CNN-Based Simultaneous Dehazing and Depth Estimation

Byeong-Uk Lee[1], Kyunghyun Lee[1], Jean Oh[2] and In So Kweon[1]

***Abstract*— It is difficult for both cameras and depth sensors to obtain reliable information in hazy scenes. Therefore, image dehazing is still one of the most challenging problems to solve in computer vision and robotics. With the development of convolutional neural networks (CNNs), lots of dehazing and depth estimation algorithms using CNNs have emerged. However, very few of those try to solve these two problems at the same time. Focusing on the fact that traditional haze modeling contains depth information in its formula, we propose a CNN-based simultaneous dehazing and depth estimation network. Our network aims to estimate both a dehazed image and a fully scaled depth map from a single hazy RGB input with end-to-end training. The network contains a single dense encoder and four separate decoders; each of them shares the encoded image representation while performing individual tasks. We suggest a novel depth-transmission consistency loss in the training scheme to fully utilize the correlation between the depth information and transmission map. To demonstrate the robustness and effectiveness of our algorithm, we performed various ablation studies and compared our results to those of state-of-the-art algorithms in dehazing and single image depth estimation, both qualitatively and quantitatively. Furthermore, we show the generality of our network by applying it to some real-world examples.**

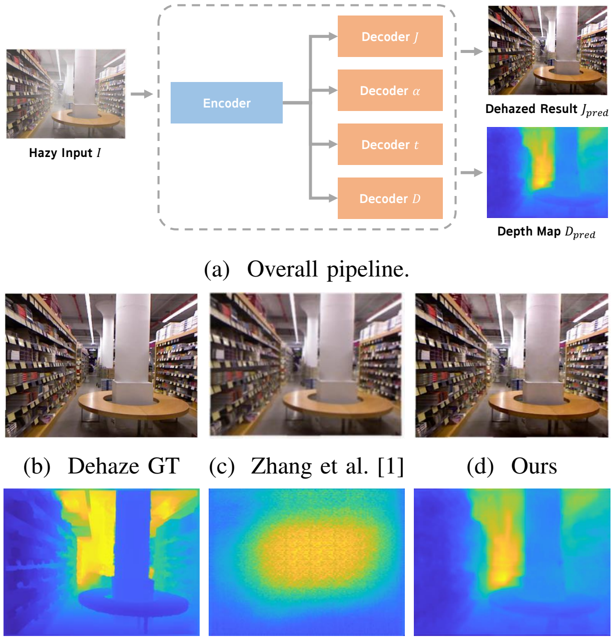
I. INTRODUCTION

What is the point in developing the best vision algorithm in the world, if you cannot clearly see through the scene? Most vision-based algorithms, such as object detection, semantic segmentation, depth estimation, and so forth, use images or videos taken by camera. However, if the dataset is obtained from a scene where there is fire/smoke, fog, or serious air pollution like smog, these algorithms are likely to fail. Haze, which is the term that describes all of the examples mentioned above, refers to an atmospheric phenomenon in which particles in the air cause light absorption and scattering. This results in image degradation with a loss of contrast and color clarity. Therefore, haze removal, i.e., dehazing is one of the most critical, yet ill-posed problem to solve in computational photography, computer vision and robotics.

Traditionally, dehazing algorithms have been based on the atmospheric scattering model proposed in [3], [4]. The atmospheric scattering model is defined as

$$I(x) = J(x)t(x) + \alpha(1 - t(x)), \tag{1}$$

where $I(x)$ is the observed hazy image, $J(x)$ is the true scene radiance, α is the global atmospheric light, and $t(x)$ refers to the medium transmission map. Therefore, many dehazing algorithms focus on obtaining atmospheric lighting and a transmission map to recover the original image of the scene by solving the following equation:



(a) Overall pipeline.

(b) Dehaze GT (c) Zhang et al. [1] (d) Ours

(e) Depth GT (f) Ma et al. [2] (g) Ours

Fig. 1: Overall pipeline and results comparisons to state-of-the-art methods.

$$J(x) = \frac{I(x) - \alpha(1 - t(x)}{t(x)} \tag{2}$$

However, this is a challenging problem because the depth information of a hazy scene is usually unknown, while the transmission map $t(x)$ is a range-based value:

$$t(x) = e^{-\beta d(x)} \tag{3}$$

with β as the scattering coefficient of the atmosphere and $d(x)$ as the depth map of the scene. Moreover, current widely-used depth sensors, such as laser-based LiDAR or stereo matching-based Kinect, etc., are not reliable in a hazy scene because the scattering particles in the atmosphere obstruct the depth acquirement.

This paper proposes an end-to-end convolutional neural network designed for simultaneous dehazing and depth estimation, focusing on the fact that a transmission map has a correlation with depth information. The network consists of a shared encoder and four decoders, where each decoder is assigned for an individual task. Some of the tasks are to

[1] Byeong-Uk Lee, Kyunghyun Lee and In So Kweon are with the School of Electrical Engineering, KAIST, Daejeon, Republic of Korea. E-mail: {byeonguk.lee, kyunghyun.lee, iskweon77}@kaist.ac.kr

[2] Jean Oh is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA 15213. E-mail: jeanoh@nrec.ri.cmu.edu

recover the true radiance of the scene, while others are to make the training procedure more efficient. The technical details of the network and training scheme are presented more fully in Sec. III.

Thus, our simultaneous dehazing and depth estimation network outperforms some of the state-of-the-art algorithms in both dehazing and single image depth estimation. In Sec. IV, we demonstrated the robustness and effectiveness of our algorithm via quantitative and qualitative comparison in various scenes.

II. RELATED WORK

A. Single Image Dehazing

Single image dehazing has always been a tricky problem to solve because a lot of information taken from a hazy scene is unknown or distorted. Therefore, various methods, such as those introduced in [5], [6], [7] tried to tackle this problem with handcrafted prior/constraint guidance. With the emergence of deep learning, Cai et al. proposed Dehazenet in [8], introducing an end-to-end CNN network with a novel bilateral rectified linear unit (BReLU) which mainly focused on acquiring a high-quality transmission map. Another algorithm that focuses on obtaining an expressive transmission map was proposed by Ren et al. in [9], via a multi-scale deep neural network.

The appearance and success of the Generative Adversarial Network (GAN), which was first proposed by Goodfellow et al. in [10] to synthesize realistic images by competitive generator/discriminator training via a min-max optimization framework, enabled many other approaches to dehazing. In [11], the authors suggested a joint discriminator based on GAN to incorporate mutual information that the transmission map and dehazed result hold. Zhang et al. [1] proposed a multi-scale image dehazing method that uses a perceptual pyramid deep network consisting of dense blocks and residual blocks. In [12], without the ground truth atmospheric light and transmission map information, a network was proposed to predict all the information. In the loss function, dehaze reconstruction loss and a haze reconstruction loss term were added.

B. Single Image Depth Estimation

Single image depth estimation was and still a difficult problem to solve because there are not many geometric cues in a monocular image. Recently, various methods have been introduced to solve this problem without any geometric cues, thanks to CNN. Eigen et al. [13] suggested a two-stack convolutional neural network, with one predicting the global coarse scale and the other refining local details. Liu et al. combined a continuous conditional random field with the CNN in [14], obtaining sharp boundaries and details in the depth map. Laina et al. [15] developed a very effective deep residual network based on the ResNet [16].

Another line of related work is depth completion from sparse/semi-dense initial depth information. Recently, Ma et al. [2] proposed feeding sparse depth information to an additional channel with an RGB image as an input to a ResNet [16]-based depth predictor. This resulted in an outstanding results, even with a very small number of input depth samples. However, this kind of setup is not valid in a hazy scene, because most of the depth sensors commonly used will fail and thus cannot provide any reliable initial guidance.

III. METHODOLOGY

The overall pipeline of the proposed simultaneous dehazing and depth estimation network is shown in Fig. 2. It contains a shared encoder and four decoders for individual tasks. Each decoder is designed to predict the following: 1) directly-regressed true scene radiance, 2) atmospheric light, 3) transmission map, and 4) depth map, respectively. The technical details of the individual modules and the loss function for the training will be explained thoroughly in the following section.

A. Simultaneous Dehazing and Depth Estimation Network

The encoder of the network follows the structure of the Densely Connected Convolutional Network (DenseNet) [17], with four layers of dense blocks. Between every two dense blocks, there exists a transition block. A transition block is a combination of a 1×1 convolution layer and a 2×2 average pooling layer. From the first dense block to the third transition block, we initialized the network using the pre-trained weight of DenseNet201 [17]. This weight was trained on the ImageNet [18], which is a dataset for image classification. Although image classification and dehazing or depth estimation are different tasks, the rich image representation learned from image classification helps to train the proposed network more effectively. Since we have to preserved some spatial resolution of the encoded image feature, we modified the last dense and transition block from the original DenseNet201 [17]. The network architecture of the encoder is described in detail in Table I.

There are four decoders in our simultaneous dehazing and depth estimation network. These decoders share the same image representation extracted from the encoder. Each decoder outputs predicted directly-regressed true scene radiance J_{direct}, atmospheric light α_{pred}, transmission map $t_{pred}(x)$, and depth map of the scene $D_{pred}(x)$. The overall architecture of each decoder is similar to the encoder, but it has transition blocks with upsampling layers instead of average pooling layers. Transition blocks help the network to effectively reorder and expand the spatial size of the encoded feature. Also, additional residual blocks suggested in [19] and refinement blocks suggested in [1] were added. The residual blocks have two consecutive dense blocks with two 3×3 convolutional layers, which gives the ability to recover more high-frequency information. The refinement blocks takes a dense pyramid-like structure with four different spatial scales of average pooling and upsampling. This helps the network to preserve both the global and local information of the image, where global information refers to high-level scene description, and local information means semantic features and spatial location in the image.

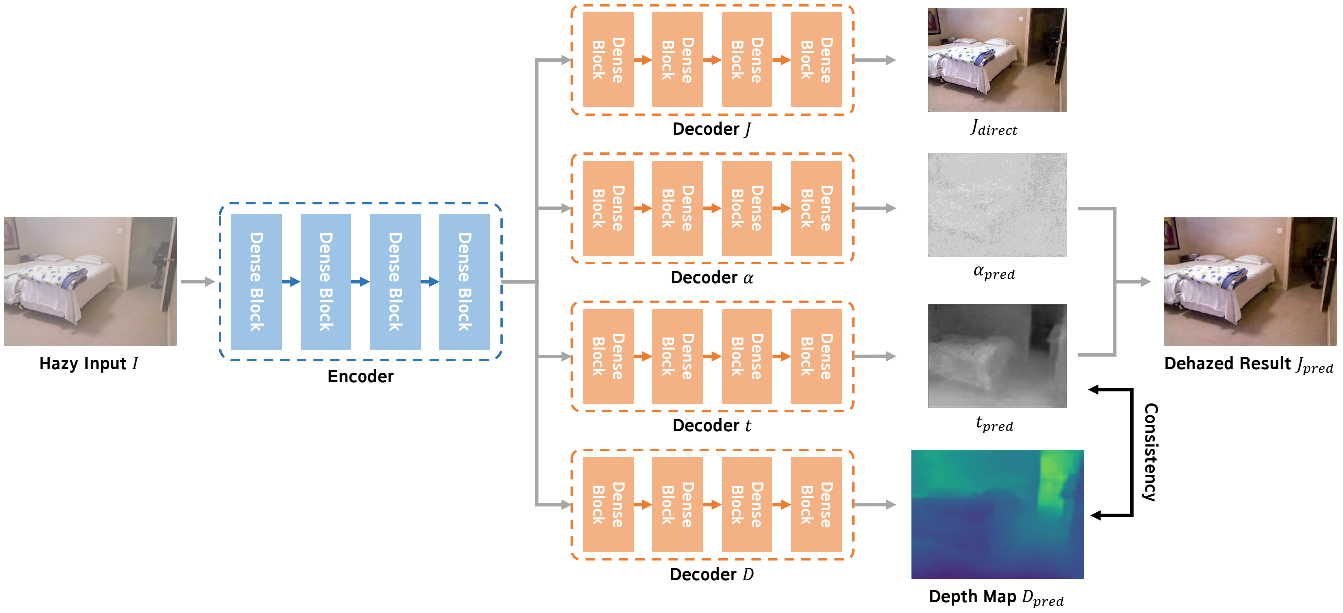


Fig. 2: Overview of the proposed simultaneous dehazing and depth estimation network.

Layers	Base	Dense Block 1	Transition Block 1	Dense Block 2	Transition Block 2
Structure	7×7 conv 3×3 max pool	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 6$	1×1 conv 2×2 avg pool	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 12$	1×1 conv 2×2 avg pool
Output Size	$128\times128\times64$	$128\times128\times256$	$64\times64\times512$	$64\times64\times512$	$32\times32\times256$

Layers	Dense Block 3	Transition Block 3	Dense Block 4	Transition Block 4
Structure	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 48$	1×1 conv 2×2 avg pool	1×1 conv 3×3 conv	1×1 conv 2×2 upsample
Output Size	$32\times32\times1792$	$16\times16\times896$	$16\times16\times1344$	$32\times32\times256$

TABLE I: Detailed architecture of the encoder.

Layers	Refinement Block 9	Refinement Block 10	Refinement Block 11	Refinement Block 12	Final
Structure	32×32 avg pool 1×1 conv 32×32 upsample	16×16 avg pool 1×1 conv 16×16 upsample	8×8 avg pool 1×1 conv 8×8 upsample	4×4 avg pool 1×1 conv 4×4 upsample	$\begin{bmatrix}3\times3 \text{ conv}\\ 7\times7 \text{ conv}\end{bmatrix}\times 2$

TABLE II: Detailed architecture of refinement blocks.

The detailed architecture of each refinement block is shown in Table II.

The outputs of all four refinement blocks are then concatenated and fed into the final dense block to predict the information for which each decoder was designed. The final dehazed result that we used is the dehazed image reconstructed using estimated α and t. The output of Decoder J is only used to help the whole network's training. This idea was first proposed in [20], and was proven to be effective.

B. Loss Function

Our loss function E is a linear combination of reconstruction loss E_R, atmospheric light loss E_α, transmission map loss E_t, depth loss E_D, and depth-transmission consistency loss E_C as follows:

$$E = E_R + E_\alpha + E_t + E_D + \lambda E_C, \tag{4}$$

where the loss weight λ was set empirically. Each term will be described thoroughly in the following subsections, and we denote $\| \cdot \|_2$ as L_2-norm.

1) Reconstruction loss: We define two types of reconstruction loss for effective training. One is dehazing loss, and the other is rehazing loss.

Dehazing loss is L_2-norm between the dehazed image and the true radiance J. There are two terms for dehazing loss, namely, the dehazing loss between reconstructed dehazed image J_{recon} and the original clean image J, and another dehazing loss between directly-regressed dehazed result J_{direct} and the true radiance J.

With the predicted transmission map, atmospheric light, and the original image, we can reconstruct the hazy image I_{recon} as well, following Eq. (1). Rehazing loss is L_2 norm as well, between the rehazed image I_{recon} and the input hazy image I.

The final reconstruction loss can be formally expressed as

$$\begin{aligned} E_R =& \| J_{recon} - J \|_2 + \| J_{direct} - J \|_2 \\ &+ \| I_{recon} - I \|_2, \end{aligned} \tag{5}$$

2) Atmospheric light, transmission map & depth loss: In a CNN-based algorithm, a dataset with input/ground truth pairs is often inevitable in network training. However, there are not a lot of options for dehazing. It is even more difficult to prepare a dataset with ground truth atmospheric light, a transmission map, and depth information. In this work, we use the NYU-Depth V2 dataset [21] to synthesize original/hazy image pairs with ground truth atmospheric light, a transmission map, and a depth map. A detailed explanation of synthetic haze dataset generation will be discussed in Sec. IV-A.

Atmospheric light, a transmission map, and a depth map are individual outputs from the decoders of our network. Atmospheric light and the transmission map are trained with L_2 loss as in Eq. (6). The training loss for depth map estimation is L_2 loss between the predicted depth map D_{pred} and the ground truth depth D_{gt}.

Each loss term can be written as

$$\begin{aligned} E_\alpha &= \| \alpha_{pred} - \alpha_{gt} \|_2 \\ E_t &= \| t_{pred} - t_{gt} \|_2 \\ E_D &= \| D_{pred} - D_{gt} \|_2, \end{aligned} \tag{6}$$

where α_{gt} and t_{gt} are the ground truth values of atmospheric light α and transmission map t.

3) Depth-transmission consistency loss: As Eq. (3) notes, the transmission map and the depth information have a correlation. To be more specific, the transmission map inside a log function ln is a scaled version of the depth map. To give the decoder t and the decoder D guidance regarding this correlation, we modeled depth-transmission consistency loss.

First, we divide the log-scaled transmission map predicted from the decoder t, t_{pred}, with the predicted depth map D_{pred}, and name it consistency term, noting as C. This term should have uniform value over all pixels because C equals to β times depth-normalizing constant, following Eq. (3). Therefore, a standard deviation of C should be 0 for each transmission map/depth map pair.

We simply use the standard deviation of C as loss and add it to the total training loss. Thus, depth-transmission consistency loss is given by

$$E_C = \| std(C) \|_2, \tag{7}$$

with C being $\ln t_{pred}/D_{pred}$, and $std(\cdot)$ being standard deviation.

IV. EXPERIMENTS

A. Datasets & Training

As introduced in other dehazing algorithms via deep learning, we synthetically generated the dataset for dehazing. The set has input/ground truth tuples which contain the following: 1) a hazy image, 2) ground truth atmospheric light, 3) a ground truth transmission map, 4) a ground truth depth map, and 5) true radiance. We used the fully-labeled NYU-Depth V2 dataset [21], which contains 1449 pairs of RGB and depth images taken from indoor scenes. It was not possible to use the raw dataset of the NYU-Depth V2 dataset, since the depth map was retrieved by Kinect. The raw depth information from Kinect has sparsity, which would make the synthesized transmission map and hazy image noisy.

For the training set, 1000 randomly chosen pairs were used. In generating atmospheric light, a random value between 0.5 and 1 was chosen as α and applied uniformly for all pixels. This is due to the assumption that the global lighting would be the same in an indoor scene. For the transmission map, the depth map normalized by the maximum value of each scene was used, and the scattering coefficient β was selected randomly between 0.4 and 1.6. Each pair was used for four different combinations of α and β. Therefore, the total number of items in the training set was 4000. Similarly, a test set was generated from 100 pairs, which were not selected in the training set, and it was sampled with four pairs of α and β, which makes the total test set 400 pairs.

The original image resolution of the NYU-Depth V2 dataset is 680×480. When fed to the network, all the information was resized to 512×512, and this resolution was also used in quantitative evaluation for dehazing. For depth estimation evaluation, the resolution was resized back to 680×480, for easier comparison with other algorithms. We applied some data augmentation as well, with random scaling by the factor within $[1, 1.5]$, random rotation by degree within $[-5, 5]$, and a 50% chance of horizontal flips.

In the training procedure, the whole network was trained in an end-to-end manner, and every decoder was trained from scratch. We used the ADAM optimizer with an initial learning rate of $1e^{-4}$ and a weight decay of $1e^{-8}$. The learning rate was reduced to 70% for every 25 epochs. We used a batch size of 4 and trained for 100 epochs. Our network was implemented by using PyTorch on a machine equipped with four NVidia 1080 Ti GPUs.

B. Ablation Study

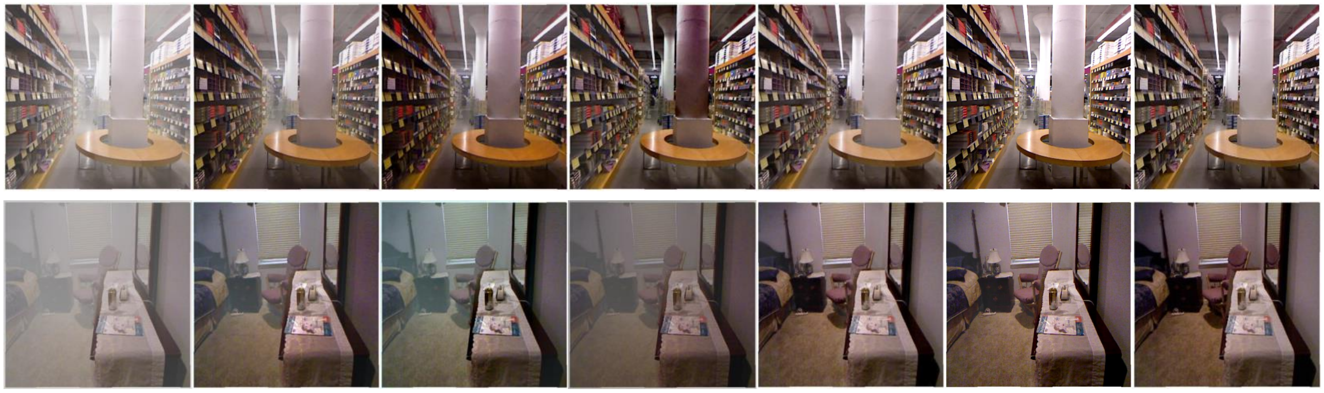
To demonstrate the effectiveness of each module of our network, we performed ablation studies. We trained the network with and without some components of the network, respectively. The results are shown in Table III.

As seen in Table III, the addition of Decoder D itself significantly improved the performance of the network. This proves that when the correlation between the transmission map and the depth map was given in the training scheme, the performance of the network improved. Depth-transmission consistency also helped the network to acquire a more accurate and refined result.

C. Comparison with State-of-the-arts

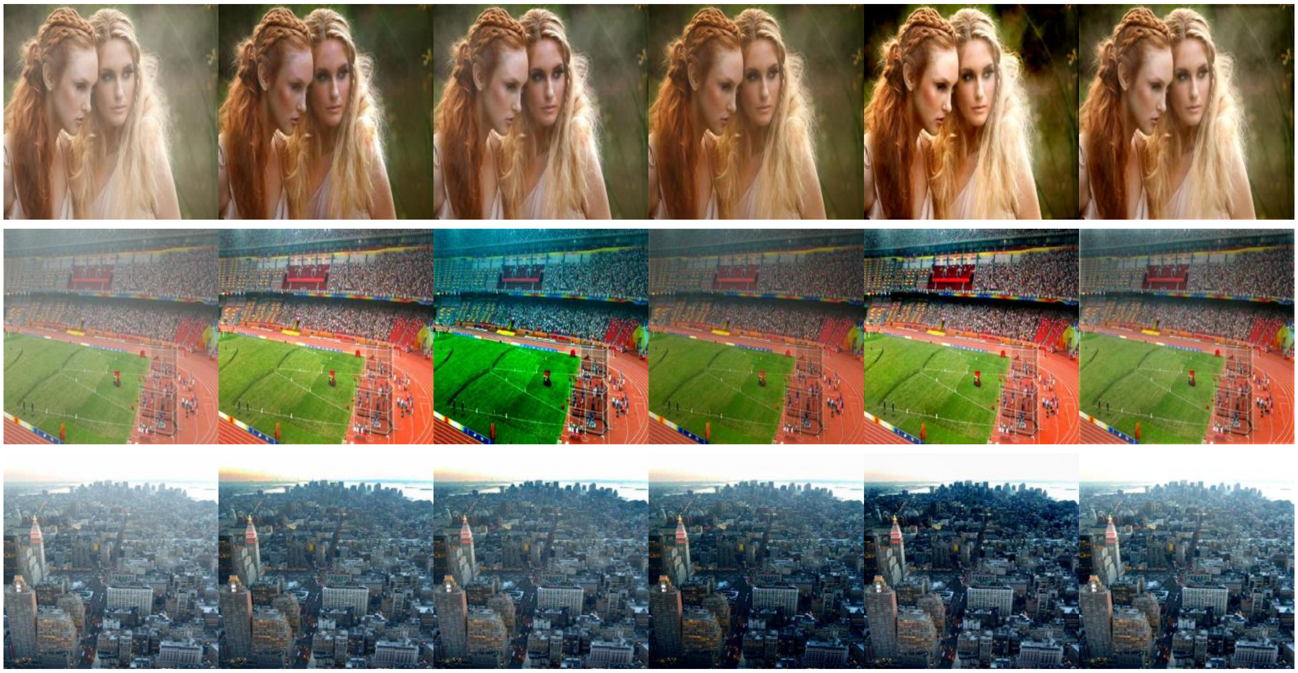
To demonstrate the robustness of our algorithm, we compare both our qualitative and quantitative results to those of other state-of-the-art methods. For dehazing comparison, we selected He et al. [6], Ren et al. [9], Li et al. [22] and Zhang et al. [1]. Note that Ren et al. [9], Li et al. [22] and Zhang et al. [1] are CNN-based algorithms, while He et al. [6] solves the problem with a dark channel prior. All CNN-based algorithms were trained with the same dataset as the proposed network.

For quantitative evaluation and comparison, we use the structural similarity index (SSIM) [23], which can be calcu-



(a) Input (b) He et al. [6] (c) Ren et al. [9] (d) Li et al. [22] (e) Zhang et al. [1] (f) Ours (g) GT

Fig. 3: Dehazing results evaluated on hazed NYU-Depth V2 dataset.



(a) Input (b) He et al. [6] (c) Ren et al. [9] (d) Li et al. [22] (e) Zhang et al. [1] (f) Ours

Fig. 4: Dehazing results evaluated on real-world images.

Model	Dehazed Image (SSIM)	Transmission Map (SSIM)	Depth (RMSE)
Baseline	0.8827	0.8203	N/A
Baseline + Decoder D	0.9410	0.9403	0.653
Baseline + Decoder D + Depth-Transmission Consistency	**0.9547**	**0.9801**	**0.622**

TABLE III: Ablation study: performance changes with and without each component of our network.

lated as

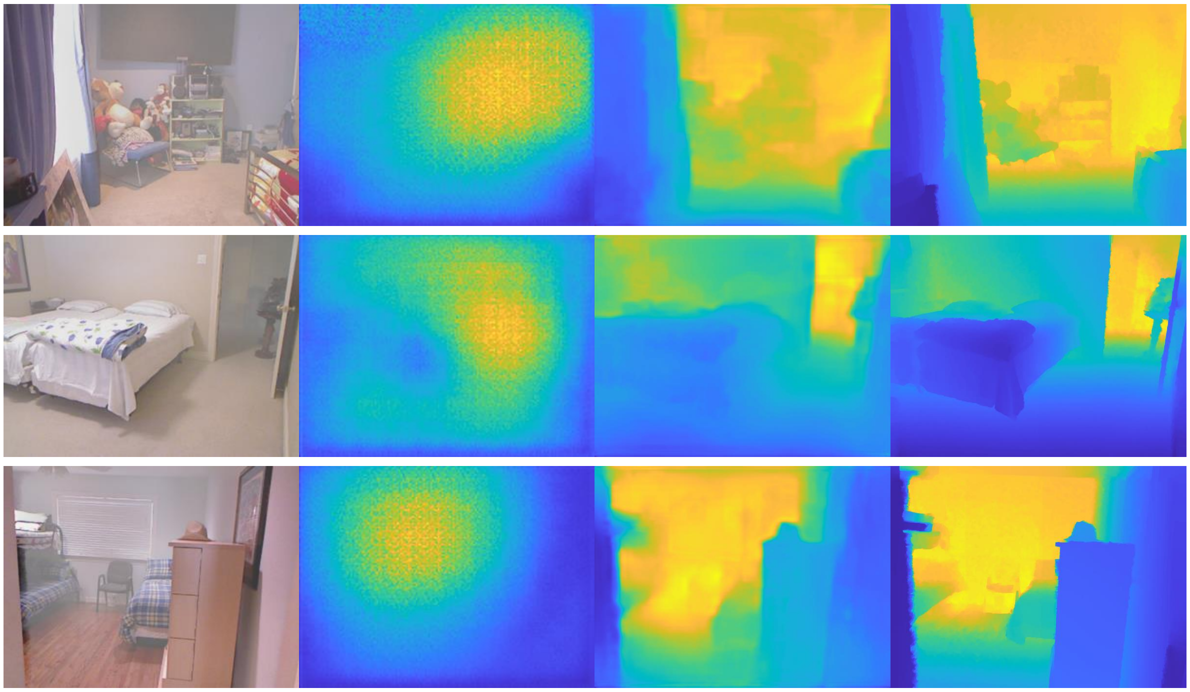
$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{8}$$

where μ and σ^2 are the average and variance of the individual input, and σ_{xy} is the covariance of x and y. The SSIM metric is commonly used in image enhancement and computational photography because the structural similarity may refer to local and detailed information of the image.

The quantitative results obtained using the test dataset generated with NYU-Depth V2 dataset [21] are shown in Table IV. Although the algorithm with the best dehazing result in the SSIM metric was Zhang et al. [1], the difference between Zhang et al. [1] and ours is relatively small in

	[6]	[9]	[22]	[1]	Ours
Image	0.8642	0.8203	0.8842	**0.9560**	0.9547
Transmission	0.8739	N/A	N/A	0.9776	**0.9801**

TABLE IV: Quantitative SSIM evaluation on hazed NYU-Depth V2 dataset.



(a) Input (b) Ma et al. [2] (c) Ours (d) GT

Fig. 5: Qualitative depth estimation results evaluated on hazed NYU-Depth V2 dataset.

comparison to other methods. Our network performed the best in transmission map estimation. Both Ren et al. [9] and Li et al. [22] do not output a transmission map from the network, so their results were neglected.

Some of the qualitative result in dehazing tasks are shown in Fig. 3 and Fig. 4. Fig. 3 shows the qualitative results obtained using the synthetic test dataset. It is visually clear that our results are better than those of He et al. [6], Ren et al. [9] and Li et al. [22]. In comparison to Zhang et al. [1], the result seems similar.

Fig. 4 shows the qualitative evaluation on real hazy images collected from previous methods [9], [8]. In every example, our method recovered more detailed information of the image, while preserving the natural color. For example, the dehazed result of Zhang et al. [1] on the first sample or the result of Li et al. [22] on the third sample shows too much of an artificial contrast and color shift.

For depth estimation, we chose the RGB version of Ma et al. [2]. We used three kinds of metrics for quantitative evaluation, namely, root mean square error (RMSE), mean absolute relative error (Rel), and δ_n. Here, δ_n refers to the portion of pixels for which the error between the predicted value and the ground truth is below a certain threshold; δ_n can be formulated as

$$\mathrm{delta}_n(x,y) = \frac{n(\{x : \max(\frac{x}{y}, \frac{y}{x}) < 1.25^n\})}{n(y)}, \tag{9}$$

with $n(\cdot)$ as the cardinality of a set.

The results are presented in Table V. The qualitative results are also shown in Fig. 5. These results show that our algorithm outperforms Ma et al. [2] in every metric. With a general approach to single image depth estimation, it is hard to get the good image representation that would help the network to configure the depth information of the scene. However, in our method, by simultaneously trying to solve both dehazing and depth estimation, the depth quality is significantly improved.

	RMSE	Rel	δ_1	δ_2	δ_3
Ma et al. [2]	0.811	0.250	0.495	0.820	0.963
Ours	**0.622**	**0.219**	**0.635**	**0.898**	**0.981**

TABLE V: Quantitative depth evaluation on hazed NYU-Depth V2 dataset (RMSE & Rel: lower the better, δ_n: higher the better).

V. CONCLUSIONS

In this work, we addressed the difficulty of CNN-based dehazing as well as the depth estimation from hazy scenes. By fully utilizing the principal of the haze model, we propose a CNN-based simultaneous dehazing and depth estimation network. Our network was trained with multi-tasking loss, helping the decoders be guided to each other. Moreover, with the depth-transmission consistency loss, we maximized the correlation between the decoders and were able to output the best result. We showed that our algorithm performs in a promising manner, in both dehazing and depth estimation, achieving a performance comparable to or better than other state-of-the-art algorithms.

Acknowledgement This work is in part supported by the U.S. Air Force Office of Scientific Research under award number FA2386-17-1-4660.

References

[1] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 3194–3203.

[2] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[3] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 820–827 vol.2, 1999.

[4] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 06, pp. 713–724, jun 2003.

[5] R. T. Tan, "Visibility in bad weather from a single image," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[6] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, Dec 2011.

[7] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 13:1–13:14, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2651362

[8] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, Nov 2016.

[9] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 154–169.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[11] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[12] T. Guo, X. Li, V. Cherukuri, and V. Monga, "Dense scene information estimation network for dehazing," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[13] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2366–2374. [Online]. Available: http://papers.nips.cc/paper/5539-depth-map-prediction-from-a-single-image-using-a-multi-scale-deep-network.pdf

[14] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[15] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth International Conference on 3D Vision (3DV)*, Oct 2016, pp. 239–248.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[19] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1646–1654.

[20] T. Guo, X. Li, V. Cherukuri, and V. Monga, "Dense scene information estimation network for dehazing," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

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

[22] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.