This ACCV 2020 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2020
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Dehazing Cost Volume for Deep Multi-view Stereo in Scattering Media

Yuki Fujimura[0000−0002−7225−8452], Motoharu Sonogashira[0000−0001−7429−4011], and Masaaki Iiyama[0000−0002−7715−3078]

Kyoto University, Kyoto-shi, 606–8501 Japan
{fujimura,sonogashira,iiyama}@mm.media.kyoto-u.ac.jp

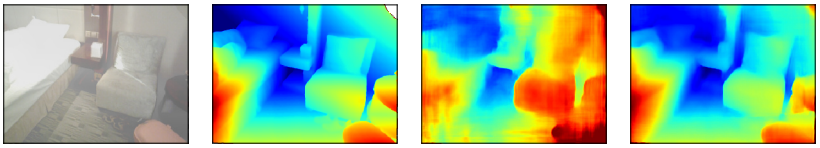
Abstract. We propose a learning-based multi-view stereo (MVS) method in scattering media such as fog or smoke with a novel cost volume, called the dehazing cost volume. An image captured in scattering media degrades due to light scattering and attenuation caused by suspended particles. This degradation depends on scene depth; thus it is difficult for MVS to evaluate photometric consistency because the depth is unknown before three-dimensional reconstruction. Our dehazing cost volume can solve this chicken-and-egg problem of depth and scattering estimation by computing the scattering effect using swept planes in the cost volume. Experimental results on synthesized hazy images indicate the effectiveness of our dehazing cost volume against the ordinary cost volume regarding scattering media. We also demonstrated the applicability of our dehazing cost volume to real foggy scenes.

1 Introduction

Three-dimensional (3D) reconstruction from 2D images is important in computer vision. However, images captured in scattering media, such as fog or smoke, degrade due to light scattering and attenuation caused by suspended particles. For example, Fig. 1(a) shows a synthesized hazy image, the contrast of which is reduced due to light scattering. Traditional 3D reconstruction techniques that exploit observed pixel intensity cannot work in such environments.

We propose a learning-based multi-view stereo (MVS) method in scattering media. MVS [1] is a method for reconstructing the 3D geometry of a scene from multiple images. Recently, learning-based MVS methods have been proposed and provided highly accurate results [2–4]. The proposed method is based on MVDepthNet [5], which is one such MVS method.

MVDepthNet estimates scene depth by taking a cost volume as input for the network. The cost volume is based on a plane sweep volume [6], i.e., it is constructed by sweeping a fronto-parallel plane to a camera in the scene and evaluates the photometric consistency between multiple cameras under the assumptions that the scene lies on each plane. As described above, however, an image captured in scattering media degrades; thus, using the ordinary cost volume leads to undesirable results, as shown in Fig. 1(c).



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

Fig. 1. (a) synthesized hazy image due to scattering medium. (b) ground truth depth. (c) output depth of fine-tuned MVDepthNet [5] with ordinary cost volume. (d) output depth of network with our dehazing cost volume.

To address this problem, we propose a novel cost volume for scattering media, called *the dehazing cost volume*. In scattering media, light bouncing off a scene is attenuated exponentially relative to the depth. On the other hand, scattered light observed with a camera increases with depth. This means that the degradation due to a scattering medium depends on the scene depth. Our dehazing cost volume can restore images with such depth-dependent degradation and compute the effective cost of photometric consistency simultaneously. It enables robust 3D reconstruction in scattering media, as shown in Fig. 1(d).

In summary, the primary contribution of this paper is to design a novel cost volume for scattering media, which *avoids the chicken-and-egg problem of depth and scattering estimation by computing degradation with the depth of each swept plane in the cost volume.* Accordingly, our dehazing cost volume will accelerate the real-time applicability of 3D reconstruction in scattering media.

2 Related work

2.1 Multi-view stereo

As mentioned above, MVS [1] is a method of reconstructing 3D geometry using multiple cameras. In general, it exploits the dense pixel correspondence between multiple images for 3D reconstruction. The correspondence is referred to as photometric consistency and computed on the basis of the similarity measure of pixel intensity. One of the difficulties in the computation of photometric consistency is occlusion, i.e., the surface of a target object is occluded from certain cameras. This leads to incorrect correspondence and inaccurate 3D reconstruction. To address this problem, methods have been proposed for simultaneous view selection to compute effective photometric consistency and 3D reconstruction with MVS, achieving highly accurate 3D reconstruction [7, 8].

Along with the above issue, there are many cases in which it is difficult to obtain accurate 3D geometry with traditional MVS methods. A textureless surface and an object with a view-dependent reflectance property, such as specular reflection, are typical cases. Learning-based MVS methods have recently been used to learn semantic information on large-scale training data and enable robust 3D reconstruction in such scenes.

Learning-based MVS methods often construct a cost volume to constrain 3D geometry between multiple cameras. For example, Wang and Shen [5] proposed MVDepthNet, which constructs a cost volume from multi-view images setting one of the images as a reference image. It can take an arbitrary number of input images to construct the cost volume. The convolutional neural network takes the reference image and cost volume as input then estimates the depth map of the reference camera. DeepMVS proposed by Huang et al. [3] first constructs a plane sweep volume, then the patch matching network is applied to the reference image and each slice of the volume to extract features to measure the correspondence, which is followed by feature aggregation networks and depth refinement with a fully connected conditional random field. Yao et al. [2] and Im et al. [4] respectively proposed MVSNet and DPSNet, in which input images are first passed through the networks to extract features , then the features are warped instead of constructing the cost volume in the image space. Our proposed method is based on MVDepthNet [5], which is the simplest and light-weight method, and we extended the ordinary cost volume for scattering media.

2.2 Dehazing

In scattering media, a captured image degrades due to light scattering and attenuation. To enhance the quality of an image captured in scattering media, dehazing and defogging methods have been proposed [9–12]. These studies introduced the priors of latent clear images to solve the ill-posed nature of the problem. For example, He et al. [9] proposed a dark channel prior with which a clear image having a dark pixel in a local image patch is assumed. Berman et al. [12] proposed a haze-line prior with which the same intensity pixels of the latent clear image forms a line in RGB space. Many learning-based methods using neural networks have also been proposed recently [13–18]. Dehazing can improve computer vision tasks in scattering media such as object detection [19].

2.3 3D reconstruction in scattering media

Our goal is to reconstruct 3D geometry directly from degraded images by scattering media instead of recovering the latent clear images. There has been research focusing on the same problem as in our study. For example, Narasimhan et al. [20] proposed a 3D reconstruction method using structured light in scattering media. Photometric stereo methods have also been proposed for scattering media [21–23]. However, these methods require active light sources, which limits real-world applicability. Instead of using an ordinary camera, Heide et al. [24] and Satat et al. [25] respectively used a time-of-flight camera and single photon avalanche diode for scattering media. Wang et al. [26] combined a line sensor and line laser to generate a programmable light curtain that can suppress the backscatter effect. However, the use of these methods is hindered due to the requirement of expensive sensors or special hardware settings.

The proposed method is based on stereo 3D reconstruction requiring neither active light sources nor special hardware settings. Caraffa et al. [27] proposed

a binocular stereo method in scattering media. With this method, image enhancement and stereo reconstruction are simultaneously modeled on the basis of a Markov random field. Song et al. [28] proposed a learning-based binocular stereo method in scattering media, where dehazing and stereo reconstruction are trained as multi-task learning. The features from the networks of each task are simply concatenated at the intermediate layer. The most related method to ours is the MVS method proposed by Li et al. [29]. They modeled dehazing and MVS simultaneously, and the output depth was regularized using an ordering constraint, which was based on a transmission map that was the output of dehazing with Laplacian smoothing. With all these methods, homogeneous scattering media is assumed; thus, we followed the same assumption. It is left open to apply these methods to inhomogeneous media.

These previous studies [27, 29] designed photometric consistency measures considering the scattering effect. However, this requires scene depth because degradation due to scattering media depends on this depth. Thus, they relied on iterative implementation of an MVS method and dehazing, which leads to large computation cost. In contrast, our dehazing cost volume can solve this chicken-and-egg problem by computing the scattering effect in the cost volume. The scene depth is then estimated effectively by taking the cost volume as input for a convolutional neural network, making fast inference possible.

3 Multi-view stereo in scattering media

In this section, we describe MVS in scattering media with our dehazing cost volume. First, we introduce an image formation model in scattering media then give an overview of the proposed method, followed by a discussion on an ordinary cost volume and our dehazing cost volume.

3.1 Image formation model

We use an atmospheric scattering model [30] for image observation in scattering media. This model is used for many dehazing methods and describes the degradation of an observed image in scattering media in daylight. Let an RGB value at the pixel (u, v) of a degraded image captured in scattering media and its latent clear image be $I(u, v) \in \mathbb{R}^3$ and $J(u, v) \in \mathbb{R}^3$, respectively. We assume that the pixel value of each color channel is within 0 and 1. The observation process of this model is given as

$$I(u, v) = J(u, v)e^{-\beta z(u,v)} + A(1 - e^{-\beta z(u,v)}), \tag{1}$$

where $z(u, v) \in \mathbb{R}$ is the depth at pixel (u, v), $\beta \in \mathbb{R}$ is a scattering coefficient that represents the density of a medium, and $A \in \mathbb{R}$ is global airlight. The first term is a component that describes reflected light in a scene. This reflected component becomes attenuated exponentially with respect to the scene depth. The second term is a scattering component, which consists of scattered light that arrives at a

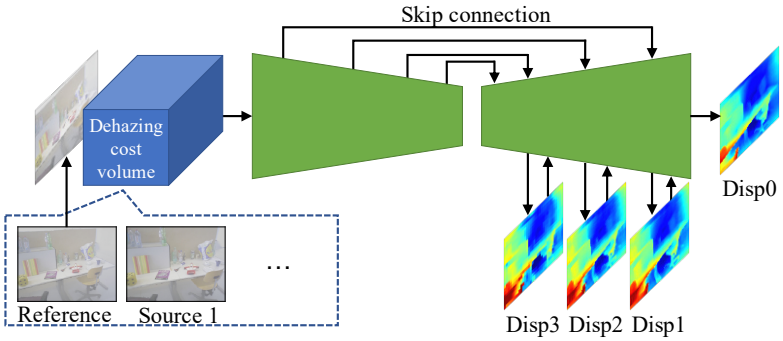


Fig. 2. Input of network is reference image captured in scattering medium and our dehazing cost volume. Our dehazing cost volume is constructed from reference image and source images. Network architecture of our method is same as that of MVDepthNet [5], which has encoder-decoder with skip connections. Output of network is disparity maps (inverse depth maps) at different resolutions.

camera without reflecting on objects. In contrast to the reflected component, this component increases with depth. Therefore, image degradation due to scattering media depends on the scene depth.

In the context of image restoration, we aim to estimate unknown parameters J, z, scattering coefficient β, and airlight A from an observed image I, and the estimation of all these parameters at the same time is an ill-posed problem. Previous studies developed methods for estimating A from a single image [9, 31]. In addition, Li et al. [29] estimated β under a multi-view setting at a structure-from-motion (SfM) step. This is the same problem setting as in our study. In the rest of this paper, therefore, we assume that A and β have already been estimated unless otherwise noted. At the end of Section 4, we discuss the the effect of the estimation error of these parameters.

3.2 Overview

MVS methods are roughly categorized by output representation, e.g., point-cloud, volume, or mesh-based reconstruction. The proposed method is formulated as a depth map estimation, i.e., given multiple cameras, we estimate a depth map for one of the cameras. In this paper, a target camera to estimate a depth map is referred to as a reference camera r and the other cameras are referred to as source cameras $s \in \{1, \cdots, S\}$, and images captured with these cameras are denoted as a reference image I_r and source images I_s. We assume that the camera parameters are calibrated beforehand.

An overview of the proposed method is shown in Fig. 2. Our dehazing cost volume is constructed from a hazy reference image and source images captured in a scattering medium. The network takes the reference image and our dehazing cost volume as input then outputs a disparity map (inverse depth map) of the

reference image. The network architecture is the same as that of MVDepthNet [5], while the ordinary cost volume used in MVDepthNet is replaced with our dehazing cost volume for scattering media.

3.3 Dehazing cost volume

In this section, we explain our dehazing cost volume, which is taken as input to the network. The dehazing cost volume enables effective computation of photometric consistency in scattering media.

Before explaining our dehazing cost volume, we show the computation of the ordinary cost volume in Fig. 3(a). We first sample the 3D space in the reference camera coordinate by sweeping a fronto-parallel plane. We then back-project source images onto each sampled plane. Finally, we take the residual between the reference image and each warped source image, which corresponds to the cost of photometric consistency on the hypothesis that the scene exists on the plane. Let the image size be $W \times H$ and number of sampled depths be N. We denote the cost volume by $\mathcal{V} : \{1, \cdots, W\} \times \{1, \cdots, H\} \times \{1, \cdots, N\} \to \mathbb{R}$, and each element of the cost volume is given as follows:

$$\mathcal{V}(u, v, i) = \frac{1}{S} \sum_{s} \| I_r(u, v) - I_s(\pi_{r \to s}(u, v; z_i)) \|_1, \tag{2}$$

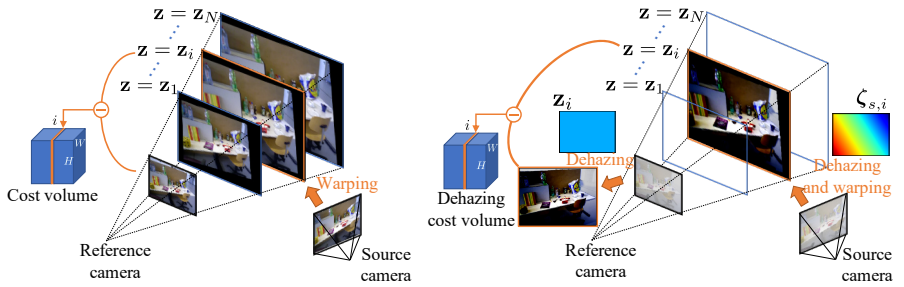
where z_i is the depth value of the i-th plane. The operator $\pi_{r \to s} : \mathbb{R}^2 \to \mathbb{R}^2$ projects the camera pixel (u, v) of the reference camera r onto the source image I_s with the given depth, which is defined as follows:

$$\begin{bmatrix} \pi_{r \to s}(u, v; z) \\ 1 \end{bmatrix} \sim z \mathbf{K}_s \mathbf{R}_{r \to s} \mathbf{K}_r^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} + \mathbf{K}_s \mathbf{t}_{r \to s}, \tag{3}$$

where $\mathbf{K}_r$ and $\mathbf{K}_s$ are the intrinsic parameters of r and the source camera s, and $\mathbf{R}_{r \to s}$ and $\mathbf{t}_{r \to s}$ are a rotation matrix and translation vector from r to s. The cost volume evaluates the photometric consistency of each pixel with respect to the sampled depth; thus, the element of the cost volume with correct depth ideally becomes zero.

An observed image captured in scattering media degrades in the manner described in Eq. (1), and the ordinary cost volume defined in Eq. (2) leads to undesirable results. In contrast, our dehazing cost volume dehazes the image and computes photometric consistency cost simultaneously. As described in Section 3.1, degradation due to scattering media depends on scene depth; thus, our dehazing cost volume restores degraded images using the depth of a swept plane.

Figure 3(b) shows the computation of our dehazing cost volume. A reference image is dehazed directly using the depth of a swept plane. A source image is dehazed using the swept plane from a source camera view, then the dehazed source image is warped to the reference camera coordinate. Similar to the ordinary cost



(a) Cost volume (b) Dehazing cost volume

Fig. 3. (a) Ordinary cost volume is constructed by sweeping fronto-parallel plane in reference-camera coordinate. Cost of photometric consistency is simply computed as residual between reference image and warped source image on each swept plane $\mathbf{z} = \mathbf{z}_i$. (b) In our dehazing cost volume, reference image is dehazed using sampled depth, $\mathbf{z}_i$, which is constant over all pixels. Source image is dehazed using depth of swept plane from source-camera view, then dehazed source image is back-projected onto plane. Cost is computed by taking residual between both dehazed images.

volume, we define our dehazing cost volume as $\mathcal{D} : \{1, \cdots, W\} \times \{1, \cdots, H\} \times \{1, \cdots, N\} \rightarrow \mathbb{R}$, and each element of our dehazing cost volume is given as

$$\mathcal{D}(u, v, i) = \frac{1}{S} \sum_{s} \| J_r(u, v; z_i) - J_s(\pi_{r \rightarrow s}(u, v; z_i)) \|_1, \tag{4}$$

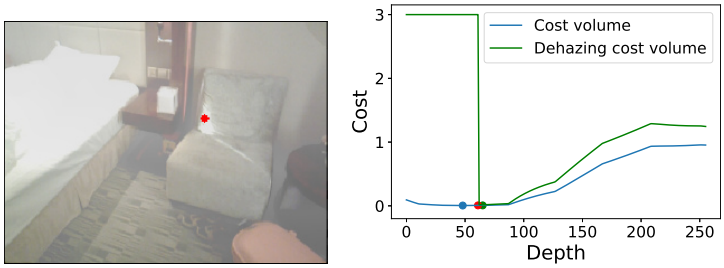
where $J_r(u, v; z_i)$ and $J_s(\pi_{t \rightarrow s}(u, v; z_i))$ are dehazed reference and source images, and from Eq. (1), they are computed as follows:

$$J_r(u, v; z_i) = \frac{I_r(u, v) - A}{e^{-\beta z_i}} + A, \tag{5}$$

$$J_s(\pi_{r \rightarrow s}(u, v; z_i)) = \frac{I_s(\pi_{r \rightarrow s}(u, v; z_i)) - A}{e^{-\beta \zeta_{s,i}(\pi_{r \rightarrow s}(u, v; z_i))}} + A. \tag{6}$$

As shown in Fig. 3(b), the reference image is dehazed using the swept plane with depth z_i, whose depth map is denoted as $\mathbf{z}_i$. On the other hand, the source image is dehazed using $\boldsymbol{\zeta}_{s,i}$, which is a depth map of the swept plane from the source camera view. The depth $\zeta_{s,i}(\pi_{r \rightarrow s}(u, v; z_i))$ is used for the cost computation of the pixel (u, v) of the reference camera because the pixel $\pi_{r \rightarrow s}(u, v; z_i)$ on the source camera corresponds to pixel (u, v) of the reference camera. Our dehazing cost volume exploits the dehazed images with much more contrast than the degraded ones; thus, the computed cost is robust even in scattering media. According to this definition of our dehazing cost volume, the photometric consistency between the latent clear images is preserved.

Our dehazing cost volume computes photometric consistency with dehazed images in the cost volume. This is similar to the previous methods [27, 29] that compute photometric consistency considering scattering effect. However, this is



(a) (b)

Fig. 4. Visualization of our dehazing cost volume. (b) shows values of ordinary cost volume and our dehazing cost volume at red point in (a), respectively.

a chicken-and-egg problem because the effect of scattering media depends on scene depth, and they rely on iterative implementation of MVS and dehazing to compute the scattering effect. Our method, on the other hand, can compute the scattering effect using a depth hypothesis of a swept plane without an explicit scene depth, which can eliminate the iterative optimization.

Our dehazing cost volume restores an image using all depth hypotheses; thus, image dehazing with depth that greatly differs from the correct scene depth results in an unexpected image. The extreme case is when a dehazed image has negative values at certain pixels. This includes the possibility that a computed cost using Eq. (4) becomes very large. To avoid such cases, we revise the definition of our dehazing cost volume as follows:

$$\mathcal{D}(u,v,i) = \frac{1}{S}\sum_{s} \begin{cases} \|J_r(u,v;z_i) - J_s(\pi_{r\to s}(u,v;z_i))\|_1 \\ \quad if\ 0 \le J_r^c(u,v;z_i) \le 1\ and \\ \quad 0 \le J_s^c(\pi_{r\to s}(u,v;z_i)) \le 1\ c \in \{r,g,b\} \\ \gamma\ otherwise, \end{cases} \tag{7}$$

where $J_r^c(u,v;z_i)$ and $J_s^c(\pi_{r\to s}(u,v;z_i))$ are the pixel values of the channel $c \in \{r,g,b\}$ of the reconstructed clear images. A constant γ is a parameter that is set as a penalty cost when the dehazed result is not contained in the domain of definition. This makes the training of the network stable because our dehazing cost volume is upper bounded by γ. We can also reduce the search space of depth by explicitly giving the penalty cost. In this study, we set $\gamma = 3$, which is the maximum value of the ordinary cost volume defined in Eq. (2) when the pixel value of each color channel is within 0 and 1.

Figure 4 (b) visualizes the ordinary cost volume and our dehazing cost volume at the red point in (a). Each point in (b) indicates a minimum, and the red point in (b) indicates ground truth depth. The curve of the cost volume is smoother than our dehazing cost volume due to the degradation of the image contrast, which leads to a depth error. In addition, our dehazing cost volume can reduce the search space with the dehazing constraint γ on the left part in (b).

3.4 Network architecture and loss function

As shown in Fig. 2, the network takes a reference image and our dehazing cost volume as input. To compute our dehazing cost volume, we should predetermine the target 3D space for scene reconstruction and number of depth hypotheses for plane sweep. We uniformly sample the depth on the disparity space between 0.02 and 2 and set the number of samples to $N = 256$. The network architecture is the same as that of MVDepthNet [5], which has an encoder-decoder architecture with skip connections. The network outputs disparity maps at different resolutions. The training loss is defined as the sum of L1 loss between these estimated disparity maps and the ground truth disparity map. For more details, please refer to [5].

4 Experiments

In this study, we used MVDepthNet [5] as a baseline method. As mentioned previously, the ordinary cost volume is replaced with our dehazing cost volume in the proposed method, so we can directly evaluate the effect of our dehazing cost volume by comparing our method with this baseline method. We also compared the proposed method to simple sequential methods of dehazing and 3D reconstruction using the baseline method. DPSNet [4], whose architecture is more complicated such as a multi-scale feature extractor, 3D convolutions, and a cost aggregation module, was also trained on hazy images for further comparison. In addition to the experiments with synthetic data, we give an example of applying the proposed method to actual foggy scenes. At the end of this section, we discuss the effect of the estimation errors of scattering parameters.

4.1 Dataset

We used the DeMoN dataset [32] for training. This dataset consists of the SUN3D [33], RGB-D SLAM [34], and MVS datasets [35], which have sequences of real images. The DeMoN dataset also has the Scenes11 dataset [36, 32], which consists of synthetic images. Each image sequence in the DeMoN dataset includes RGB images, depth maps, and camera parameters. In the real-image datasets, most of the depth maps have missing regions due to sensor sensibility. As we discuss later, we synthesized hazy images from the DeMoN dataset for training the proposed method, which requires a dense depth map without missing regions. Therefore, we first trained the baseline method using clear images then compensated for the missing regions of each depth map with the output depth of the baseline method. To suppress boundary discontinuities and sensor noise around missing regions, we applied a median filter after inpainting each depth map. For the MVS dataset, which has larger noise than other datasets, we reduced the noise simply by thresholding before inpainting. Note that the training loss was computed using only pixels that originally had valid depth values. We generated 419,046 and 8,842 samples for training and test data, respectively. Each sample contained one reference image and one source image. All images were resized to 256×192.

We synthesized a hazy-image dataset for training the proposed method from clear images. The procedure of generating a hazy image is based on Eq. (1). For A, we randomly sampled $A \in [0.7, 1.0]$ for each data sample. For β, we randomly sampled $\beta \in [0.4, 0.8], [0.4, 0.8], [0.05, 0.15]$ for the SUN3D, RGB-D SLAM, and Scenes11 datasets, respectively. We found that for the MVS dataset, it was difficult to determine the same sampling range of β for all images because it contains various scenes with different depth scales. Therefore, we determined the sampling range of β for each sample of the MVS dataset as follows: first, we set the range of a transmission map $e^{-\beta z}$ to $e^{-\beta z} \in [0.2, 0.4]$ for all samples then computed the median of a depth map z_{med} for each sample. Finally, we determined the β range for each sample as $\beta \in [-\log(0.4)/z_{med}, -\log(0.2)/z_{med}]$.

Similar to [5], we adopted data augmentation to enable the network to reconstruct a wide depth range. The depth of each sample was scaled by a factor between 0.5 and 1.5 together with the translation vector of the camera. Note that when training the proposed method, β should also be scaled by the inverse of the scale factor.

4.2 Training details

All networks were implemented in PyTorch. The training was done on a NVIDIA V100 GPU with 32-GB memory. The size of a minibatch was 32 for all training.

We first trained the baseline method from scratch on the clear image dataset. We used Adam with a learning rate of 1.0×10^{-4}. After the initial 100K iterations, the learning rate was reduced by 20% after every 20K iterations. The method was trained for about 260K iterations in total.

We then fine-tuned the baseline method on hazy images and trained the proposed method with our dehazing cost volume. The parameters of both methods were initialized by that of the trained baseline method on clear images. The initial learning rate was set to 1.0×10^{-4} and reduced by 20% after every 20K iterations. The fine-tuned baseline and proposed methods were trained for about 196K and 144K iterations, respectively.

We also trained the dehazing methods, AOD-Net [19] and FFA-Net [18], and DPSNet [4] on our hazy image dataset for comparison. The dehazing networks were followed by the baseline method trained on clear images for depth estimation. DPSNet was trained with the same loss function and learning schedule as in the original paper [4].

4.3 Results

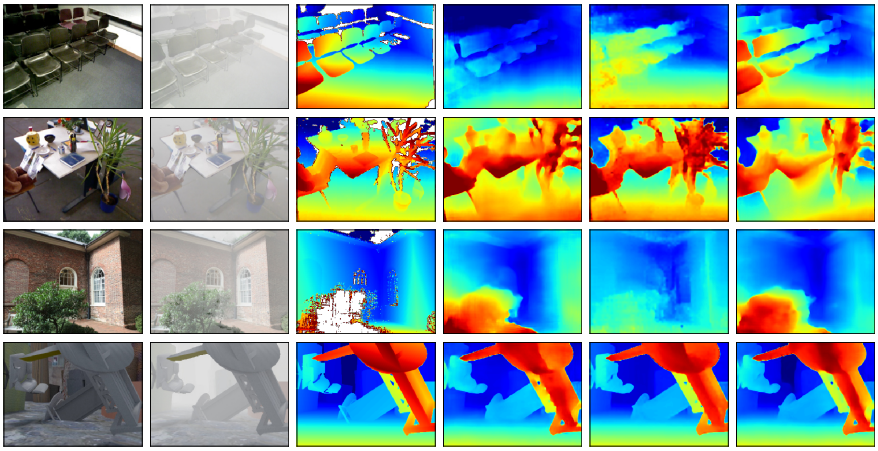
Table 1 shows the quantitative evaluation of each method. We used four evaluation metrics following [5]: L1-rel is the mean of the relative L1 error between the ground truth depth and estimated depth, L1-inv is the mean of the L1 error between ground truth inverse depth and estimated inverse depth, sc-inv is the scale-invariant error of depth proposed by Eigen et al. [37], and correctly estimated depth percentage (C.P.) [38] is the percentage of pixels whose relative L1 error is within 10%. The red and blue values are the best and second-best,

Table 1. Quantitative results. We compared proposed method to baseline method [5] fine-tuned on hazy images, simple sequential methods of dehazing [19, 18] and depth estimation with baseline method, and DPSNet [4] trained on hazy images. Red and blue values are best and second-best, respectively.

Dataset	Method	L1-rel	L1-inv	sc-inv	C.P. (%)
SUN3D	AOD-Net [19] + Baseline [5]	0.249	0.132	0.250	47.8
	FFA-Net [18] + Baseline [5]	0.180	0.111	0.211	55.5
	Fine-tuned [5]	0.155	0.093	0.184	60.3
	DPSNet [4]	0.145	0.082	0.183	64.7
	Proposed	0.100	0.058	0.161	79.0
RGB-D SLAM	AOD-Net [19] + Baseline [5]	0.205	0.127	0.315	58.9
	FFA-Net [18] + Baseline [5]	0.179	0.114	0.288	65.0
	Fine-tuned [5]	0.157	0.091	0.254	70.7
	DPSNet [4]	0.152	0.090	0.234	71.6
	Proposed	0.162	0.089	0.231	68.8
MVS	AOD-Net [19] + Baseline [5]	0.323	0.123	0.309	51.9
	FFA-Net [18] + Baseline [5]	0.215	0.112	0.288	55.6
	Fine-tuned [5]	0.184	0.100	0.241	57.1
	DPSNet [4]	0.191	0.088	0.239	67.9
	Proposed	0.160	0.091	0.222	58.1
Scenes11	AOD-Net [19] + Baseline [5]	0.330	0.036	0.539	52.3
	FFA-Net [18] + Baseline [5]	0.377	0.041	0.600	51.3
	Fine-tuned [5]	0.151	0.022	0.279	64.0
	DPSNet [4]	0.105	0.018	0.381	81.8
	Proposed	0.134	0.019	0.216	72.3

respectively. The proposed method was compared to the baseline method [5] fine-tuned on hazy images, the sequential method of dehazing [19] and baseline method [5], and DPSNet [4] trained on hazy images. In most evaluation metrics, the proposed method outperformed the fine-tuned baseline method, demonstrating the effectiveness of our dehazing cost volume. For the RGB-D SLAM dataset, the fine-tuned baseline method was comparable to the proposed method. This is because many scenes in the RGB-D SLAM dataset are close to a camera. In this case, the degradation of an observed image is small and exists uniformly in the image, which has little effect on photometric consistency. The proposed method also performed better than the sequential methods of dehazing [19, 18] and baseline method [5]. Therefore, we can see that the simultaneous modeling of dehazing and 3D reconstruction based on our dehazing cost volume is effective. DPSNet [4] first extracts feature maps from input images, and then constructs a cost volume in the feature space. Thus, the feature extractor might be able to deal with image degradation caused by light scattering. Nevertheless, our dehazing cost volume allows considering image degradation with a simple network architecture.

The output depth of each method is shown in Fig. 5. From top to bottom, each row shows the results of the input images in the SUN3D, RGB-D SLAM, MVS, and Scenes11 datasets, respectively. DPSNet failed to construct correspon-



(a) Clear image (b) Hazy input (c) Ground truth (d) Fine-tuned [5] (e) DPSNet [4] (f) Proposed

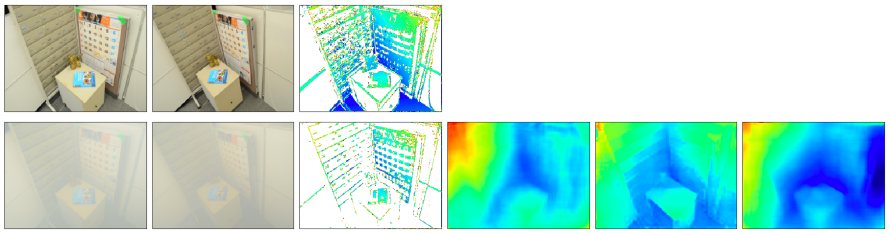
Fig. 5. Qualitative results. (a) clear image, (b) hazy input, (c) ground-truth depth, (d) output of fine-tuned baseline [5], (e) output of DPSNet [4], and (f) output of proposed method. From top to bottom, each row shows results of input images in SUN3D, RGB-D SLAM, MVS, and Scenes11 datasets, respectively. Values below each estimated depth represent error values (L1-rel/L1-inv/sc-inv/C.P.).

dence in some scenes, although it has the multi-scale feature extractor. Note that the results from the Scenes11 dataset indicate that the proposed method can reconstruct the 3D geometry of a distant scene where the image is heavily degraded due to scattering media.

4.4 Experiments with actual data

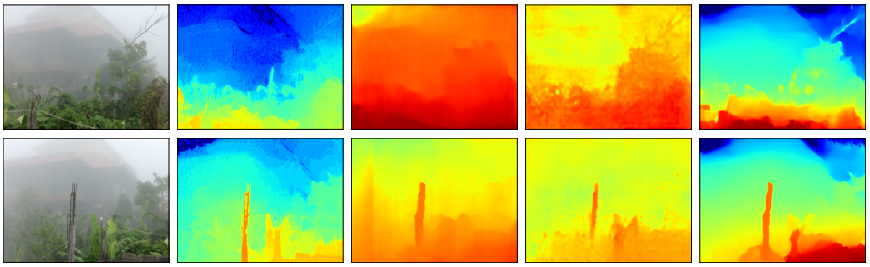
We applied the proposed method to actual scenes including scattering media. The captured images are shown in Figs. 6(a)(b). We generated fog artificially with a fog generator. Differing from the synthetic data, A and β were unknown. We applied a previous method [31] to both the reference and source images to estimate A as pre-processing. We then applied COLMAP [39, 8] to estimate extrinsic parameters and an initial depth map, which was very sparse due to image degradation, as shown in Fig 6(c). This sparse depth was used for the estimating β in a similar manner to [29]. The results of depth estimation are shown in Figs. 6(d)–(f). The proposed method also estimated depth effectively in these actual hazy scenes. DPSNet estimated edge-preserved depth, which was achieved due to its cost aggregation module.

We also applied the proposed method to actual outdoor scenes including scattering media. We used the image sequence *bali* [29] for the actual data. We rescaled the camera parameters so that the scene depth is contained within the target 3D space of our dehazing cost volume. In this experiment, the network



(a) Reference (b) Source (c) COLMAP [8] (d) Fine-tuned [5] (e) DPSNet [4] (f) Proposed

Fig. 6. Experimental results with actual foggy scenes. Top and bottom rows show clear and hazy scenes, respectively. (a) Reference image, (b) source image, (c) output of COLMAP [8], (d) output of fine-tuned baseline [5], (e) output of DPSNet [4], and (f) output of proposed method.

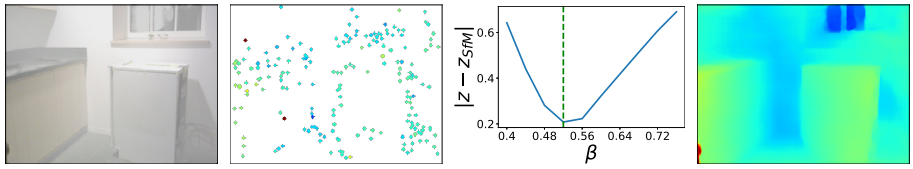


(a) Foggy input (b) Li et al. [29] (c) Fine-tuned [5] (d) DPSNet [4] (e) Proposed

Fig. 7. Experimental results on actual outdoor foggy scenes. (a) foggy input, (b) estimated depth of [29], (c) output of fine-tuned baseline method [5], (d) output of DPSNet [4], and (e) output of proposed method.

took five images as input, one as a reference image and the others as source images. For A and β, we used the estimated values presented in a previous paper [29].

The results are shown in Fig. 7. These scenes are very difficult for learning-based methods due to a large domain gap. The fine-tuned baseline method and DPSNet did not perform well in these scenes. In contrast, the proposed method is more robust and the estimated depth is the closest to that of [29], though the details of the near objects were lost. However, the method proposed by Li et al. [29] requires iterative graph-cut optimization, so it takes a few minutes to estimate depth for one image. Our method, on the other hand, requires only a few seconds to estimate depth for one reference image.



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

Fig. 8. Discusssion on errors of β. β can be adjusted so that output depth corresponds to sparse SfM depth.

4.5 Discusssion on errors of scattering parameters

In this study, scattering parameters A and β are assumed to be estimated beforehand. However, this assumption is sometimes too strict especially for β because the estimation method [29] that uses sparse point clouds obtained at a SfM step and the corresponding pixel intensity is not necessarily numerically stable. However, our dehazing cost volume is parameterized by β, that is, the output depth can be regarded as a function of one variable β. Thus, β can be easily adjusted so that the output depth corresponds to the sparse SfM depth. Figure 8(b) is the SfM depth of (a). (c) shows the L1 error between the sparse depth and output depth with each β. The green dashed line, which represents the ground truth β, corresponds to the global minimum. The final output depth (d) is obtained with this value.

5 Conclusion

We proposed a learning-based MVS method with a novel cost volume, called the dehazing cost volume, which enables MVS methods to be used in scattering media. Differing from the ordinary cost volume, our dehazing cost volume can compute the cost of photometric consistency by taking into account image degradation due to scattering media. This is the first paper to solve the chicken-and-egg problem of depth and scattering estimation by computing the scattering effect using each swept plane in the cost volume without explicit scene depth. The experimental results on synthesized hazy images indicate the effectiveness of our dehazing cost volume in scattering media. We also demonstrated its applicability using the images captured in actual foggy scenes. For future work, we will include the estimation of the scattering coefficient and airlight in our method. We will also extend the proposed method to depth-dependent degradation, other than light scattering, such as defocus blur [40, 41].

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 18H03263 and 19J10003.

References

1. Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. Foundations and Trends® in Computer Graphics and Vision **9** (2015) 1–148
2. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. The European Conference on Computer Vision (ECCV) (2018) 767–783
3. Huang, P., Matzen, K., Kopf, J., Ahuja, N., Huang, J.: Deepmvs: Learning multi-view stereopsis. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2821–2830
4. Im, S., Jeon, H., Lin, S., Kweon, I.S.: Dpsnet: End-to-end deep plane sweep stereo. International Conference on Learning Representations (ICLR) (2019)
5. Wang, K., Shen, S.: Mvdepthnet: real-time multiview depth estimation neural network. International Conference on 3D Vision (3DV) (2018) 248–257
6. Collins, R.T.: A space-sweep approach to true multi-image matching. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (1996) 358–363
7. Zheng, E., Dunn, E., Jojic, V., Frahm, J.: Patchmatch based joint view selection and depthmap estimation. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1510–1517
8. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.: Pixelwise view selection for unstructured multi-view stereo. The European Conference on Computer Vision (ECCV) (2016) 501–518
9. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Transaction on Pattern Analysis and Machine Intelligence **33** (2011) 2341–2353
10. Nishino, K., Kratz, L., Lombardi, S.: Bayesian defogging. International Journal of Computer Vision **98** (2012) 263–278
11. Fattal, R.: Dehazing using color-lines. ACM Transactions on Graphics (TOG) **34** (2014)
12. Berman, D., Treibitz, T., Avidan, S.: Non-local image dehazing. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1674–1682
13. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE Transaction on Image Processing **25** (2016) 5187–5198
14. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.: Single image dehazing via multi-scale convolutional neural networks. European Conference on Computer Vision (ECCV) (2016) 154–169
15. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3194–3203
16. Yang, D., Sun, J.: Proximal dehaze-net: A prior learning-based deep network for single image dehazing. The European Conference on Computer Vision (ECCV) (2018) 702–717
17. Liu, Y., Pan, J., Ren, J., Su, Z.: Learning deep priors for image dehazing. The IEEE International Conference on Computer Vision (ICCV) (2019) 2492–2500
18. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20) (2020) 11908–11915
19. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. The IEEE International Conference on Computer Vision (ICCV) (2017) 4770–4778

20. Narasimhan, S.G., Nayar, S.K., Sun, B., Koppal, S.J.: Structured light in scattering media. Proceedings of the Tenth IEEE International Conference on Computer Vision **I** (2005) 420–427
21. Tsiotsios, C., Angelopoulou, M.E., Kim, T., Davison, A.J.: Backscatter compensated photometric stereo with 3 sources. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 2259–2266
22. Murez, Z., Treibitz, T., Ramamoorthi, R., Kriegman, D.J.: Photometric stereo in a scattering medium. IEEE Transaction on Pattern Analysis and Machine Intelligence **39** (2017) 1880–1891
23. Fujimura, Y., Iiyama, M., Hashimoto, A., Minoh, M.: Photometric stereo in participating media considering shape-dependent forward scatter. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 7445–7453
24. Heide, F., Xiao, L., Kolb, A., Hullin, M.B., Heidrich, W.: Imaging in scattering media using correlation image sensors and sparse convolutional coding. Optics Express **22** (2014) 26338–26350
25. Satat, G., Tancik, M., Rasker, R.: Towards photography through realistic fog. The IEEE International Conference on Computational Photography (ICCP) (2018) 1–10
26. Wang, J., Bartels, J., Whittaker, W., Sankaranarayanan, A.C., Narasimhan, S.G.: Programmable triangulation light curtains. The European Conference on Computer Vision (ECCV) (2018) 19–34
27. Caraffa, L., Tarel, J.: Stereo reconstruction and contrast restoration in daytime fog. Asian Conference on Computer Vision (ACCV) (2012) 13–25
28. Song, T., Kim, Y., Oh, C., Sohn, K.: Deep network for simultaneous stereo matching and dehazing. British Machine Vision Conference (BMVC) (2018)
29. Li, Z., Tan, P., Tang, R.T., Zou, D., Zhou, S.Z., Cheong, L.: Simultaneous video defogging and stereo reconstruction. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 4988–4997
30. Tan, R.T.: Visibility in bad weather from a single image. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008) 1–8
31. Berman, D., Treibitz, T., Avidan, S.: Air-light estimation using haze-lines. The IEEE International Conference on Computational Photography (ICCP) (2017)
32. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 5038–5047
33. Xiao, J., Owens, A., Torralba, A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. The IEEE International Conference on Computer Vision (ICCV) (2013) 1625–1632
34. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. The International Conference on Intelligent Robot Systems (IROS) (2012)
35. Fuhrmann, S., Langguth, F., Goesel, M.: Mve: a multi-view reconstruction environment. Eurographics Workshop on Graphics and Cultural Heritage (2014) 11–18
36. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. arXiv:1512.03012 (2015)
37. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Twenty-eighth Conference on Neural Information Processing Systems (NeurIPS) (2014)

38. Tateno, K., Tombari, F., Laina, I., Navab, N.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 6243–6252
39. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 4104–4113
40. Gur, S., Wolf, L.: Single image depth estimation trained via depth from defocus cues. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 7683–7692
41. Maximov, M., Galim, K., Leal-Taixe, L.: Focus on defocus: Bridging the synthetic to real domain gap for depth estimation. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 1071–1080