially for safety and health-oriented visual perception. In this paper, a bidirectional process of image synthesis and decomposition (BPISD) approach is proposed, and thus a novel self-supervised multi-task learning framework, to simultaneously estimate depth map, atmospheric visibility, airlight, and $PM_{2.5}$ mass concentration, in which depth map and visibility are considered highly associated with traffic safety, while airlight and $PM_{2.5}$ mass concentration are directly correlated with human health. Both the training and testing phases of the proposed system solely require a single image as input. Due to the innovative training pipeline, the depth estimation network can automatically manage various levels of visibility conditions and overcome diverse inherent problems in current image-synthesis-based self-supervised depth estimation, thereby generating high-quality depth maps even in low-visibility situations and further benefiting accurate estimations of visibility, airlight, and $PM_{2.5}$ mass concentration. Extensive experiments on the original and synthesized data from the KITTI dataset and real-world data collected in Beijing demonstrate that the proposed method can (1) achieve performance comparable in self-supervised depth estimation as compared with other state-of-the-art methods when taking clear images as input; (2) predict vivid depth map for images contaminated by various levels of haze when the network trained with previous framework fails; and (3) accurately estimate visibility, airlight, and $PM_{2.5}$ mass concentrations. Beneficial applications can be developed based on the presented work to contribute to high-precise and dynamic geoinformation reconstruction, transportation, meteorology, and smart city.

1. Introduction

In the past decade, communication technologies have undergone substantial development. In transportation systems, connected vehicle (CV) technology enables information exchanges between different system components. The running connected vehicles can continuously collect data of interest and send them back to the data center for further analysis. Taking CVs as mobile sensors opens new possibilities to solve transportation problems using the shared trajectory data (Jia et al., 2023). It provides us insights to conduct driving environment perception tasks in a connected environment. Specifically, the image data captured by the onboard cameras installed on the CVs can also be shared with other CVs and stored in the data centers. As the road network generally matches the whole city, CVs, therefore, are distributed in large-scale urban areas due to regular traffic demands. Together with the fact that CV flows are continuous in time, the quantity estimations can then be spatiotemporally continuous. This may promote paradigm shifts in driving environment perception.

The following four types of information are considered important for driving environment perception. First, depth estimation is crucial for 3D scene understanding and the driver's decision-making. While current depth estimation methods require either truthful labels for supervision or image sequence for self-supervised training. The former is expensive. The latter may fail when the static scene assumption is violated in a driving environment. Second, visibility is highly associated with traffic safety. Moreover, $PM_{2.5}$ mass concentration and airlight are related to air quality, and thus human health. However,

* Corresponding author at: Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong.
E-mail address: wei.hn.yao@polyu.edu.hk (W. Yao).

https://doi.org/10.1016/j.jag.2024.103753
Received 2 October 2023; Received in revised form 31 January 2024; Accepted 4 March 2024
Available online 6 March 2024

visibility, $PM_{2.5}$ mass concentration, and airlight are often measured with a professional instrument outfitted on the roads. Due to the high capital costs, it is impossible to obtain dense estimates in a large-scale city. Moreover, substantial differences in these quantities can exist across various urban areas, raising from diverse land uses, special geographical properties, and complex weather conditions. Therefore, spatiotemporally continuous estimates are expected for achieving more accurate driving environment perception. The dynamic CV signals offer a unique channel for solving these problems.

Taking the image data collected by CVs as input, a self-supervised co-training framework is proposed to simultaneously estimate depth map, airlight, clear image, and visibility using four convolutional neural networks (CNNs). The forward inference process is an image decomposition phase. The estimated results are coupled with Koschmieder's law to reconstruct the input image. This reconstruction process is an image synthesis phase. The difference between the input image and the reconstructed image is used to train the system. It is noted that after co-training each sub-network can be used separately. Upon the estimated visibility, a statistical model is further deployed to map visibility with $PM_{2.5}$ mass concentration, where low visibility is assumed to be solely caused by $PM_{2.5}$. The estimated $PM_{2.5}$ mass concentrations are projected back to the physical world in terms of the location information for precise air quality monitoring. As traffic is highly dynamic, the estimated local $PM_{2.5}$ mass concentration can be continuously updated. The expectation of multiple estimates for the same area can be taken as the final outcome. Visibility, depth map, and airlight, can also be projected onto the physical map and used in driving environment reconstruction or air component analysis.

Such a bidirectional process of image synthesis and decomposition (BPISD) is radically different from the previous training pipeline wherein only image synthesis is used. Furthermore, the reference images used for reconstruction in the previous training pipeline have certain time and space shifts as compared with the target image, i.e., the image sequence is used. This introduces issues of occlusion, moving objects, lighting changes, etc. The presented work solely requires a single image as input. The corresponding clear image (i.e., dehazed image in this study) will be the reference image for reconstruction. All the above-mentioned issues are automatically resolved.

To validate the proposed system, a large number of driving-view images with various visibilities caused by different degrees of $PM_{2.5}$ mass concentrations are needed for training the deep neural networks. Such a dataset, however, is unavailable to the best of our knowledge. To address this issue, a novel synthetic method based on Koschmieder's law has been proposed and applied to Zhou et al.' split (Zhou et al., 2017) in self-supervised depth estimation on the KITTI dataset (Geiger et al., 2013). Eigen's split (Zhou et al., 2017) of the KITTI dataset was used for evaluating the performance of depth estimation. The trained visibility estimation model was directly applied to the real-world data collected in Beijing without any refinement. Adopting simple polynomial fitting, the estimated visibilities can be readily transformed to $PM_{2.5}$ mass concentrations. Comparisons show that the proposed method achieves competitive and robust performance in self-supervised depth estimation under various visibility conditions. For estimation of visibility and $PM_{2.5}$ mass concentrations, it was found that the mean absolute percentage errors (MAPE) were respectively confined within 5% and 8% across various levels of relative humidity with the order of polynomial fitting beyond 6. These promising performances clearly demonstrate the effectiveness of the proposed method. It is noted that the said approach does not require any additional devices or change the vehicle configurations, being a non-intrusive method. As such, it has great potential to be implemented for monitoring real-time particle conditions and promoting paradigm shifts on many applications, e.g., starting and embracing health-aware navigation and travel planning.

In short, the contributions of this paper are fourfold:

- A CV-based framework is proposed for estimating spatiotemporally continuous visibility, airlight, and PM2.5 mass concentration in large-scale cities and possibly establishing dynamical 3D driving maps by means of accurate depth maps irrespective of diverse visibility conditions.
- A novel bidirectional process of image synthesis and decomposition (BPISD) paradigm is proposed, and thus a unified self-supervised multi-task learning framework, to simultaneously estimate depth map, visibility, airlight, and $PM_{2.5}$ mass concentration by taking a single image as input.
- Applying the proposed methods to the KITTI dataset and the real-world Beijing data affords excellent performance in all four subtasks. In particular, the proposed single-image and self-supervised depth (SSD) method can achieve performance competitive compared to state-of-the-art methods and significantly outperform the method trained with the traditional self-supervised pipeline in the case of low-visibility situations.
- This study showcases how advanced vehicle technologies (e.g., CVs) help solve problems beyond transportation and opens new possibilities for developing health and safety-oriented applications.

In particular, the presented research is fundamentally different from DMRVisNet (You et al., 2022) in the following aspects: (1) this study focuses on simultaneously estimating depth map, airlight, visibility, and $PM_{2.5}$ mass concentration, while DMRVisNet only targets on visibility estimation; (2) this study proposes a self-supervised and multi-task learning framework leveraging Koschmieder's law, in contrast, DMRVisNet adopts supervised learning for all tasks; (3) this study presents a framework for developing safety and health-oriented applications in large-scale cities by combining connected vehicle technologies and the proposed self-supervised multi-task learning paradigm, however, DMRVisNet is unrelated.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 defines the problem. Section 4 presents the proposed method. Section 5 reports experimental results. Section 6 concludes the paper.

2. Related work

2.1. Estimation of depth map and visibility

Depth map offers important three-dimensional (3D) information for the given image, and thus in conjunction with LiDAR point clouds (Wang and Yao, 2022) is widely used in 3D reconstruction, scene understanding, and autonomous driving. However, it is nontrivial to accurately estimate the depth map from a single image, as monocular depth estimation is an inherently ill-posed problem. Prior to the prosperity of deep learning, hand-crafted features need to be extracted from the input raw images. Then, regression or classification process is used to estimate the depth map (Saxena et al., 2008; Baig and Torresani, 2016; Choi et al., 2015; Furukawa et al., 2017; Zoran et al., 2015). This type of method heavily relies on the feature design. It is challenging to devise abstract and comprehensive features manually.

Deep learning, especially convolutional neural networks (CNN), can automatically extract abstract and deep features from the input data (Jia et al., 2020b; Weng et al., 2021; Polewski et al., 2021; Li et al., 2023; Wang et al., 2023), and thus provide new channels to conduct monocular depth estimation. Given the ground truth depth map, the end-to-end networks are trained by minimizing the difference between the estimated and truthful depth maps. Tremendous work has been conducted focusing on exploring various CNN structures (Chen et al., 2016; Eigen et al., 2014; Eigen and Fergus, 2015; Laina et al., 2016; Li et al., 2017) and capturing global information of the images (Cao et al., 2017; Eigen and Fergus, 2015; Li et al., 2015; Liu et al., 2015a; Mousavian et al., 2016; Xu et al., 2017, 2018; Almalioglu et al., 2019;

C.S. Kumar et al., 2018; Grigorev et al., 2017; Mancini et al., 2017; Tananaev et al., 2018; Wang et al., 2019a; Jia et al., 2020a). However, such supervised depth map methods require a large number of truthful depth maps. This hinders the models' universal application.

Self-supervised depth estimation does not require any labeled data for training the networks. The previous training pipeline requires continuous image sequences as input. By defining target and reference images, the difference between the reconstructed and original target images is computed for training the whole system, in which the reconstruction of target images is based on reference images, the depth map of the target image, and the ego-motion between the target image and reference images. This pipeline has attracted many researchers (Zhou et al., 2017; Chen et al., 2019b; Garg et al., 2016; Ranjan et al., 2019; Yin and Shi, 2018; Zhan et al., 2018; Zhou et al., 2019, 2017; Godard et al., 2017; Kuznietsov et al., 2017; Almalioglu et al., 2019; Feng and Gu, 2019; Guizilini et al., 2020; Jia et al., 2021, 2022; Jia and Yao, 2023). However, the underlying assumption is that all objects in the scene are static. This assumption is often violated in a driving environment. Moreover, the common occurrence of occlusion and disocclusion in the course of vehicle moving also brings difficulty in finding pixel correspondence. Although some works are devoted to mitigating such issues (Godard et al., 2019; Casser et al., 2019; Klingner et al., 2020; Shu et al., 2020), the problems can hardly be completely resolved.

Different from the previous training pipeline used in depth estimation, this work proposes a novel self-supervised multi-task learning framework based on a bidirectional process of image synthesis and decomposition (BPISD), which solely requires a single image as input for training the whole system. Thus, the above-described issues are automatically resolved, thereby possibly achieving better performance. A single image input lets the system training be more flexible as well. Multi-task learning is considered more cost-effective in driving environment perception.

In visibility estimation, three types of methods are often adopted, i.e., traditional methods, statistical methods, and deep neural network (DNN)-based methods. Traditionally, visibility is measured by either manual observation or professional instruments. For the former, the estimate could be varied from observer to observer. For the latter, professional instrument is generally expensive and is impossible to be densely installed in large-scale cities (Chaabani et al., 2017; Pomerleau, 1997). Statistical methods estimates visibility by definition or modeling the relationship between the collected data and visibility (Dietz et al., 2019; Cheng et al., 2018), which generally need to perform geographic calibration and thus are difficult for universal application. DNN-based methods can establish an end-to-end model for conventionally estimating visibility. While a large number of labeled data are required for training the networks (Palvanov and Im Cho, 2018; You et al., 2022).

The presented visibility estimation method differs from the previous works in two aspects: (1) the proposed method does not need any labeled data as supervision for training the networks, and (2) the spatiotemporally continuous visibility across a large-scale city can be estimated via the active CVs. This provides a unique opportunity for developing many real-time weather-related applications.

2.2. Estimation of $PM_{2.5}$ mass concentration and airlight

$PM_{2.5}$ refers to particulate matters (PM) that own aerodynamic diameters of no more than 2.5 μm. Heterogeneous chemical compositions of $PM_{2.5}$ significantly impact aerosol light extinction including aerosol absorption and scattering (Xu et al., 2020; Cao et al., 2012; Tao et al., 2019), and thus degenerate visibility (Renhe et al., 2014; Wang et al., 2009; Watson, 2002) and influence transportation. Most importantly, $PM_{2.5}$ can be readily breathed into the lungs and further penetrate deep into the brain from the blood streams, causing serious health problems, e.g., cardiovascular disease, respiratory disease, and premature death (Agency, 2016). With the rapid industrial development in many countries, such as China, India, and Nepal, quickly increasing energy consumption leads to deteriorating air quality in urban areas (Chan and Yao, 2008; Kan et al., 2012), especially in the level of $PM_{2.5}$ mass concentration. Regular haze weather draws special attention of both the general public and academia (Huang et al., 2020). Accurate and dynamical detection of $PM_{2.5}$ mass concentration becomes crucial for air quality monitoring and travel planning so as to protect public health.

Various methods have been used in the estimation of $PM_{2.5}$ mass concentration. The commonest type of method is to devise and improve professional instruments based on various chemical principles (Chen et al., 2019a; Zhao et al., 2019; Pandolfi et al., 2018; Malm and Day, 2001). While the professional instrument can be costly and difficult for universal application. Considering the wide spatiotemporal distribution of $PM_{2.5}$, satellite-based remote sensing data were considered advantageous (Zheng et al., 2017; Chelani, 2019; Shelton et al., 2021; Van Donkelaar et al., 2006; Sun et al., 2019). However, the satellite data may not be economically and timely available. Moreover, several empirical models have also been proposed to estimate the $PM_{2.5}$ mass concentration from the atmospheric visibility (Wang et al., 2006; Ji et al., 2020). Nevertheless, the atmospheric visibility still needs to be measured using professional instruments; this limits the models' intensive use in dense estimations. Due to sophisticated meteorological changes and geographical differences between various urban areas, $PM_{2.5}$ mass concentrations exhibit high spatial–temporal variations in large-scale cities. The above-described instrument-based methods are hard to capture such variations in $PM_{2.5}$ mass concentration in light of the limited and fixed detector installations. Taking CVs as mobile "sensors", the proposed method can dynamically and densely estimate $PM_{2.5}$ mass concentrations for a whole city.

In airlight estimation, there are two types of methods in general: prior-based methods and learning-based methods. For the former, different priors have been studied, e.g., the brightest pixel prior (He et al., 2010; Fattal, 2008; Tan, 2008), color constancy prior (Gautam et al., 2020), color attenuation prior (Zhu et al., 2015), statistical priors (Berman et al., 2016; Bahat and Irani, 2016; Fattal, 2014), etc. While various priors may have distinct application conditions; it is challenging for common use. The latter is to take the raw image as input and then directly estimate airlight through DNN (Cai et al., 2016; Zhang and Patel, 2018; Wang et al., 2019b). The presented method is also a learning-based method but, it does not require any labeled data and is simultaneously estimated along with the other three tasks. In this paper, airlight is considered associated with chemical components in the air, so that it can be used in air quality analysis for protecting human health.

3. Problem statement

Given an image captured by the onboard camera, the goal is to simultaneously estimate the scene's depth map, visibility, airlight, and $PM_{2.5}$ mass concentration. Previous methods need to separately train the networks for estimating these quantities. For depth estimation, the image sequences are required for conducting self-supervised training. For the estimation of visibility, airlight, and $PM_{2.5}$ mass concentration, truthful labels are generally used for supervised training. One target of this study is to simultaneously estimate these quantities by devising a unified, self-supervised, and single-image training framework.

Traditional methods for measuring visibility, airlight, and $PM_{2.5}$ mass concentration are based on fixed meteorological stations, which are sparsely distributed in the city, and thus are difficult to capture the spatial variations. Another goal of this study is to find a solution for densely estimating visibility, airlight, and $PM_{2.5}$ mass concentration in a megacity, thereby providing more precise perception information.

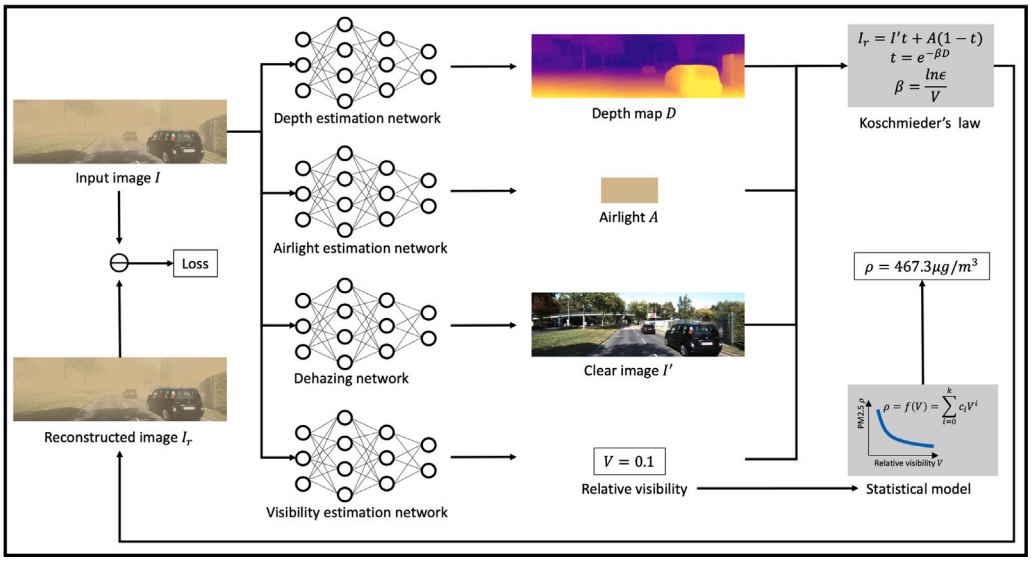


Fig. 1. Architecture of the proposed framework.

4. Method

This section introduces the detailed method. The overall architecture, hazy image synthesis model, self-supervised multi-task learning framework, and polynomial correlation model for $PM_{2.5}$ mass concentration estimation are presented, respectively.

4.1. Architecture overview

The whole system is trained in a self-supervised manner, meaning that no labeled data is required. Moreover, the system only takes a single image as input instead of an image sequence used in the previous self-supervised training framework. As shown in Fig. 1, the input image is passed to various sub-networks for estimating diverse information, including depth map, airlight, dehazed image, and visibility. This process is to conduct image decomposition. In depth estimation, the depth net used in Monodepth2 (Godard et al., 2019) was utilized for a fair comparison. The only difference is that the single-scale output was adopted in this study. Simply changing the number of output channels to 3 in depth estimation network offers the dehazing network. Moreover, ResNet-18 (He et al., 2016) and four convolutional layers with kernel sizes of $(1, 3, 3, 1)$ were respectively used as encoder and decoder in airlight and visibility estimation.

The estimated depth map, airlight, clear image, and visibility are then coupled with Koschmieder's law for reconstructing the input image. This is apparently an image synthesis process (to be introduced in the next part). The difference between the reconstructed image and the original input image, which is measured by L1 norm and structure similarity (SSIM), is used to train the system. Upon the completion of visibility estimation, a statistical polynomial correlation model is further utilized to map visibility to $PM_{2.5}$ mass concentration. It is noted that after training each sub-network can be independently used by taking a single image as input.

4.2. Hazy image synthesis model

A large-scale hazy image dataset is required for training deep neural networks, but it is not available currently. Nevertheless, large-scale clear image dataset, e.g., the KITTI dataset, is popular and publicly available. The KITTI dataset has been used to validate a bunch of computer vision algorithms and autonomous driving models. Multi-source data have been collected in real-world scenarios, including images, point clouds, and global positioning systems (GPS). The image data were collected by the onboard cameras. This meets the requirement of the presented work. Thus, the KITTI dataset was chosen to synthesize the hazy images, and thus train the proposed framework. Proposition 1 is proposed for generating hazy images as follows.

Proposition 1. *Given an image without haze, $I' \in \mathbb{R}^{H\times W\times 3}$, the corresponding depth map, $D \in \mathbb{R}^{H\times W\times 1}$, visibility, $V \in \mathbb{R}^1$, airlight, $A \in \mathbb{R}^{1\times 1\times 3}$, and the minimal observable contrast, ϵ, the hazy image with respect to the specified conditions, $I \in \mathbb{R}^{H\times W\times 3}$, can be obtained by the following*

$$I = I' \odot e^{\frac{\ln\epsilon}{V}R} + A \odot (1 - e^{\frac{\ln\epsilon}{V}R}), \tag{1}$$

$$R^{i,j} = \|P^{i,j}\|_2, R \in \mathbb{R}^{H\times W\times 1}, \tag{2}$$

$$P^{i,j} = D^{i,j}K^{-1}p, \tag{3}$$

where $R \in \mathbb{R}^{H\times W\times 1}$, $R^{i,j}$, $\odot$, $P \in \mathbb{R}^{H\times W\times 3}$, $P^{i,j}$, $K^{-1} \in \mathbb{R}^{3\times 3}$, and $p \in \mathbb{R}^{3\times 1}$ represent the range image of the scene, the range value at row i and column j, element-wise multiplication, the point clouds of the scene, the point cloud at row i and column j, the inverse of camera intrinsic matrix, and the pixel coordinate. The broadcasting technique will be used in the course of computation for Eq. (1).

Proof. Light intensity will suffer an attenuation along with the increase of travel distance, because of a series of physical effects, e.g., reflection and scattering. This causes that the observed luminance of objects that are located at various places can be different, even if the actual light intensities of them are identical. With Koschmieder's Law, the effect can be mathematically stated as

$$I = I' \odot T + A \odot (1 - T), \tag{4}$$

wherein $T \in \mathbb{R}^{H\times W\times 1}$ represents the transmission map of the scene, and is defined as

$$T = e^{-\beta R}. \tag{5}$$

In Eq. (5), e and $\beta \in R^1$ represent the natural base and extinction coefficient. In Eq. (4), the airlight, A, is defined as the color of light which has been scattered or diffused in the air by dust, haze, or fog (haze is considered in this paper). This color, in general, is regarded as the location-free variable, i.e., the haze colors over different locations in a scene are homogeneous. With priors to the color of haze, it is convenient to generate various A which are close to five possible haze colors: (1) white, (2) blue gray, (3) yellow, (4) gray, and (5) sepia.

Let the RGB representation for color (i) be $(R^i, G^i, B^i), \forall i \in [1,5]$. The airlight, (R, G, B), for an arbitrary image is determined by the following two steps:

- Randomly choose the haze color from the given five possible colors. Each color has the same probability of 1/5=0.2 being selected.
- For any selected haze color $(R^i, G^i, B^i), \forall i \in [1,5]$, the airlight, (R, G, B), is generated by

$$R = R^i + \Delta R, \Delta R \sim U(-r, r). \tag{6}$$

$$G = G^i + \Delta G, \Delta G \sim U(-g, g). \tag{7}$$

$$B = B^i + \Delta B, \Delta B \sim U(-b, b). \tag{8}$$

where $U(a, b)$ represents the uniform distribution with the support $[a, b]$; ΔR, ΔG, and ΔB are samples from the respective uniform distributions.

Given that R, G, and B are measured with the range of 0 to 255, this paper sets $r = g = b = 10$. Note that all values cannot exceed the valid range $[0, 255]$. Then, the key is to determine the transmission map, T, which depends on β and R.

Using any depth estimation model, e.g., Monodepth2 (Godard et al., 2019), the depth map can then be estimated. It is noted that a depth estimation method is solely used in haze image synthesis for validating the proposed framework. In practice, the haze images are directly obtained from the environment. Thus, the proposed approach is fully independent of depth estimation models. Given the depth map of the scene, we can easily get the range map, R, by computing the L_2 norm for each space coordinate in the point cloud matrix, P, which can be obtained by reprojection described in Eq. (3). Therefore, the problem comes to find the extinction coefficient, β.

As the contrast, C, is defined as

$$C = \frac{\hat{p_o} - \hat{p_b}}{\hat{p_b}}, \tag{9}$$

wherein $\hat{p_o}$ and $\hat{p_b}$ represent the light intensity of object and the baseline intensity. Generally, the horizon sky is chosen for $\hat{p_b}$, i.e., $\hat{p_b} = A$, the contrasts for the clear and hazy images, $C_{I'}$ and C_I, is written as

$$C_{I'} = \frac{I' - A}{A}, \tag{10}$$

$$\begin{aligned} C_I = \frac{I - A}{A} &= \frac{I' \odot T + A \odot (1 - T) - A}{A} \\ &= \frac{I' - A}{A} \odot T = C_{I'} \odot T, \end{aligned} \tag{11}$$

where pixel values are directly used as the measurement of light intensity for simplicity.

Given that the minimal observable contrast, ϵ, is defined as the absolute contrast between the black object and airlight, it follows

$$|C_{I'}| = |\frac{0 - A}{A}| = 1, \tag{12}$$

$$|C_I| = |C_{I'} \odot T| = |-T| \geq \epsilon. \tag{13}$$

Substituting Eq. (5) into Eq. (13) furnishes

$$e^{-\beta R} \geq \epsilon. \tag{14}$$

When equality holds in Eq. (14), the range becomes visibility, i.e.,

$$e^{-\beta R} = \epsilon. \tag{15}$$

Thus, β can be derived accordingly,

$$\beta = -\frac{\ln \epsilon}{V}. \tag{16}$$

Substituting Eqs. (16) and (5) into Eq. (4) offers Eq. (1). □

Remark 1. Although Proposition 1 is used to generate haze images using clear images, it also can be used in mimicking other types of contaminated images by simply changing the value of airlight, such as foggy images, snowy images, night images, and so on.

With Proposition 1, different hazy images, as shown in Fig. 3, can be gracefully generated based on the images without haze, which will be further used to train the neural networks in the proposed framework.

4.3. Self-supervised multi-task learning framework

This subsection introduces a novel bidirectional process of image synthesis and decomposition (BPISD) training pipeline, and thus a self-supervised multi-task learning framework for driving environment perception. Using Proposition 1, a hazy image can be synthesized using the given clear image, depth map, airlight, and visibility. Inspired by this image synthesis process, the opposite image decomposition course can be deployed to estimate those components used for image synthesis. Thus, the following corollary is proposed.

Corollary 1. *Given a hazy image, $I \in \mathbb{R}^{H\times W\times 3}$, and its corresponding clear image, $I' \in \mathbb{R}^{H\times W\times 3}$, the range image, $R \in \mathbb{R}^{H\times W\times 1}$, airlight, $A \in \mathbb{R}^{1\times 1\times 3}$, and visibility, $V \in \mathbb{R}^1$, of the given hazy image can be estimated by solving the following minimization problem:*

$$\begin{aligned} &min \quad \Phi(I_r, I) \\ &s.t. \quad I_r = I' \odot e^{\frac{\ln \epsilon}{V} R} + A \odot (1 - e^{\frac{\ln \epsilon}{V} R}), \end{aligned} \tag{17}$$

where $\Phi(\cdot)$ represents the function that is to evaluate the difference between the reconstructed image I_r and the input image I.

Proof. Based on Proposition 1, the target quantities, R, A, and V, and the given clear image, I', can synthesize haze images. Given a specific set of values $\{R, A, V\}$, the corresponding haze image is determined. By updating $\{R, A, V\}$ towards the direction of minimizing the distance between the synthesized image I_r and the original haze image I, the resulting solution, $\{R^*, A^*, V^*\}$, will converge to the range image, airlight, and visibility of the given haze image. Therefore, the range image, airlight, and visibility of the given haze image can be estimated. □

Remark 2. The problem stated in Corollary 1 is a typically ill-posed problem. If I' is not given, which thus has to be estimated, the solution of the presented minimization can collapse to some unexpected cases, e.g., $I' = I$, $R = \mathbf{0}$, $A = \mathbf{0}$, and V can be any non-zero number. To address this issue, some constraints can be imposed on the estimated either I', R, A, or V. I' was considered known under this circumstance for successfully estimating other quantities, because (1) the corresponding clear image of I can be readily obtained by retrieving historical image data of the same location captured by the on-board cameras outfitted on the CVs in a city; and (2) dehazing has been well studied, and a number of supervised and unsupervised methods (He et al., 2010; Engin et al., 2018; Yang and Sun, 2018) have been proposed, which can be borrowed to estimate the clear images.

Remark 3. As the range image has the same resolution as compared with the input image and thus is high-dimensional, it is challenging to solve Eq. (17) through typical algorithms in the field of optimization. Nevertheless, it is apparent that R, A, and V are dependent on I, meaning that R, A, and V can somehow be derived from I. Despite the impossibility of deriving an explicit relationship between I and $R/A/V$ (at least it is impossible at this stage), powerful DNNs can be deployed to fit such correlations. Then, it is convenient to train DNNs using gradient-based optimizers.

Based on Corollary 1, the proposed framework is constituted by multiple tasks, including estimations of range image, airlight, and visibility. Various deep neural networks were used to estimate these quantities. Consider the range estimation network as Ψ_R, which is defined as $\Psi_R : I \in \mathbb{R}^{H\times W\times 3} \rightarrow R \in \mathbb{R}^{H\times W\times 1}$. A popular encoder–decoder architecture presented in Monodepth2 (Godard et al., 2019) was adopted for a fair comparison. Differently, single-scale output was utilized instead of four-scale output in the original paper. Similarly, the airlight and visibility estimation networks are denoted as $\Psi_A : I \in \mathbb{R}^{H\times W\times 3} \rightarrow A \in \mathbb{R}^{1\times 1\times 3}$ and $\Psi_V : I \in \mathbb{R}^{H\times W\times 3} \rightarrow V \in \mathbb{R}^1$, respectively. Ψ_A and Ψ_V can either be two separate networks or share the same network with two separate estimation heads. The latter strategy was adopted here. Specifically, ResNet-18 (He et al., 2016) and four convolutional layers were respectively taken as the encoder and decoder in this paper. To obtain a clear image, the depth estimation network with minor changes to the last layer was used to dehaze in a supervised manner. Consider the dehazing network as $\Psi_C : I \in \mathbb{R}^{H\times W\times 3} \rightarrow I' \in \mathbb{R}^{H\times W\times 3}$. The training is to solve the following minimization problem:

$$\begin{aligned} &min \quad \Theta(I', I^c) \\ &s.t. \quad \Theta(I', I^c) = \alpha\|I' - I^c\|_1 + (1-\alpha)\frac{1 - SSIM(I', I^c)}{2}, \end{aligned} \tag{18}$$

where $SSIM(\cdot)$ and I^c represent the structure similarity function and the ground truth clear image; α is set to 0.15 following Jia et al. (2022). It is noted that the trained dehazing model will be directly used when performing multi-task learning and no longer be trained. As mentioned, clear images can also be estimated by other means, even they can be directly obtained by searching in the historical data.

For each forward process, Φ_R, Φ_A, and Φ_V take a single hazy image as input to estimate its range image, airlight, and visibility. The trained Φ_C also takes the hazy image as input and outputs the clear image for further actions. Then, Proposition 1 is used to reconstruct the input hazy images based on the estimated R, A, I' and V, as shown in Eq. (17). Let R_1 and R_2 be the estimated range images for I and I'. Correspondingly, denote the reconstructed images based on R_1 and R_2 as $I_{r,1}$ and $I_{r,2}$. The reconstruction error can be written as

$$L_1 = \sum_{i=1}^{2} \Theta(I_{r,i}, I). \tag{19}$$

The estimated ranges images R_1 and R_2 are expected to be consistent. Thus, the consistency loss is given by

$$L_2 = \Theta(R_1, R_2). \tag{20}$$

Finally, the edge-aware smoothness loss, as follows, is also adopted.

$$L_3 = \beta \sum_{i=1}^{2} |\partial_x \hat{R}_i^*| e^{-|\partial_x I'|} + |\partial_y \hat{R}_i^*| e^{-|\partial_y I'|}, \tag{21}$$

where $\hat{R}_i^* = \frac{\hat{R}_i}{E(\hat{R}_i)}$ is the mean-normalized inverse range following Godard et al. (2019); $E(\cdot)$ and ∂ represent the expectation operation and the partial derivative operator, respectively; and β is set to 0.001. The final loss is given by

$$L = \Phi(I_r, I) = \sum_{i=1}^{3} L_i. \tag{22}$$

The Adam optimizer (Kingma and Ba, 2014) with the initial learning rate of 0.0001, batch size of 12, and other default parameters was used to train the networks. The numbers of epochs for dehazing model and multi-task learning framework were set to 100 and 20, during which the initial learning rates will be decreased by a factor of 10 at the 95th and the 15th epochs. The pretrained model on ImageNet was loaded to initialize the parameters of ResNet-18.

4.4. Polynomial correlation model for $PM_{2.5}$ mass concentration estimation

Applying the trained visibility estimation model, the visibility of a given image can be obtained. To further estimate the $PM_{2.5}$ mass concentration, a simple polynomial correlation model was adopted to model the relationship between visibility and $PM_{2.5}$ mass concentration. It follows that

$$\hat{\rho} = \sum_{i=0}^{k} c_i V^i, \tag{23}$$

where $\hat{\rho}$, c_i, and k represent the estimated $PM_{2.5}$ mass concentration, the coefficients of monomials, and the order of the polynomial, respectively. Thus, the key is to estimate the coefficients, $c_i, \forall i \in [1, k]$. The problem can be formulated as

$$min \sum_{j=1}^{N} (\sum_{i=0}^{k} c_i V_j^i - \rho_j)^2, \tag{24}$$

where N and ρ_j represent the number of samples and truthful $PM_{2.5}$ mass concentration. The least square method was utilized to solve the above minimization. The fitted model can then be used to estimate $PM_{2.5}$ mass concentration.

Remark 4. It is noted that a small set of truthful $PM_{2.5}$ mass concentrations is available from the sparse meteorological stations, which can be used to calibrate the polynomial correlation model. The calibration process only needs to be conducted once.

5. Experiment

In this section, datasets, definitions of metrics, and experimental results are introduced, respectively. All experiments are implemented with PyTorch 1.9.1 on a single TITAN RTX card and same set of hyperparameters to have fair comparisons.

5.1. Datasets

For self-supervised depth estimation, a large-scale and popular dataset, KITTI (Geiger et al., 2013), was used to train the networks. Based on Zhou et al.' split (Zhou et al., 2017) on the KITTI dataset, 40,109 images with various visibility conditions were first generated in terms of Proposition 1 to mimic the different levels of $PM_{2.5}$ mass concentrations. Five visibility scales, i.e., Terrible, Bad, Middle, Good, and Perfect, were considered with respect to the corresponding relative visibilities of 0.1, 0.3, 0.5, 0.8, and 1, where 1 represents the clear images without haze. In each visibility group, 697 images were used for testing. The image resolution is set to 640 × 192.

It is noted that some images in the original test set of Eigen's split are not suitable for evaluating visibility and airlight as the scenes are extremely close to the cameras or are full of buildings and pedestrians. Thus, in visibility and airlight estimation, 377 images in Eigen's split (Zhou et al., 2017) of the KITTI dataset were chosen for generating testing images, which are synthesized based on the given clear images and randomly generated relative visibilities in the range of 0 to 1.

To validate the effectiveness of the proposed method in estimating $PM_{2.5}$ mass concentration, real-world images with various $PM_{2.5}$ mass concentrations were collected. The image data were manually selected on the Beijing Tour website, on which the real-time image data and detailed weather information are provided, including temperature, humidity, wind, etc. The corresponding $PM_{2.5}$ mass concentration of each image is obtained from the U.S. Embassy in Beijing. The numbers of images in relative humidity 0 to 0.5, 0.5 to 0.7, 0.7 to 0.9, and 0.9 to 1 are 39, 17, 13, and 24, respectively, i.e., 93 images with different $PM_{2.5}$ mass concentrations were used for validation.

Table 1
Model training configurations. √ and × represent with and without corresponding truthful information during training.

Model	Clear image?	Airlight?	Visibility?
SSD-A	√	√	√
SSD-B	√	√	×
SSD-C	√	×	×
SSD-D	×	×	×

5.2. Definitions of metrics

This subsection introduces the metrics used for evaluations. Denote the ground truth depth map and the predicted depth map as $D^{gt} \in \mathbb{R}^{h\times w}$ and $D^{pre} \in \mathbb{R}^{h\times w}$, where h and w are the height and width of the depth map, respectively. N is the number of valid pixels in the ground truth depth map. Thus, the metrics used for evaluating depth estimation are defined as follows:

- Absolute relative (AbsRel) error (Eq. (25)):

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i^{pre} - D_i^{gt}|}{D_i^{gt}}; \tag{25}$$

- Square relative (SqRel) error (Eq. (26)):

$$SqRel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i^{pre} - D_i^{gt}|^2}{D_i^{gt}}; \tag{26}$$

- Root mean square (RMS) error (Eq. (27)):

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|D_i^{pre} - D_i^{gt}|^2}; \tag{27}$$

- Root mean square logarithm (RMSlog) error (Eq. (28)):

$$RMSlog = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|log\, D_i^{pre} - log\, D_i^{gt}|^2}; \tag{28}$$

- δ_T accuracy (Eq. (29)):

$$\delta_T = \frac{\sum_{i=1}^{N}(max(\frac{D_i^{pre}}{D_i^{gt}}, \frac{D_i^{gt}}{D_i^{pre}}) < T)}{N}. \tag{29}$$
$$T = 1.25, 1.25^2, 1.25^3$$

In the evaluation of visibility, airlight, and $PM_{2.5}$ mass concentration estimation, RMS error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were adopted. For each test item, the metrics stated above are computed. Then, the final results are obtained by averaging all testing data.

5.3. Results

The proposed co-training system is constituted by four sub-tasks: estimations of the depth map, airlight, visibility, and $PM_{2.5}$ mass concentration. All these four tasks are evaluated.

5.3.1. Self-supervised depth estimation

Current self-supervised depth estimation systems are based on view synthesis and use reference images to reconstruct the target image. The reconstruction error is taken as the loss to train the networks. Such a view-synthesis approach requires an image sequence as input for training. However, videos may not be always available, due to limited transmission bandwidth and energy supply on edge devices. Instead, the presented self-supervised depth estimation system solely takes a single image as input, i.e., performing single-image and self-supervised depth (SSD) estimation. The detailed model training configurations are shown in Table 1. This provides a new paradigm for conducting monocular depth estimation.

Table 2 compares the proposed method with other state-of-the-art methods on the original KITTI dataset. It should be noted that (1) "SSD-B, C, and D" perform self-supervised multi-task learning, which are much more challenging than other methods of only conducting depth estimation and (2) other methods require image sequence (video) as input to train the network while SSD only takes a single image as input. The numbers in Table 2 indicate that SSD-A outperforms other state-of-the-art methods almost on all metrics by clear margins. "SSD-B, C, and D" only exhibit very minor performance drops compared with "SSD-A". "SSD-A, B, C, and D" show the ablation studies by gradually relaxing the training conditions, demonstrating that the proposed system can accurately estimate the depth map as well as the byproducts (i.e., visibility and airlight), even with the estimated clear images.

In particular, Table 3 and Fig. 2 reports the quantitative results of SSD under various visibility conditions on the KITTI dataset. Detailed ablation studies on the training conditions were also presented. "SSD–A" model was trained with truthful clear images, airlight, and visibility, which outperforms Monodepth2 (Godard et al., 2019) on various visibility conditions. In particular, low-visibility conditions significantly degrade the performance of Monodepth2. In contrast, the proposed method solely undergoes neglectable performance drops. "SSD–B, C, and D" gradually relax the training condition to without truthful visibility, without truthful visibility and airlight, and finally without truthful visibility, airlight, and clear image. All these unknown data were simultaneously estimated through the proposed framework. The results indicate that the performance only slightly drops as compared with that of "SSD–A". For a perfect visibility situation, the performance of the final model "SSD–D" is still on par with that of Monodepth2 (Godard et al., 2019). In other hazy conditions, "SSD–D" exhibits huge advantages on all metrics as compared with Monodepth2 (Godard et al., 2019). This clearly demonstrates the effectiveness and robustness of the proposed framework.

Some qualitative comparisons across various visibility conditions are presented in Fig. 3. Given the perfect-visibility images, the qualitative results from the proposed method and Monodepth2 (Godard et al., 2019) are highly consistent. Nevertheless, with the innovative training framework, the proposed method performs much better than Monodepth2 on some difficult regions, e.g., windows. With visibility deteriorating, many details and far scenes could not be estimated in Monodepth2. The proposed method still provides accurate estimations with vivid details.

5.3.2. Self-supervised visibility and airlight estimation

Visibility for a given image is represented by a scalar, which is considered highly associated with traffic management and safety. Applying the trained self-supervised visibility estimation model to the selected KITTI data, it was found that the RMSE, MAE, and MAPE of the proposed method are 0.032, 0.025, and 4.9%, respectively.

Airlight, which is represented by a 3D vector (i.e., red, green, and blue components), is caused by light scattering and diffusion by particulate matters, e.g., dust and haze. Airlight owns spatiotemporal variations due to the changes in chemical components in the air. Conversely, knowledge of airlight plays a critical role in analyzing the chemical components in the air and thus provides important guidance to improve air quality and more precise warnings to the public. Applying the trained airlight estimation model to the same set of 337 images used for self-supervised visibility estimation, the RMSE, MAE, and MAPE were found to be 0.067, 0.049, and 9.2%, respectively. Fig. 4a, b, c, and d show the estimated quantities versus the ground truth quantities graphs. The dots are well distributed along the 45-degree lines, meaning that the estimations are sufficiently accurate. Fig. 4e presents some estimation examples. The minor prediction deviations clearly demonstrate the effectiveness of the proposed framework on self-supervised visibility and airlight estimation.

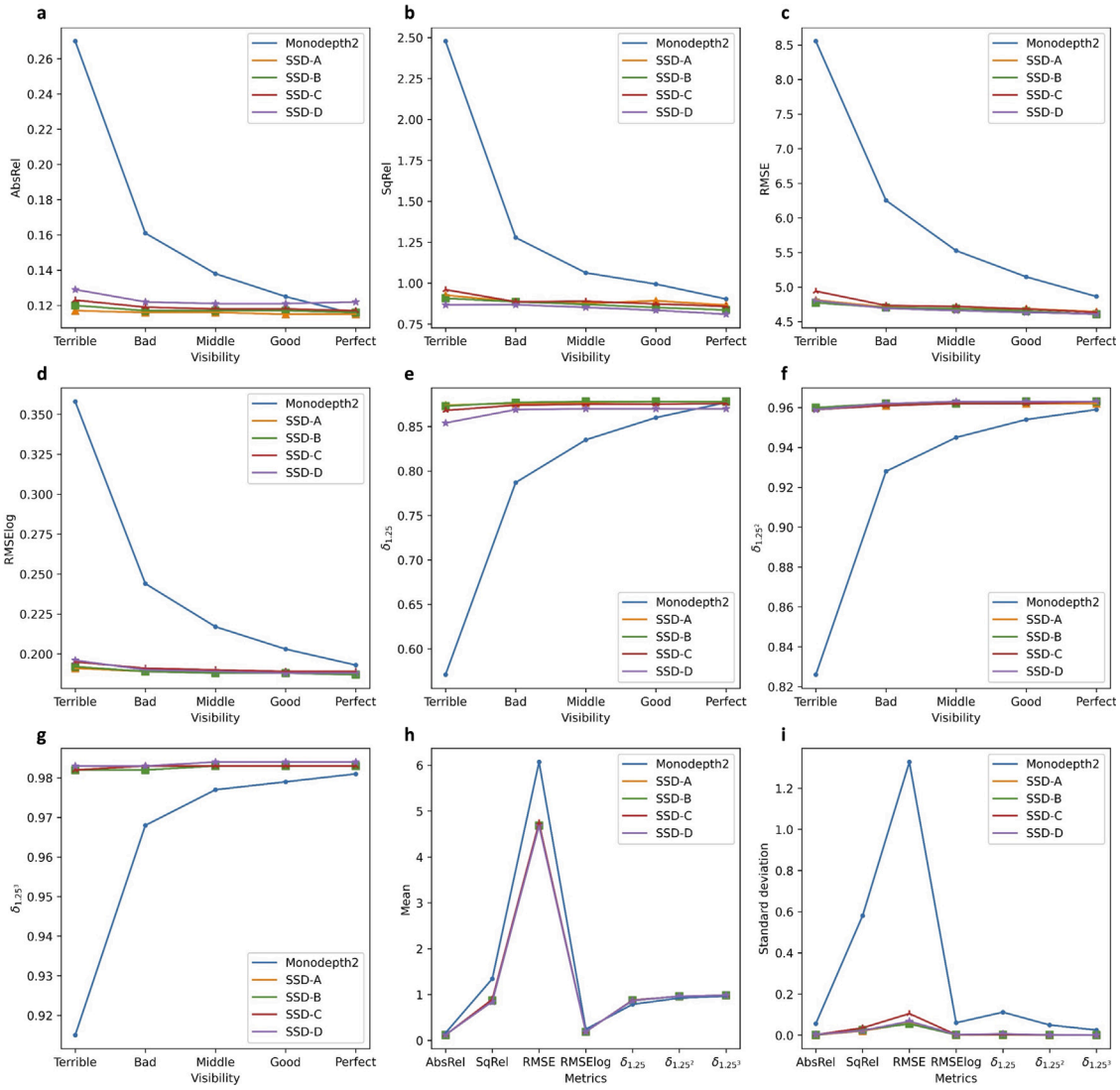


Fig. 2. Results of self-supervised depth estimation on the KITTI dataset (Geiger et al., 2013). The results on various evaluation metrics, including AbsRel, SqRel, RMSE, RMSElog, $\delta_{1.25}$, $\delta_{1.25^2}$, $\delta_{1.25^3}$, are shown in a, b, c, d, e, f, and g, respectively. h and i represent the means and standard deviations of different metrics across various visibility conditions.

5.3.3. $PM_{2.5}$ mass concentration estimation

Assuming that low visibility is solely caused by $PM_{2.5}$, $PM_{2.5}$ mass concentration can be derived from the atmosphere visibility. By directly applying the trained visibility estimation model on the KITTI dataset to real-world data, the relative visibilities were first estimated. Then, the polynomial correlation model was used to match the estimated relative visibilities with the truthful $PM_{2.5}$ mass concentrations.

As shown in Table 4, Fig. 5a, b, c, and d, the estimation performance gradually gets improved and being asymptotically stable with the increase of polynomial order. The absolute percentage errors were well confined within 8% as the polynomial order is equal to or greater than 6. RMSE and MAE were respectively confined within 41 and 31 over various humidity conditions. Fig. 5e presents some estimation results by setting the polynomial order to 10. Excellent performance indicates that the proposed method can accurately estimate $PM_{2.5}$ mass concentrations under various humidity conditions. It has great potential to be implemented in an urban system to dynamically monitor $PM_{2.5}$ mass concentrations.

6. Conclusion

This paper proposes a novel framework to simultaneously conduct the estimations of the depth map, airlight, visibility, and $PM_{2.5}$ mass concentrations leveraging the CV technologies. The on-road CVs can share local information with other vehicles and data centers. Image data collected by the onboard cameras are used in this study. Due to the different trips of CVs, it is assumed that CVs are well distributed in the city. Consider a specific time instant; the images in different urban regions can be obtained via CVs. Along with the traverse of CVs in the whole city, the local images can be regularly updated. Given multiple observations reported by CVs at the same location, the average of these multiple estimates can be used as the final estimation to enhance robustness and improve accuracy. Moreover, the spatial resolution can be readily adjusted by setting various data report frequencies of CVs. Real-time precise airlight, visibility, $PM_{2.5}$ mass concentrations, and depth maps can be estimated accordingly.

The proposed framework solely requires a single input image without any labels for all sub-tasks, working in a self-supervised manner. Moreover, it is a non-intrusive method, meaning that no additional equipment, professional instruments, and special adjustments to vehicles are involved. Thus, it is considered more convenient and flexible as compared with other methods in estimating those quantities. Comprehensive experiments demonstrate the effectiveness and superiority of the proposed method. In depth estimation, the proposed method

Table 2
Quantitative comparisons of depth estimation on the KITTI dataset (Geiger et al., 2013). GT and PT represent ground truth and pretraining, respectively. – represents that the situation is unclear. Res. represents resolution. The best performances are marked **bold**.

Res.	Methods	GT?	PT?	Errors ↓				Errors ↑		
				AbsRel	SqRel	RMS	RMSlog	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
–	Eigen et al. (2014), coarse	√	–	0.214	1.605	6.563	0.292	0.673	0.884	0.957
	Eigen et al. (2014), fine	√	–	0.203	1.548	6.307	0.282	0.702	0.890	0.958
	Liu et al. (2015b)	√	–	0.202	1.614	6.523	0.275	0.678	0.895	0.965
	Kuznietsov et al. (2017)	√	√	0.113	0.741	4.621	0.189	0.862	0.960	0.986
	Fu et al. (2018)	√	√	**0.072**	**0.307**	**2.727**	**0.120**	**0.932**	**0.984**	**0.994**
416 × 128	Yin and Shi (2018) (VGG)		–	0.164	1.303	6.090	0.247	0.765	0.919	0.968
	Yin and Shi (2018) (ResNet)		–	0.155	1.296	5.857	0.233	0.793	0.931	0.973
	Wang et al. (2018)		√	0.151	1.257	5.583	0.228	0.810	0.936	0.974
	Bian et al. (2019)		√	0.149	1.137	5.771	0.230	0.799	0.932	0.973
	Casser et al. (2019)		–	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	Jia et al. (2021)		√	0.144	**0.966**	5.078	0.208	0.815	0.945	**0.981**
	Klingner et al. (2020)		√	**0.128**	1.003	5.085	0.206	0.853	0.951	0.978
	Godard et al. (2019)		√	**0.128**	1.087	5.171	0.204	**0.855**	0.953	0.978
	Jia et al. (2022) (CC)		√	**0.128**	0.990	5.064	**0.202**	0.851	**0.955**	0.980
	Jia et al. (2022) (CL)		√	**0.128**	0.979	**5.033**	**0.202**	0.851	0.954	0.980
	Zhou et al. (2017)			0.208	1.768	6.856	0.283	0.678	0.885	0.957
	Yang et al. (2017)			0.182	1.481	6.501	0.267	0.725	0.906	0.963
	Godard et al. (2019)[a]			0.144	1.059	5.289	0.217	0.824	0.945	0.976
	Jia et al. (2022) (LL)			0.141	1.060	5.247	0.215	0.830	0.944	0.977
	Jia and Yao (2023)-S			0.135	0.973	5.084	0.208	0.840	0.948	0.978
	Jia and Yao (2023)-L			**0.128**	**0.897**	**4.905**	**0.200**	**0.852**	**0.953**	**0.980**
640 × 192	Godard et al. (2019)			**0.132**	1.044	5.142	0.210	**0.845**	0.948	0.977
	Jia et al. (2022) (LL)			0.135	**0.979**	**5.078**	**0.209**	0.841	**0.949**	**0.978**
	Klingner et al. (2020)		√	0.117	0.907	4.844	0.196	0.875	0.958	0.980
	Godard et al. (2019)		√	**0.115**	0.903	4.863	0.193	0.877	0.959	0.981
	Jia et al. (2022) (CL)		√	0.116	0.886	4.787	0.192	0.876	0.959	0.981
	Jia et al. (2022) (CC)		√	0.116	0.842	4.708	0.190	0.876	0.961	0.982
	Jia et al. (2022) (CC)		√	0.116	0.842	4.708	0.190	0.876	0.961	0.982
	Our (SSD-A)		√	**0.115**	0.866	4.643	0.188	**0.878**	0.962	0.983
	Our (SSD-B)		√	0.116	0.836	**4.607**	**0.187**	**0.878**	**0.963**	0.983
	Our (SSD-C)		√	0.117	0.858	4.635	0.189	0.876	**0.963**	0.983
	Our (SSD-D)		√	0.122	**0.810**	4.611	0.188	0.870	**0.963**	**0.984**

[a] The results are reproduced by Jia et al. (2022).

Table 3
Quantitative results of self-supervised depth estimation under different visibility conditions.

Methods	Visibility	Errors ↓				Errors ↑		
		AbsRel	SqRel	RMS	RMSlog	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Godard et al. (2019)	Perfect	**0.115**	0.903	4.863	0.193	0.877	0.959	0.981
SSD-A	Perfect	**0.115**	0.866	4.643	0.188	**0.878**	0.962	0.983
SSD-B	Perfect	0.116	0.836	**4.607**	**0.187**	**0.878**	**0.963**	0.983
SSD-C	Perfect	0.117	0.858	4.635	0.189	0.876	**0.963**	0.983
SSD-D	Perfect	0.122	**0.810**	4.611	0.188	0.870	**0.963**	**0.984**
Godard et al. (2019)	Good	0.125	0.994	5.147	0.203	0.860	0.954	0.979
SSD-A	Good	**0.115**	0.892	4.686	**0.188**	**0.878**	0.962	0.983
SSD-B	Good	0.117	0.851	4.653	**0.188**	**0.878**	**0.963**	0.983
SSD-C	Good	0.118	0.873	4.682	0.189	0.875	0.962	0.983
SSD-D	Good	0.121	**0.834**	**4.632**	**0.188**	0.870	**0.963**	**0.984**
Godard et al. (2019)	Middle	0.138	1.063	5.526	0.217	0.835	0.945	0.977
SSD-A	Middle	**0.116**	0.876	4.683	**0.188**	0.877	0.962	0.983
SSD-B	Middle	0.117	0.870	4.688	**0.188**	**0.878**	0.962	0.983
SSD-C	Middle	0.118	0.889	4.720	0.190	0.875	0.962	0.983
SSD-D	Middle	0.121	**0.852**	**4.660**	0.189	0.870	**0.963**	**0.984**
Godard et al. (2019)	Bad	0.161	1.278	6.253	0.244	0.787	0.928	0.968
SSD-A	Bad	**0.116**	0.887	4.709	**0.189**	0.876	0.961	**0.983**
SSD-B	Bad	0.117	0.886	4.703	**0.189**	**0.877**	**0.962**	0.982
SSD-C	Bad	0.119	0.886	4.733	0.191	0.874	0.961	**0.983**
SSD-D	Bad	0.122	**0.868**	**4.691**	0.190	0.869	**0.962**	**0.983**
Godard et al. (2019)	Terrible	0.270	2.478	8.554	0.358	0.571	0.826	0.915
SSD-A	Terrible	**0.117**	0.926	4.813	**0.191**	**0.874**	**0.960**	0.982
SSD-B	Terrible	0.120	0.907	**4.772**	0.192	0.873	**0.960**	0.982
SSD-C	Terrible	0.123	0.960	4.940	0.195	0.868	0.959	0.982
SSD-D	Terrible	0.129	**0.867**	4.803	0.196	0.854	0.959	**0.983**

Table 4
Quantitative results of $PM_{2.5}$ mass concentration estimation on the real-world data. RH represents relative humidity. Order represents the polynomial fitting order.

Order	$0 \le RH < 0.5$			$0.5 \le RH < 0.7$			$0.7 \le RH < 0.9$			$0.9 \le RH < 1$		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
1	55.8	36.8	16.0	58.2	49.7	13.0	61.0	48.7	13.4	42.6	33.3	5.9
2	37.4	26.4	11.5	47.9	38.5	10.3	43.0	36.4	10.6	37.6	30.9	5.4
3	22.4	18.5	8.9	43.9	36.8	9.7	18.3	14.6	3.9	37.6	30.9	5.4
4	21.2	17.2	8.2	40.9	34.3	9.0	18.1	14.3	3.8	32.9	25.9	4.5
5	20.1	15.1	7.1	40.8	34.2	9.0	16.7	13.9	4.0	32.9	26.0	4.5
6	19.2	13.5	6.3	40.3	30.7	7.9	12.6	9.9	2.9	30.4	24.1	4.1
7	18.9	13.4	6.2	40.1	31.0	8.0	11.6	7.1	2.9	30.3	23.8	4.1
8	18.9	13.4	6.2	35.1	27.5	7.3	11.6	7.1	2.1	30.2	23.7	4.1
9	18.8	13.0	6.0	34.2	27.0	7.2	11.6	7.1	2.1	30.1	23.4	4.0
10	18.7	13.3	6.2	33.7	26.9	7.3	10.7	6.1	1.8	29.1	22.6	3.8

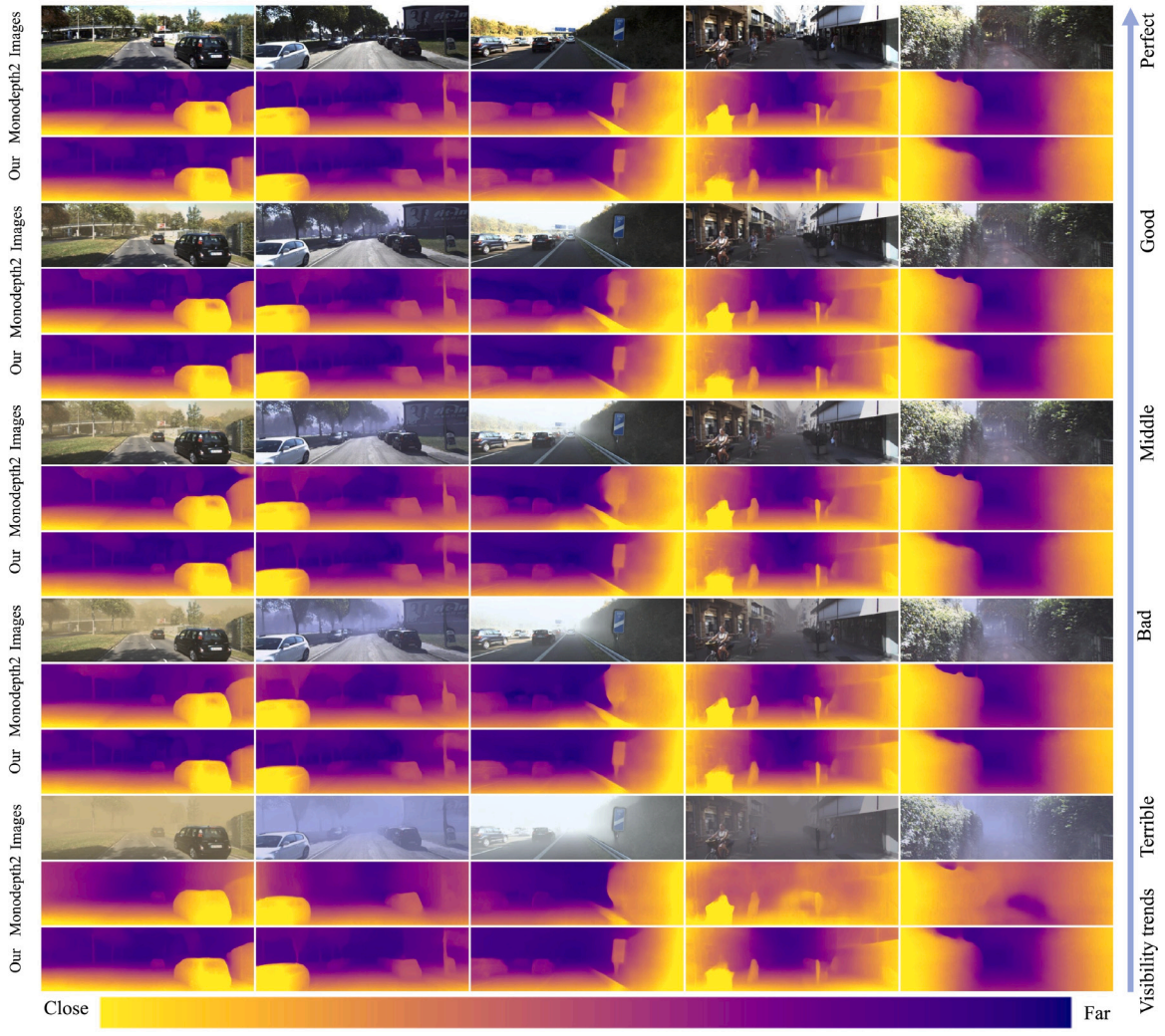


Fig. 3. Qualitative comparisons of depth estimation under various visibility conditions on the KITTI dataset (Geiger et al., 2013).

achieves performance competitive to current self-supervised methods when taking clear images as input, and significantly outperforms current methods when taking hazy images as input. Minor estimation derivations on airlight, visibility, and $PM_{2.5}$ mass concentration further show its great application potential. Nevertheless, the proposed method has the following limitations: (1) given the monocular image as input, the estimated depth map and visibility are relative values, and (2) considering different climate situations in different cities, the calibration process for $PM_{2.5}$ mass concentration estimation may need to be reconducted in different cities. We expect to develop real-world applications in the future.

CRediT authorship contribution statement

Shaocheng Jia: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review &

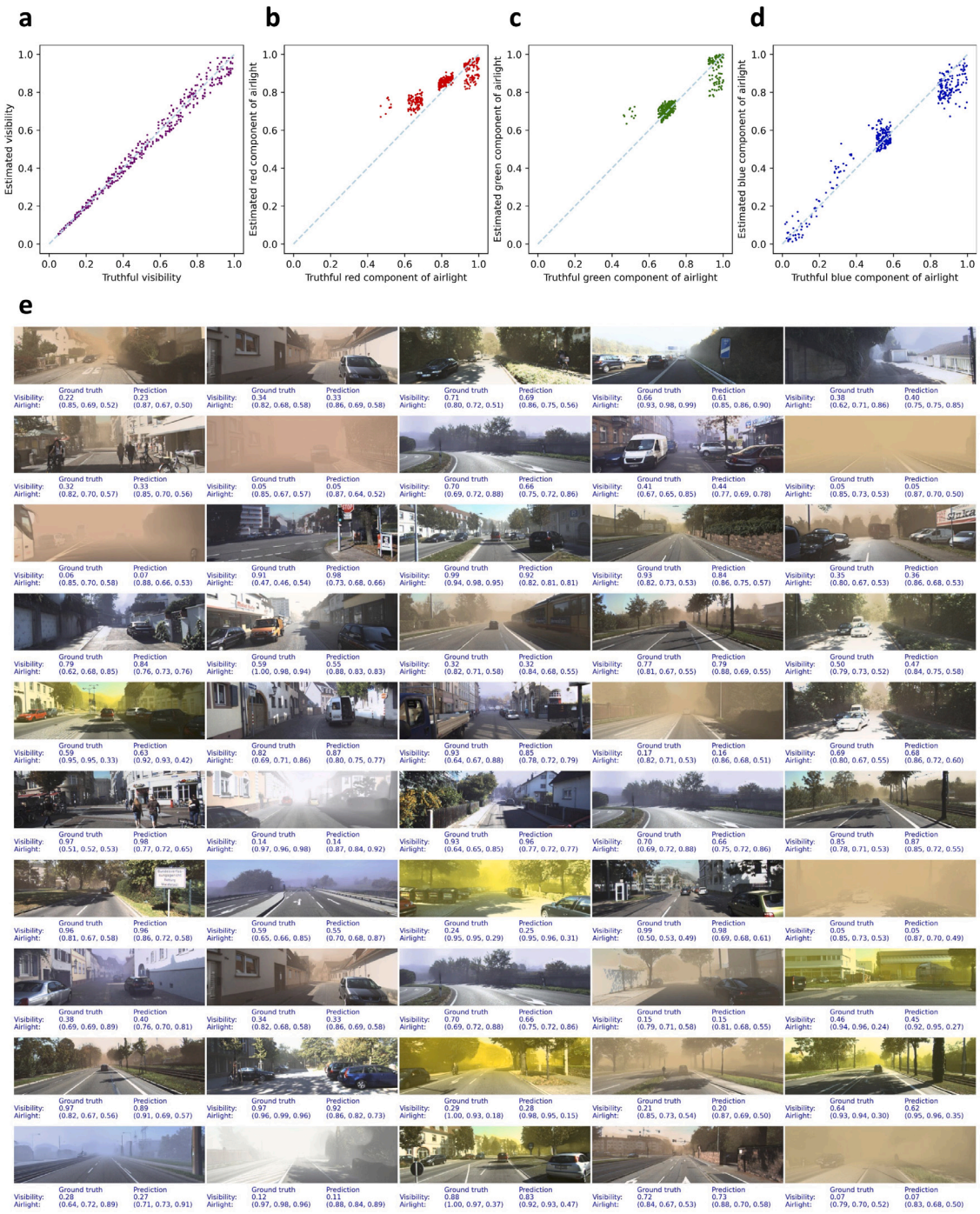


Fig. 4. Results of self-supervised visibility and airlight estimations on the KITTI dataset (Geiger et al., 2013). a: the estimated visibility versus ground truth visibility. b/c/d: the estimated red/green/blue component of airlight versus truthful red/green/blue component. e: some estimation samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

editing. **Wei Yao:** Conceptualization, Validation, Project administration, Funding acquisition, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work described in this paper was partially supported by the National Natural Science Foundation of China (Project No. 42171361), The Hong Kong Polytechnic University under Projects 1-ZVN6, 1-ZECE, and Q-CDAU.

Fig. 5. Results of $PM_{2.5}$ mass concentration estimation on the real-world data. a, b, c, d: estimation errors with the relative humidity of 0 to 0.5, 0.5 to 0.7, 0.7 to 0.9, and 0.9 to 1. e: some estimation examples (the unit for numbers is $\mu g/m^3$).

References

Agency, U.E.P., 2016. Health and Environmental Effects of Particulate Matter (PM). US Environmental Protection Agency.

Almalioglu, Y., Saputra, M.R.U., de Gusmao, P.P., Markham, A., Trigoni, N., 2019. Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In: Proc. IEEE Int. Conf. Rob. Autom.. pp. 5474–5480.

Bahat, Y., Irani, M., 2016. Blind dehazing using internal patch recurrence. In: 2016 IEEE International Conference on Computational Photography. ICCP, IEEE, pp. 1–9.

Baig, M.H., Torresani, L., 2016. Coupled depth learning. In: IEEE Winter Conf. Appl. Comput. Vis.. WACV, pp. 1–10.

Berman, D., Avidan, S., et al., 2016. Non-local image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1674–1682.

Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.-M., Reid, I., 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In: Adv. Neur. in. (NeurIPS). pp. 35–45.

Cai, B., Xu, X., Jia, K., Qing, C., Tao, D., 2016. Dehazenet: An end-to-end system for single image haze removal. IEEE Trans. Image Process. 25 (11), 5187–5198.

Cao, J.-j., Wang, Q.-y., Chow, J.C., Watson, J.G., Tie, X.-x., Shen, Z.-x., Wang, P., An, Z.-s., 2012. Impacts of aerosol compositions on visibility impairment in Xi'an, China. Atmos. Environ. 59, 559–566.

Cao, Y., Wu, Z., Shen, C., 2017. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Trans. Circuits Syst. Video Technol. 28 (11), 3174–3182.

Casser, V., Pirk, S., Mahjourian, R., Angelova, A., 2019. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, pp. 8001–8008.

Chaabani, H., Kamoun, F., Bargaoui, H., Outay, F., et al., 2017. A neural network approach to visibility range estimation under foggy weather conditions. Proc. Comput. Sci. 113, 466–471.

Chan, C.K., Yao, X., 2008. Air pollution in mega cities in China. Atmosp. Environ. 42 (1), 1–42.

Chelani, A.B., 2019. Estimating PM2. 5 concentration from satellite derived aerosol optical depth and meteorological variables using a combination model. Atmosp. Pollut. Res. 10 (3), 847–857.

Chen, W., Fu, Z., Yang, D., Deng, J., 2016. Single-image depth perception in the wild. In: Adv. Neur. in. (NeurIPS). pp. 730–738.

Chen, J., Li, Z., Lv, M., Wang, Y., Wang, W., Zhang, Y., Wang, H., Yan, X., Sun, Y., Cribb, M., 2019a. Aerosol hygroscopic growth, contributing factors, and impact on haze events in a severely polluted region in northern China. Atmos. Chem. Phys. 19 (2), 1327–1342.

Chen, P.-Y., Liu, A.H., Liu, Y.-C., Wang, Y.-C.F., 2019b. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 2624–2632.

Cheng, X., Liu, G., Hedman, A., Wang, K., Li, H., 2018. Expressway visibility estimation based on image entropy and piecewise stationary time series analysis. CoRR abs/1804.04601 arXiv:1804.04601 URL http://arxiv.org/abs/1804.04601.

Choi, S., Min, D., Ham, B., Kim, Y., Oh, C., Sohn, K., 2015. Depth analogy: Data-driven approach for single image depth estimation using gradient samples. IEEE Trans. Image Process. 24 (12), 5953–5966.

C.S. Kumar, A., Bhandarkar, S.M., Prasad, M., 2018. Depthnet: A recurrent neural network architecture for monocular depth prediction. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 283–291.

Dietz, S.J., Kneringer, P., Mayr, G.J., Zeileis, A., 2019. Forecasting low-visibility procedure states with tree-based statistical methods. Pure Appl. Geophys. 176 (6), 2631–2644.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proc. IEEE Int. Conf. Comput. Vision. pp. 2650–2658.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. In: Adv. Neur. in. (NeurIPS). pp. 2366–2374.

Engin, D., Genç, A., Kemal Ekenel, H., 2018. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 825–833.

Fattal, R., 2008. Single image dehazing. ACM Trans. Graph. (TOG) 27 (3), 1–9.

Fattal, R., 2014. Dehazing using color-lines. ACM Trans. Graph. (TOG) 34 (1), 1–14.

Feng, T., Gu, D., 2019. Sganvo: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. IEEE Robot. Autom. Lett. 4 (4), 4431–4437.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 2002–2011.

Furukawa, R., Sagawa, R., Kawasaki, H., 2017. Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface. In: Proc. IEEE Int. Conf. Comput. Vision. pp. 4640–4648.

Garg, R., Bg, V.K., Carneiro, G., Reid, I., 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: Lect. Notes Comput. Sci.. Springer, pp. 740–756.

Gautam, S., Gandhi, T.K., Panigrahi, B.K., 2020. An improved air-light estimation scheme for single haze images using color constancy prior. IEEE Signal Process. Lett. 27, 1695–1699.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. Int. J. Robot. Res. 32 (11), 1231–1237.

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 270–279.

Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J., 2019. Digging into self-supervised monocular depth estimation. In: Proc. IEEE Int. Conf. Comput. Vision. pp. 3828–3838.

Grigorev, A., Jiang, F., Rho, S., Sori, W.J., Liu, S., Sai, S., 2017. Depth estimation from single monocular images using deep hybrid network. Multimedia Tools Appl. 76 (18), 18585–18604.

Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A., 2020. 3D packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2485–2494.

He, K., Sun, J., Tang, X., 2010. Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. 33 (12), 2341–2353.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 770–778.

Huang, X., Ding, A., Wang, Z., Ding, K., Gao, J., Chai, F., Fu, C., 2020. Amplified transboundary transport of haze by aerosol–boundary layer interaction in China. Nat. Geosci. 13 (6), 428–434.

Ji, D., Deng, Z., Sun, X., Ran, L., Xia, X., Fu, D., Song, Z., Wang, P., Wu, Y., Tian, P., et al., 2020. Estimation of PM 2.5 mass concentration from visibility. Adv. Atmosp. Sci. 37, 671–678.

Jia, S., Pei, X., Jing, X., Yao, D., 2021. Self-supervised 3D reconstruction and ego-motion estimation via on-board monocular video. IEEE Trans. Intell. Transp. Syst. 23 (7), 7557–7569.

Jia, S., Pei, X., Yang, Z., Tian, S., Yue, Y., 2020a. Novel hybrid neural network for dense depth estimation using on-board monocular images. Transp. Res. Rec. 2674 (12), 312–323.

Jia, S., Pei, X., Yao, W., Wong, S.C., 2022. Self-supervised depth estimation leveraging global perception and geometric smoothness. IEEE Trans. Intell. Transp. Syst. 24 (2), 1502–1517.

Jia, S., Wong, S.C., Wong, W., 2023. Uncertainty estimation of connected vehicle penetration rate. Transp. Sci. 57 (5), 1160–1176.

Jia, S., Yao, W., 2023. Joint learning of frequency and spatial domains for dense image prediction. ISPRS J. Photogramm. Remote Sens. 195, 14–28.

Jia, S., Yue, Y., Yang, Z., Pei, X., Wang, Y., 2020b. Travelling modes recognition via bayes neural network with bayes by backprop algorithm. In: CICTP 2020. pp. 3994–4004.

Kan, H., Chen, R., Tong, S., 2012. Ambient air pollution, climate change, and population health in China. Environ. Int. 42, 10–19.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Klingner, M., Termöhlen, J.-A., Mikolajczyk, J., Fingscheidt, T., 2020. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: European Conference on Computer Vision. Springer, pp. 582–600.

Kuznietsov, Y., Stuckler, J., Leibe, B., 2017. Semi-supervised deep learning for monocular depth map prediction. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 6647–6655.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. In: Proc. - Int. Conf. 3D Vis. (3DV). IEEE, pp. 239–248.

Li, J., Klein, R., Yao, A., 2017. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In: Proc. IEEE Int. Conf. Comput. Vision. pp. 3372–3380.

Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M., 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 1119–1127.

Li, R., Yang, Z., Pei, X., Yue, Y., Jia, S., Han, C., He, Z., 2023. A novel one-stage approach for pointwise transportation mode identification inspired by point cloud processing. Transp. Res. C 152, 104127.

Liu, F., Shen, C., Lin, G., 2015a. Deep convolutional neural fields for depth estimation from a single image. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 5162–5170.

Liu, F., Shen, C., Lin, G., Reid, I., 2015b. Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. Pattern Anal. Mach. Intell. 38 (10), 2024–2039.

Malm, W.C., Day, D.E., 2001. Estimates of aerosol species scattering characteristics as a function of relative humidity. Atmos. Environ. 35 (16), 2845–2860.

Mancini, M., Costante, G., Valigi, P., Ciarfuglia, T.A., Delmerico, J., Scaramuzza, D., 2017. Toward domain independence for learning-based monocular depth estimation. IEEE Robot. Autom. Lett. 2 (3), 1778–1785.

Mousavian, A., Pirsiavash, H., Košecká, J., 2016. Joint semantic segmentation and depth estimation with deep convolutional networks. In: Proc. - Int. Conf. 3D Vis. (3DV). IEEE, pp. 611–619.

Palvanov, A., Im Cho, Y., 2018. Dhcnn for visibility estimation in foggy weather conditions. In: 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems. ISIS, IEEE, pp. 240–243.

Pandolfi, M., Alados-Arboledas, L., Alastuey, A., Andrade, M., Angelov, C., Artiñano, B., Backman, J., Baltensperger, U., Bonasoni, P., Bukowiecki, N., et al., 2018. A European aerosol phenomenology–6: scattering properties of atmospheric aerosol particles from 28 ACTRIS sites. Atmos. Chem. Phys. 18 (11), 7877–7911.

Polewski, P., Shelton, J., Yao, W., Heurich, M., 2021. Instance segmentation of fallen trees in aerial color infrared imagery using active multi-contour evolution with fully convolutional network-based intensity priors. ISPRS J. Photogramm. Remote Sens. 178, 297–313. http://dx.doi.org/10.1016/j.isprsjprs.2021.06.016.

Pomerleau, D., 1997. Visibility estimation from a moving vehicle using the RALPH vision system. In: Proceedings of Conference on Intelligent Transportation Systems. IEEE, pp. 906–911.

Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J., 2019. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 12240–12249.

Renhe, Z., Li, Q., Zhang, R., 2014. Meteorological conditions for the persistent severe fog and haze event over eastern China in January 2013. Sci. China Earth Sci. 57, 26–35.

Saxena, A., Chung, S.H., Ng, A.Y., 2008. 3-d depth reconstruction from a single still image. Int. J. Comput. Vis. 76 (1), 53–69.

Shelton, J., Polewski, P., Yao, W., 2021. U-Net for learning and inference of dense representation of multiple air pollutants from satellite imagery. In: Proceedings of the 10th International Conference on Climate Informatics. In: CI2020, Association for Computing Machinery, New York, NY, USA, pp. 128–133. http://dx.doi.org/10.1145/3429309.3429328.

Shu, C., Yu, K., Duan, Z., Yang, K., 2020. Feature-metric loss for self-supervised learning of depth and egomotion. In: European Conference on Computer Vision. Springer, pp. 572–588.

Sun, Y., Zeng, Q., Geng, B., Lin, X., Sude, B., Chen, L., 2019. Deep learning architecture for estimating hourly ground-level PM 2.5 using satellite remote sensing. IEEE Geosci. Remote Sens. Lett. 16 (9), 1343–1347.

Tan, R.T., 2008. Visibility in bad weather from a single image. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

Tananaev, D., Zhou, H., Ummenhofer, B., Brox, T., 2018. Temporally consistent depth estimation in videos with recurrent architectures. In: Proc. Eur. Conf. Comput. Vis..

Tao, J., Zhang, Z., Wu, Y., Zhang, L., Wu, Z., Cheng, P., Li, M., Chen, L., Zhang, R., Cao, J., 2019. Impact of particle number and mass size distributions of major chemical components on particle mass scattering efficiency in urban Guangzhou in southern China. Atmos. Chem. Phys. 19 (13), 8471–8490.

Van Donkelaar, A., Martin, R.V., Park, R.J., 2006. Estimating ground-level PM2. 5 using aerosol optical depth determined from satellite remote sensing. J. Geophys. Res.: Atmos. 111 (D21).

Wang, Y., Chau, L.-P., Ma, X., 2019b. Airlight estimation based on distant region segmentation. In: 2019 IEEE International Symposium on Circuits and Systems. ISCAS, IEEE, pp. 1–5.

Wang, K., Dickinson, R.E., Liang, S., 2009. Clear sky visibility has decreased over land globally from 1973 to 2007. Science 323 (5920), 1468–1470.

Wang, D., Jia, S., Pei, X., Han, C., Yao, D., Liu, D., 2023. DERNet: driver emotion recognition using onboard camera. IEEE Intell. Transp. Syst. Mag. 2–17.

Wang, C., Miguel Buenaposada, J., Zhu, R., Lucey, S., 2018. Learning depth from monocular videos using direct methods. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 2022–2030.

Wang, R., Pizer, S.M., Frahm, J.-M., 2019a. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 5555–5564.

Wang, P., Yao, W., 2022. A new weakly supervised approach for ALS point cloud semantic segmentation. ISPRS J. Photogramm. Remote Sens. 188, 237–254. http://dx.doi.org/10.1016/j.isprsjprs.2022.04.016, URL https://www.sciencedirect.com/science/article/pii/S0924271622001198.

Wang, J.-L., Zhang, Y.-h., Shao, M., Liu, X.-l., Zeng, L.-m., Cheng, C.-l., Xu, X.-f., 2006. Quantitative relationship between visibility and mass concentration of PM2. 5 in Beijing. J. Environ. Sci. 18 (3), 475–481.

Watson, J.G., 2002. Visibility: Science and regulation. J. Air Waste Manage. Assoc. 52 (6), 628–713.

Weng, P., Jia, S., Pei, X., Yue, Y., 2021. Bayes neural network with a novel pictorial feature for transportation mode recognition based on GPS trajectories. In: CICTP 2021. pp. 1635–1645.

Xu, W., Kuang, Y., Bian, Y., Liu, L., Li, F., Wang, Y., Xue, B., Luo, B., Huang, S., Yuan, B., et al., 2020. Current challenges in visibility improvement in southern China. Environ. Sci. Technol. Lett. 7 (6), 395–401.

Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N., 2017. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 5354–5362.

Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E., 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 3917–3925.

Yang, D., Sun, J., 2018. Proximal dehaze-net: A prior learning-based deep network for single image dehazing. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 702–717.

Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R., 2017. Unsupervised learning of geometry with edge-aware depth-normal consistency. arXiv preprint arXiv:1711.03665.

Yin, Z., Shi, J., 2018. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 1983–1992.

You, J., Jia, S., Pei, X., Yao, D., 2022. DMRVisNet: Deep multihead regression network for pixel-wise visibility estimation under foggy weather. IEEE Trans. Intell. Transp. Syst. 23 (11), 22354–22366.

Zhan, H., Garg, R., Saroj Weerasekera, C., Li, K., Agarwal, H., Reid, I., 2018. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 340–349.

Zhang, H., Patel, V.M., 2018. Densely connected pyramid dehazing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3194–3203.

Zhao, C., Yu, Y., Kuang, Y., Tao, J., Zhao, G., 2019. Recent progress of aerosol light-scattering enhancement factor studies in China. Adv. Atmosp. Sci. 36, 1015–1026.

Zheng, C., Zhao, C., Zhu, Y., Wang, Y., Shi, X., Wu, X., Chen, T., Wu, F., Qiu, Y., 2017. Analysis of influential factors for the relationship between PM 2.5 and AOD in Beijing. Atmos. Chem. Phys. 17 (21), 13473–13489.

Zhou, T., Brown, M., Snavely, N., Lowe, D.G., 2017. Unsupervised learning of depth and ego-motion from video. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 1851–1858.

Zhou, J., Wang, Y., Qin, K., Zeng, W., 2019. Unsupervised high-resolution depth learning from videos with dual networks. In: Proc. IEEE Int. Conf. Comput. Vision. pp. 6872–6881.

Zhu, Q., Mai, J., Shao, L., 2015. A fast single image haze removal algorithm using color attenuation prior. IEEE Trans. Image Process. 24 (11), 3522–3533.

Zoran, D., Isola, P., Krishnan, D., Freeman, W.T., 2015. Learning ordinal relationships for mid-level vision. In: Proc. IEEE Int. Conf. Comput. Vision. pp. 388–396.