Deep Network for Simultaneous Stereo Matching and Dehazing

Taeyong Song[1]
sty37@yonsei.ac.kr
Youngjung Kim[1]
read12300@yonsei.ac.kr
Changjae Oh[2]
c.oh@qmul.ac.uk
Kwanghoon Sohn*[1]
khsohn@yonsei.ac.kr

[1] School of Electrical and Electronic Engineering
Yonsei University
Seoul, South Korea

[2] School of Electronic Engineering and Computer Science
Queen Mary University of London
London, England

Abstract

Unveiling the image structure and dense correspondence under the haze layer remains a challenging task, since the scattering effects cause image features to be less distinctive. In this paper, we introduce a deep network that simultaneously estimates a clear latent image and disparity from a hazy stereo image pair. To this end, inspired by a physical model of hazy image acquisition, we propose a dehazing loss function which serves as an additional cue for establishing dense correspondence. We show that stereo matching and dehazing can be synergistically formulated by incorporating depth information from haze transmission into the stereo matching process, and *vice versa*. As a result, our method estimates high-quality disparity for scenes in scattering media, and produces appearance images with enhanced visibility. We quantitatively evaluate the proposed method on synthetic datasets and provide an extensive ablation study. Experimental results demonstrate that our approach outperforms the recent state-of-the-art methods on both dehazing and stereo matching tasks.

1 Introduction

Dense depth information is indispensable to computer vision applications, including 3D reconstruction [28], object recognition [10], intrinsic image decomposition [4], and autonomous driving for vehicles [8]. Although active 3D scanner such as LiDAR and structured light can be used for direct depth acquisition, sensing depth from stereo camera is a more cost effective solution [19]. Specifically, for a pixel $\mathbf{x}=(x,y)$ in left image, its correspondence may be found at location $\tilde{\mathbf{x}}=(x-d,y)$ in the right, where d is often referred to as *disparity*. Since depth is inversely proportional to d with calibration parameters, a stereo matching method is instead targeted for generating dense disparity [27]. Stereo matching has been traditionally cast as an energy minimization problem with several stages of optimization [21, 27]. Nonetheless, it is difficult to select the correct disparity at ill-posed regions,

* Corresponding Author.

Figure 1: (From left to right) Clear left of the stereo pair, hazy left, depth maps with clear and hazy stereo pairs. We use the stereo algorithm of [20] trained with clear stereo pairs.

i.e., occluded and texture-less regions. The correspondence cannot be decided for a pixel appearing in one image but occluded in the other. While for texture-less regions, many possible correspondences can exist leading to systematic errors in disparity estimation. Taking advantages of a large amount training data and more efficient computing hardware, recent methods formulate stereo matching as a supervised learning task [20]. Stereo matching with convolutional neural networks (CNNs) achieves significant gain compared to traditional approaches in terms of both accuracy and speed [19, 20, 31].

In outdoor scenes, we are often faced with opaque objects covering over more distant regions of an image. The presence of tiny particles in the atmosphere causes the deflections of light, which travels from objects to camera. This physical phenomenon is known as *haze* or *fog*, and makes the appearance to be attenuated along its path [16]. However, the current stereo matching methods based on CNNs [19, 20, 31] are only designed for images captured in clear weather. Consequently, bad atmospheric conditions present a significant difficulty in disparity estimation as shown in Fig. 1. One reason for this is that hazy images are associated with visibility decrease that causes image features to be less distinctive and confuses the matching cost computation. In the best case, the existing methods [19, 20, 31] are correct only up to a critical distance as the visibility is an exponentially decreasing function of depth [16]. A naive solution is to apply dehazing and stereo matching algorithms sequentially, or to fine-tune the CNNs with hazy images. We will show that such methods are marginally helpful for stereo matching in scattering media.

While haze poses a challenge for stereo matching, it may provide a depth cue in the gray level of far-away objects [3]. The depth cue from haze transmission is particularly interesting since it is complementary to the stereo depth [3]. The former provides more accurate depth ordering for remote objects[1] [17], and the latter is more reliable for nearby objects [12]. Thus, stereo matching also can help dehazing. Previous dehazing methods are mainly designed to restore a single image, which is an under-constrained problem, i.e., there exists a large number of valid solutions. This holds for every pixel and can not be resolved independently at each pixel given a single image [5]. Recent techniques employ additional heuristic assumptions, including dark channel prior [13] and maximum local contrast [29]. However, all of these can be fooled in sky regions or objects with saturated colors. It is known that even rough depth information from single image can improve the dehazing performance [5, 24]. A precise depth estimation from stereo image will reduce the ambiguity in dehazing.

In this paper, we devise a multi-task CNN that simultaneously predicts scene depth and clear appearance from a hazy stereo pair. The key insight is that the depth cues from stereo and haze transmission are complementary to each other. Building on top of two-stream CNN, we enforce that depth information from stereo matching to be incorporated into the dehazing process, and *vice versa*. This feature enables the underlying model to effectively predict depth and dehazed images with high consistency and corresponding accuracy. Extensive evaluations on synthetic dataset demonstrate the effectiveness and flexibility of the proposed

[1] Thicker haze is associated with larger distance [16].

method. Finally, we illustrate some further examples of our model generalizing to real-world scenes.

2 Related Work

2.1 Stereo matching

There is a large body of literature on stereo matching. Among various methodologies, we review a few of them with emphasis placed on recent methods using CNNs. In early stage, CNNs are used to measure the similarity between images patches. Han *et al.* introduced a Siamese network which extracts features from a pair of patches followed by similarity measure [11]. Similarly, Zbontar *et al.* [31] presented a series of CNN architectures called MC-CNN for pairwise matching, and applied these in disparity estimation. Luo *et al.* [19] proposed to learn a probability distribution over all disparity values. They replaced the concatenation and subsequent processing layers by a single product layer, measuring costs in less than a second. The learned similarity metrics [11, 19, 31] outperform the traditional handcrafted ones, such as sum of absolute difference and normalized cross correlation. However, a number of post-processing steps are still required to produce compelling results. More recent works try to learn stereo matching in an end-to-end fashion by carefully designing and supervising the CNNs. Mayer *et al.* [20] proposed the DispNet, where the network is trained end-to-end using synthetically generated stereo pairs and the corresponding disparity. Pang *et al.* [23] combined the DispNet and cascade residual learning to produce disparities with more details. Kendall *et al.* [14] devised the GC-NET using 3D convolution and proposed differentiable soft-argmin operation, which allows the network to be trained end-to-end with sub-pixel accuracy. Yu *et al.* [30] realized the color-guided cost aggregation using CNNs, and collaborated with the deep cost computation. All these methods, however, are designed for images captured in clear scenes. Bad weather conditions reduce the quality of stereo pairs, and introduce artifacts in disparity estimation.

2.2 Dehazing

Numerous methods have been proposed to solve the dehazing problem. Tan *et al.* [29] formulated the image dehazing as Markov random fields (MRF), and obtained a clear image by maximizing local contrast. Fattal assumed that a hazy scene can be divided into regions of constant albedo, and inferred the transmission using such assumption [8]. Several heuristic assumptions have also been made on natural images to estimate haze transmission. For instance, the dark channel prior [13] imposes that patches of natural images contain very low intensity in at least one color channel. The color-line prior [6] relied on the regularity of natural images, where small patches typically exhibit one-dimensional distribution in RGB color space. Recent methods adopt CNNs to extract haze-relevant features from a single image. Ren *et al.* [24] proposed to use multi-scale CNN to estimate transmission maps from hazy images directly. Cai *et al.* [2] proposed an end-to-end CNN model to estimate the transmission with novel BReLU activation function. Zhang *et al.* [32] devised a densely connected architecture, and employed multi-level pyramid pooling for estimating edge-preserving transmission maps. While these single image dehazing methods work to some extent, they often regard nearby objects as far away ones with saturated color, and introduce visual artifacts [17].

2.3 Joint stereo matching and dehazing

The aforementioned methods treat stereo matching and dehazing as independent tasks, thus ignoring their complementary relationship. In the literature, only a few works have been devoted to solve stereo matching with hazy images. The most related work to ours is that of Caraffa and Tarel [3]. They proposed a MRF model of the stereo matching and dehazing, which can be optimized iteratively with α-expansion. Roser *et al.* [25] iterated the dense stereo matching and estimation of clear images according to the scattering equation. The matting Laplacian filter was used to enhance the overall quality of disparity. Li *et al.* [17] improved the data consistency term in multi-view stereo by explicitly modeling the appearance change due to the scattering effects. They further enforced the ordering consistency between scene depth and hazy transmission at neighboring pixels. These attempts [3, 17, 25] are conceptually similar to ours, but suffer from two main drawbacks. First, all methods are inherently iterative, and rely on global optimization techniques such as multi-label graph cut [1], leading to a huge computational overhead. Second, hand-crafted features are used to measure the matching cost for stereo reconstruction. In contrast, our learned model directly fuses the depth cues from stereo matching and dehazing, and produces stronger results.

3 Proposed Method

We denote by $\{I_L, I_R\}$ left and right images of a stereo pair. It is assumed that $\{I_L, I_R\}$ are observed after perturbation of atmospheric scattering and optics. The images without such perturbation will be denoted as J_L and J_R, respectively. The unknowns are disparity map D aligned to left image and J_L. Our goal is to fuse depth cues from stereo matching and haze transmission to achieve a better reconstruction of $\{J_L, D\}$. Before formulating our approach, we give an overview of the scattering model that is used for hazy image generation [16].

3.1 Haze model

Haze is a phenomenon that results from the scattering of light causing an attenuation in the appearance of scene. The effect of haze (atmospheric scattering) can be mathematically modeled as follows [16]:

$$I(\mathbf{x}) = J(\mathbf{x})T(\mathbf{x}) + A(\mathbf{x})(1 - T(\mathbf{x})), \tag{1}$$

where A is the atmospheric light, and T is the medium transmission determining the portion of light scattering. If the atmosphere is homogeneous, we can express the transmission as $T(\mathbf{x}) = e^{-\beta z(\mathbf{x})}$, where β is the scattering coefficient associated with the density of media, and z is the scene depth. The clear image J can then be obtained in the inverse way:

$$J(\mathbf{x}) = \frac{I(\mathbf{x}) - A(\mathbf{x})(1 - T(\mathbf{x}))}{\max(\varepsilon, T(\mathbf{x}))}, \tag{2}$$

where ε is a constant for the numerical stability. As a result, the task of dehazing can be divided into two tasks: estimations of the transmission map T and the atmospheric light A.

3.2 Stereo Matching Network

The lessons of the previous works [14, 20, 23, 30] inspire us to employ CNN for disparity estimation. Our Stereo matching network takes the hazy images $\{I_L, I_R\}$ as input, and outputs

A-Network

Layer	Kernel	Channel	Input
cnv_A1	7x7	3 / 64	I_L
cnv_A2	5x5	64 / 48	cnv_A1
cnv_A3	3x3	48 / 32	cnv_A2
cnv_A4	3x3	32 / 16	cnv_A3
$\hat{A}$	3x3	16 / 1	cnv_A4

T-network (Encoder)

Layer	Kernel	Channels	In	Out	Input
cnv_t1a	7x7	3/32	1	2	I_L
cnv_t1b	7x7	32/32	2	2	cnv_t1a
cnv_t2a	5x5	32/64	2	4	cnv_t1b
cnv_t2b	5x5	64/64	4	4	cnv_t2a
cnv_t3a	3x3	64/128	4	8	cnv_t2b
cnv_t3b	3x3	128/128	8	8	cnv_t3a
cnv_t4a	3x3	128/256	8	16	cnv_t3b
cnv_t4b	3x3	256/256	16	16	cnv_t4a
cnv_t5a	3x3	256/512	16	32	cnv_t4b
cnv_t5b	3x3	512/512	32	32	cnv_t5a
cnv_t6a	3x3	512/1024	32	64	cnv_t5b
cnv_t6b	3x3	512/1024	64	64	cnv_t6a

T-network (Decoder)

Layer	Kernel	Channels	In	Out	Input
upcnv_t6	4x4	1024/512	64	32	cnv_t6b
iconv_t6	3x3	1024/512	32	32	{upcnv_t6, cnv_t5b}
upcnv_t5	3x3	512/256	32	16	iconv_t6
iconv_t5	3x3	512/256	16	16	{upcnv_t5, cnv_t4b}
upcnv_t4	3x3	256/128	16	8	iconv_t5
iconv_t4	3x3	256/128	8	8	{upcnv_t4, cnv_t3b}
upcnv_t3	3x3	128/64	8	4	iconv_t4
iconv_t3	3x3	128/64	4	4	{upcnv_t3, cnv_t2b}
upcnv_t2	3x3	64/32	4	2	iconv_t3
iconv_t2	3x3	64/32	2	2	{upcnv_t2, cnv_t1b}
upcnv_t1	3x3	32/16	2	1	iconv_t2
$\hat{T}$	3x3	16/1	1	1	deconv_ut_6_1

Table 1: Our *A*- and *T*-network architectures for dehazing. 'Channels' is the number of input and output channels. 'In' or 'Out' is the downsampling factor relative to the input image. '$\{\cdot,\cdot\}$' denotes the concatenation operator.

the disparity D aligned to the left image. We use the DispNetC-1D [20] (here "C-1D" means that the network has a 1D correlation layer) as our baseline architecture. For a concise presentation, the detailed architecture of DispNetC is omitted here (please refer to [20]). In a nutshell, the two images $\{I_L, I_R\}$ are processed separately up to second convolution layer. The resulting activations are correlated horizontally to construct the cost volume. Disparity values are then regressed from the cost volume using the following encoder-decoder network. The original DispNetC [20] outputs disparity map at half the resolution of inputs. Differently, we append extra upconvolution and convolution layers to obtain the disparity at the same size of input images. We then follow the typical supervised learning paradigm and compute L_1 loss between the estimate $\hat{D}$ and the ground-truth disparity D_{gt}.

$$\mathcal{L}_D = \sum_{\mathbf{x}} \left| \hat{D}(\mathbf{x}) - D_{gt}(\mathbf{x}) \right|_1. \tag{3}$$

The stereo matching network outputs disparity predictions at five different scales, which double in spatial resolution at the subsequent scales [20].

3.3 Dehazing network

We also utilize CNNs to estimate the clear appearance J from hazy image I. The dehazing network takes a hazy image, and predicts the transmission T and atmospheric light A, followed by clear image estimation using (2).

Architecture The architecture of our dehazing network is presented in Table 1. Our network is inspired by the DispNet [20], but features some important modifications for the dehazing. It consists of two modules: T-network and A-network for transmission and atmospheric light, respectively. The T-network is the critical component of dehazing network, being responsible for extracting depth information and relative haze level. We thus adopt fully convolutional encoder (from cnv1a to cnv6b) and decoder (from upcnv6) architecture for ensuring high-capacity. The encoder extracts a variety of multi-scale haze-relevant features, and the decoder estimates scene transmission maps from these representations. We use skip-connections [18] from the encoder's activation blocks to compensate for information loss during convolutions and pooling. The A-network has five convolutional layers with 7×7, 5×5, and 3×3 kernels. We design the A-network to have a compact parameterization since the manifold of A is topologically much simpler than that of T. For both the networks, we use the leaky rectified linear unit $\max(0.2x, x)$ (LReLU) as the pointwise nonlinearity.

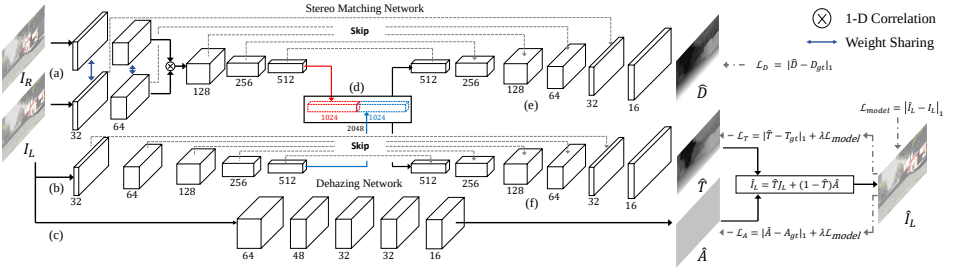

Figure 2: Our full architecture for simultaneous stereo matching and dehazing.

Loss function We define the loss functions for dehazing network as follows:

$$\mathcal{L}_T = \sum_{\mathbf{x}} \left|\hat{T}(\mathbf{x}) - T_{gt}(\mathbf{x})\right|_1 + \lambda \mathcal{L}_{\text{model}}, \quad \mathcal{L}_{\text{A}} = \sum_{\mathbf{x}} \left|\hat{A}(\mathbf{x}) - A_{gt}(\mathbf{x})\right|_1 + \lambda \mathcal{L}_{\text{model}}, \tag{4}$$

where $\hat{T}$ and $\hat{A}$ denote the estimated transmission map and atmospheric light, respectively. $\lambda > 0$ is a balancing parameter. Note that the existing methods using CNNs train T- and A-networks separately [32], or estimate A heuristically [2, 24]. These methods [24, 32] disregard the correlation between transmission map and atmospheric light, and tends to amplify the image noise. Differently, our dehazing network shares the forward model consistency loss function $\mathcal{L}_{\text{model}}$:

$$\mathcal{L}_{\text{model}} = \sum_{\mathbf{x}} \left|\left(J(\mathbf{x})\hat{T}(\mathbf{x}) + \hat{A}(\mathbf{x})(1-\hat{T}(\mathbf{x}))\right) - I(\mathbf{x})\right|_1. \tag{5}$$

The derivation is straightforward from (1). That is, we measure the sum of absolute error of forward model consistency (1). $\mathcal{L}_{\text{model}}$ helps to back propagate errors to each sub-network concurrently, and to avoid overfitting when training each sub-network individually. We will demonstrate that our dehazing loss produces more consistent results with various haze levels.

To learn the parameters in T- and A-networks, we require the partial derivatives of the dehazing loss with respect to $\hat{T}$ and $\hat{A}$. Since the input hazy image I is the convex combination of J and A, we can easily derive the following derivatives:

$$\frac{\partial \mathcal{L}_T}{\partial \hat{T}} = \sum_{\mathbf{x}} \left(\text{sgn}(\hat{T}(\mathbf{x}) - T_{gt}(\mathbf{x})) + \lambda \tau(\mathbf{x}) \left(J(\mathbf{x}) - \hat{A}(\mathbf{x})\right)\right), \tag{6}$$

$$\frac{\partial \mathcal{L}_A}{\partial \hat{A}} = \sum_{\mathbf{x}} \left(\text{sgn}(\hat{A}(\mathbf{x}) - A_{gt}(\mathbf{x})) + \lambda \tau(\mathbf{x}) \left(1 - \hat{T}(\mathbf{x})\right)\right), \tag{7}$$

where $\tau(\mathbf{x}) = \text{sgn}\left(\left(J(\mathbf{x})\hat{T}(\mathbf{x}) + \hat{A}(1-\hat{T}(\mathbf{x}))\right) - I(\mathbf{x})\right)$. These derivatives, which can be computed efficiently using point-wise operations, are further back-propagated onto T- and A-networks. Note that after $\hat{T}$ and $\hat{A}$ are estimated by our dehazing network, we solve (2) to obtain the clear appearance J.

3.4 Simultaneous stereo matching and dehazing

Full architecture We now explain our full architecture illustrated in Fig. 2 for simultaneous stereo matching and dehazing. We first use the two stream encoders to extract depth cues from stereo matching and single haze transmission, as shown in Fig. 2(a) and (b). To

Stereo Matching

Network (Training Img.)	Testing Img.	3PE (three-pixel-error)			EPE (endpoint-error)		
		FT3D [20]	Driving [20]	Avg.	FT3D [20]	Driving [20]	Avg.
Stereo matching net (Clear)	Clear	0.1598	0.2126	0.1786	2.8718	3.8793	3.2316
	Hazy	0.5421	0.5049	0.5288	22.372	16.553	20.294
	Dehazed	0.3799	0.4073	0.3897	10.606	8.3087	9.7859
Stereo matching net (Hazy)	Hazy	0.3781	0.4455	0.4022	8.4375	10.031	9.0066
Full architecture (Hazy)	Hazy	**0.2643**	**0.4052**	**0.3147**	**3.7753**	**6.1859**	**4.6362**

Dehazing

Network	PSNR (dB)		
	FT3D [20]	Driving [20]	Avg.
Dehazing net w/o $\mathcal{L}_{model}$	19.692	19.134	19.492
Dehazing net	20.225	19.541	19.981
Full architecture w/o $\mathcal{L}_{model}$	22.031	21.248	21.751
Full architecture	**22.900**	**21.744**	**22.487**

Table 2: The results of ablation study on the FT3D and Driving datasets [20]. (left) stereo matching and (right) dehazing tasks, respectively.

combine these information, we associate the intermediate activations via concatenation at the end of the encoders (Fig. 2(d)). The combined activations are then followed by the task-specific decoders for the stereo matching and dehazing, respectively (Fig. 2(e) and (f)). Each decoder in the stereo matching and T-network keeps their skip connections with the corresponding encoders (the dotted lines in Fig. 2). Note that we directly fuse the depth cues from the learned cost-volume and haze-relevant features to achieve better reconstruction of disparity and transmission. This is a noticeable difference from the previous approaches [3, 17] that are based on the iterative use of multi-label graph cuts [1] or loopy belief propagation [2]. The A-network estimates the atmospheric light along Fig. 2(c), and interacts with the T-network through the model consistency loss $\mathcal{L}_{model}$.

Loss function Finally, we jointly train our full architecture in Fig. 2. The overall loss function is the summation of three terms:

$$\mathcal{L}_{Total} = \mathcal{L}_D + \mathcal{L}_T + \mathcal{L}_A. \tag{8}$$

Consequently, all three estimates $(\hat{D}, \hat{T}, \hat{A})$ reinforce each other to optimize the whole model. It is trained by back-propagation in an end-to-end manner. We use the Adam solver [15] and their default setting. The initial step size is set to 10^{-4} which is kept constant for the first 100,000 iterations, and after that it is halved every 20,000 iterations until the end (400,000 iterations in total).

4 Experimental Results

Experimental setup and detailed analysis of the proposed method are presented in this section. We conduct extensive ablation studies to demonstrate the effectiveness of our joint model. We also compare our method with other state-of-the-art approaches, including [2, 3, 13, 24]. The proposed method are implemented with the tensorflow library [22], and are trained using NVIDIA GeForce GTX 1080. The results for the comparison, except for [3], are obtained from source codes provided by the authors. Since the source code of [3] is not available publicly, we directly take their results from the original paper. We ensure that all the learning-based methods are trained with the same procedure.

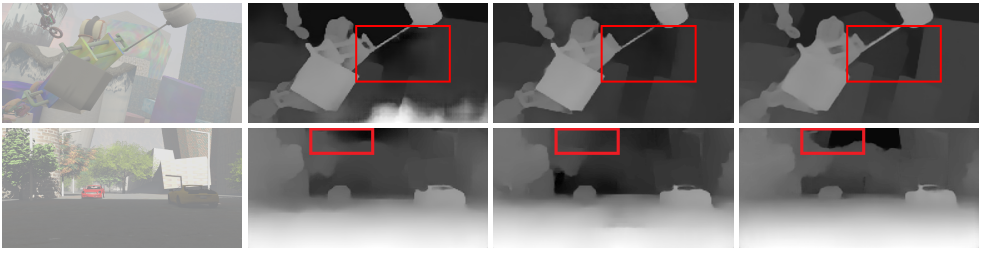

Figure 3: Visual results of ablation study for stereo matching. (From left to right) Hazy left image, disparities from the stereo network fine-tuned on hazy stereo, sequential dehazing and stereo networks, and our full architecture.

4.1 Implementation details

Dataset We use three publicly available datasets for training and testing in this work:

- *FlyingThings3D* (FT3D) [20]: it is a large scale dataset that consists of synthetic stereo pairs and the corresponding disparities. We discard samples with unreasonably large disparities, resulting in 20,000 stereo pairs for training and 3,000 for testing.
- *Driving* [20]: it provides synthetic but mostly naturalistic 4,000 street scenes from the viewpoint of a driving car. Since there is no official split, similar to the practice in [9], we divide the dataset into training and testing splits (3,750 and 250, respectively).
- *FRIDA3* [3]: a small dataset capturing synthetic outdoor scenes, which has 66 stereo pairs with ground-truth disparities. We use this for testing only.

Thus, the total numbers of training and testing samples are 23,750 and 3,316, respectively.

Haze generation For each left and right image, the depth map z is first recovered from the ground-truth disparity with $z(\mathbf{x}) = bf/D(\mathbf{x})$, where b and f denote the baseline distance and the camera focal length. We then choose random $\beta \in (1.0, 2.2)$ and $A \in (0.7, 1.0)$, and generate the transmission map T. Finally, we synthesize the input stereo pairs using (1). We do not perform any data augmentation as geometric shifts could break the epipolar constraint, and lead to negative disparities [20].

4.2 Ablation study

In this section, we provide an intensive ablation study to see how each component contributes to the performance of our model. The results for the stereo matching is reported in Table 2(left). The evaluation is performed on the 3,250 testing split of FT3D and Driving datasets [20].

We additionally train the stereo matching network only in Section 3.2 with different configurations: training with clear and hazy stereo pairs. It can be seen that the performance of the stereo matching network trained with clear image is degraded in the presence of haze, and the fine-tuning on hazy stereo pairs is marginally helpful. We also provide the result of the straightforward combination of dehazing and stereo matching networks in Table 2(left). In this case, we separately train the dehazing network in Section 3.3 and perform the stereo matching on *dehazed* images. Our full model achieves the best quantitative performance in terms of three-pixel-error (3PE) and endpoint-error (EPE). Visual comparisons of stereo matching are shown in Fig 3. Our method produces the plausible disparity even at the far-away object (see the red-box in Fig 3). These experiments demonstrate that depth information from haze transmission can serve as an additional cue for disparity estimation.

Figure 4: Visual results of ablation study for dehazing. (From left to right) Hazy image, results from the dehazing network and our full architecture, and ground-truth image.

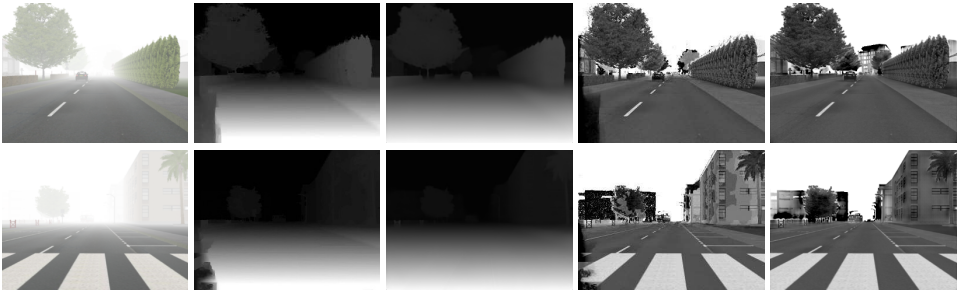

Figure 5: Simultaneous stereo matching and dehazing on the FRIDA3 dataset [3]. (From left to right) Hazy left, disparities from [3] and ours, and dehazed left images from [3] and ours.

The quantitative results for dehazing task is reported in Table 2(right). We compare the results from the dehazing network and our full model. The effect of $\mathcal{L}_{model}$ is also analyzed. Combining the depth cues from stereo and single haze transmission has the most impact on dehazing performance, and the forward model consistency loss $\mathcal{L}_{model}$ results in the further improvement. The gain from our full model exceeds about 3dB in average. We show the visual comparisons of dehazing in Fig. 4. It can be observed that the single dehazing network suffers from the airlight-albedo ambiguity [5]. That is, the transmission is overestimated for bright pixels (or underestimated for dark ones). In contrast, our full architecture estimates the transmission map, aligning to the actual depth ordering.

4.3 Comparison with state of the arts and further experiment

Here, we compare our method with the current state-of-the-art methods in terms of stereo matching or dehazing. To our best knowledge, simultaneous stereo matching and dehazing using CNNs has not been investigated earlier in the literature. For the stereo matching, the only work that is directly comparable to ours is the method of Caraffa *et al.* [3]. They formulate a joint MRF model of the both tasks, which is optimized iteratively using α-expansion [3]. We provide an evaluation comparing our method both quantitatively and qualitatively with [3] on the FRIDA3 dataset. For the dehazing evaluation, we compare to several single image dehazing methods, i.e., the dark channel prior (DCP) [13], MSCNN [24], and DehazeNet [7]. The last two methods are based on the CNNs. Finally, we show the result of our fine-tuned model on KITTI dataset [8].

Comparison with [3] Visual examples on the FRIDA3 dataset [3] are presented on Fig. 5. As can be observed, our results are much visually pleasant for both near and far regions. The

Stereo Matching (1PE)	
Caraffa [3]	Full architecture
0.172	0.124

Dehazing (PSNR in dB)				
DCP [13]	MSCNN [27]	DehazeNet [2]	Our dehazing net	Full architecture
13.388	15.512	15.723	16.124	**17.383**

Table 3: Quantitative comparison on the FRIDA3 dataset [3]. (left) stereo matching and (right) dehazing tasks, respectively. Note that "Our dehazing net" takes a single image only.

Figure 6: Simultaneous stereo matching and dehazing on the KITTI dataset [8]. (From left to right) Hazy left, estimated disparity, and dehazed image.

method [3] suffers from artifacts in lower left corner and far objects for disparity estimation. These artifacts generate large error on the dehazed images (see the road, building, and tree in Fig. 5). Quantitatively, we measure and report the one-pixel-error (1PE) in Table 3(left).

Comparison with the recent single image dehazing We perform the quantitative evaluation with recent single image dehazing methods on the FRIDA3 dataset [3]. All the learning-based methods [2, 24] including ours are trained on FT3D and Driving datasets [20]. The results are reported in Table 3(right). In this table, "Our dehazing net" denotes the dehazing network in Section 3.3 trained separately, and thus it takes a single hazy image as input. Our dehazing network outperforms the existing single image methods [2, 13, 24], and the depth cue from stereo matching significantly improves the performance of dehazing.

Further experiment Due to the absence of a real dataset with dense ground-truth disparity, our evaluation was limited to a synthetic dataset. To establish possible extension of our method to real-world scenarios, we alternatively use hazy images and disparities from *Foggy Cityscapes* [26]. We fine-tune our network on 2975 training and 500 validation images on *Foggy Cityscapes*. In the testing stage, we follow the approach introduced in [26] to simulate the haze on the KITTI dataset [8]. Figure 6 shows the results of simultaneous stereo matching and dehazing on KITTI dataset [8]. This figure demonstrates that our fine-tuned model can be successfully applied to real scenes.

5 Conclusion

In this paper, we have introduced a joint learning framework for simultaneous stereo matching and dehazing. Different from the previous methods, our deep architecture directly combines depth cues from stereo image and single haze transmission. We further enforce the estimated transmission and atmospheric light to be consistent with the scattering model. As a result, our method estimates high-quality disparities in scattering media, and produces appearance images with enhanced visibility. Experiments demonstrate the effectiveness of our method on both stereo matching and dehazing.

Acknowledgements

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7069370).

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[2] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: an end-to-end system for single image haze removal. *IEEE trans. Image Process.*, 25(11):5187–5198, 2016.

[3] L. Caraffa and J.P. Tarel. Stereo reconstruction and contrast restoration in daytime fog. *ACCV*, 2012.

[4] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. *ICCV*, 2013.

[5] R. Fattal. Single image dehazing. *ACM Trans. Graph.*, 27(3):72, 2008.

[6] R. Fattal. Dehazing using color-lines. *ACM Trans. Graph.*, 34(1):13, 2014.

[7] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient belief propagation for early vision. *Int. Journ. Comput. Vis.*, 70(1):41–54, 2006.

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *CVPR*, 2012.

[9] S. Gidaris and N. Komodakis. Detect, replace, refine: deep structured prediction for pixel wise labeling. *CVPR*, 2017.

[10] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. *ECCV*, 2014.

[11] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A.C. Berg. Matchnet: unifying feature and metric learning for patch-based matching. *CVPR*, 2015.

[12] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, 2004.

[13] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2341–2353, 2011.

[14] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. *ICCV*, 2017.

[15] D. Kingma and J. Ba. Adam: a method for stochastic optimization. *Int. Conf. Learn. Repres.*, 2015.

[16] H. Koschmieder. Theorie der horizontalen sichtweite: kontrast und sichtweite. *Keim and Nemnich*, 1925.

[17] Z. Li, P. Tan, R.T. Tan, D. Zou, S.Z. Zhou, and L.F. Cheong. Simultaneous video defogging and stereo reconstruction. *CVPR*, 2015.

[18] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.

[19] W. Luo, A.G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. *CVPR*, 2016.

[20] N. Mayer and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *CVPR*, 2016.

[21] M.G. Mozerov and J. Weijer. Accurate stereo matching by two-step energy minimization. *IEEE trans. Image Process.*, 24(3):1153–1163, 2015.

[22] Online. https://www.tensorflow.org/.

[23] J. Pang, W. Sun, J. Ren, C. Yang, and Q. Yan. Cascaded residual learning: a two-stage convolutional neural network for stereo matching. *ICCV Workshop*, 2017.

[24] W. Ren, S. Liu, H. Zhang, J. Pan, and X. Cao. Single image dehazing via multi-scale convolutional neural networks. *ECCV*, 2016.

[25] M. Roser, M. Dunbabin, and A. Geiger. Simultaneous underwater visibility assessment, enhancement and improved stereo. *ICRA*, 2014.

[26] C. Sakaridis, D. Dai, and L.V. Gool. Semantic foggy scene understanding with synthetic data. *Int. Journ. Compt. Vis.*, 2018.

[27] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. Journ. Compt. Vis.*, 47(1):7–42, 2002.

[28] S. N. Shinha and M. Pollefeys. Multi-view reconstruction using photo-consistency and exact silhouette constraints: a maximum-flow formulation. *ICCV*, 2005.

[29] R. Tan. Visibility in bad weather from a single image. *CVPR*, 2008.

[30] L. Yu, Y. Wang, Y. Wu, and Y. Jia. Deep stereo matching with explicit cost aggregation sub-architecture. *AAAI*, 2018.

[31] J. Zbontar and Y. Lecun. Stereo matching by training a convolutional neural network to compare image patches. *Journ. of Machi. Learn. Research*, 17(2):1–32, 2016.

[32] H. Zhang and V.M. Patel. Densely connected pyramid dehazing network. *CVPR*, 2018.