Densely Connected Pyramid Dehazing Network

He Zhang Vishal M. Patel
Department of Electrical and Computer Engineering
Rutgers University, Piscataway, NJ 08854
{he.zhang92,vishal.m.patel}@rutgers.edu

Abstract

We propose a new end-to-end single image dehazing method, called Densely Connected Pyramid Dehazing Network (DCPDN), which can jointly learn the transmission map, atmospheric light and dehazing all together. The end-to-end learning is achieved by directly embedding the atmospheric scattering model into the network, thereby ensuring that the proposed method strictly follows the physics-driven scattering model for dehazing. Inspired by the dense network that can maximize the information flow along features from different levels, we propose a new edge-preserving densely connected encoder-decoder structure with multi-level pyramid pooling module for estimating the transmission map. This network is optimized using a newly introduced edge-preserving loss function. To further incorporate the mutual structural information between the estimated transmission map and the dehazed result, we propose a joint-discriminator based on generative adversarial network framework to decide whether the corresponding dehazed image and the estimated transmission map are real or fake. An ablation study is conducted to demonstrate the effectiveness of each module evaluated at both estimated transmission map and dehazed result. Extensive experiments demonstrate that the proposed method achieves significant improvements over the state-of-the-art methods. Code will be made available at: https://github.com/hezhangsprinter

1. Introduction

Under severe hazy conditions, floating particles in the atmosphere such as dusk and smoke greatly absorb and scatter the light, resulting in degradations in the image quality. These degradations in turn may affect the performance of many computer vision systems such as classification and detection. To overcome the degradations caused by haze, image and video-based haze removal algorithms have been proposed in the literature [33, 5, 42, 3, 13, 21, 27, 51, 24, 57, 8, 10, 9, 34].

Figure 1: Sample image dehazing result using the proposed DCPDN method. Left: Input hazy image. Right: Dehazed result.

The image degradation (atmospheric scattering model) due to the presence of haze is mathematically formulated as

$$I(z) = J(z)t(z) + A(z)(1 - t(z)), \quad (1)$$

where I is the observed hazy image, J is the true scene radiance, A is the global atmospheric light, indicating the intensity of the ambient light, t is the transmission map and z is the pixel location. Transmission map is the distance-dependent factor that affects the fraction of light that reaches the camera sensor. When the atmospheric light A is homogeneous, the transmission map can be expressed as $t(z) = e^{-\beta d(z)}$, where β represents attenuation coefficient of the atmosphere and d is the scene depth. In single image dehazing, given I, the goal is to estimate J.

It can be observed from Eq. 1 that there exists two important aspects in the dehazing process: *(1) accurate estimation of transmission map,* and *(2) accurate estimation of atmospheric light*. Apart from several works that focus on estimating the atmospheric light [4, 40], most of the other algorithms concentrate more on the accurate estimation of the transmission map and they leverage empirical rule in estimating the atmospheric light [13, 29, 33, 41]. This is mainly due to the common belief that good estimation of transmission map will lead to better dehazing. These methods can be broadly divided into two main groups: prior-based methods and learning-based methods. Prior-based methods often leverage different priors in characterizing the transmission map such as dark-channel prior [13], contrast color-lines [10] and haze-line prior [3], while learning-based methods, such as those based on convolutional neural networks (CNNs), attempt to learn the transmission map di-

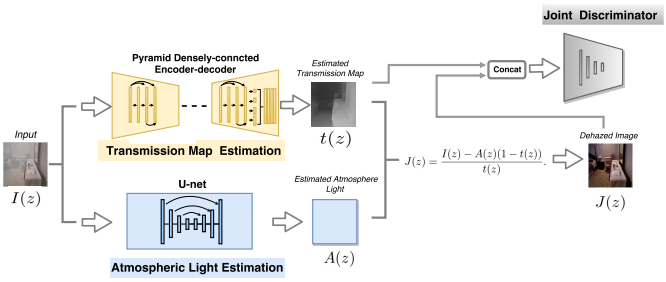


Figure 2: An overview of the proposed DCPDN image dehazing method. DCPDN consists of four modules: 1. Pyramid densely connected transmission map estimation net. 2. Atmospheric light estimation net. 3. Dehazing via Eq2. 4. Joint discriminator. We first estimate the transmission map using the proposed pyramid densely-connected transmission estimation net, followed by prediction of atmospheric light using the U-net structure. Finally, using the estimated transmission map and the atmospheric light we estimate the dehazed image via Eq. 2.

rectly from the training data [42, 33, 5, 51, 24]. Once the transmission map and the atmospheric light are estimated, the dehazed image can be recovered as follows

$$\hat{J}(z) = \frac{I(z) - \hat{A}(z)(1 - \hat{t}(z))}{\hat{t}(z)}. \quad (2)$$

Though tremendous improvements have been made by the learning-based methods, several factors hinder the performance of these methods and the results are far from optimal. This is mainly because: *1. Inaccuracies in the estimation of transmission map translates to low quality dehazed result. 2. Existing methods do not leverage end-to-end learning and are unable to capture the inherent relation among transmission map, atmospheric light and dehazed image. The disjoint optimization may hinder the overall dehazing performance.* Most recently, a method was proposed in [24] to jointly optimize the whole dehazing network. This was achieved by leveraging a linear transformation to embed both the transmission map and the atmospheric light into one variable and then learning a light-weight CNN to recover the clean image.

In this paper, we take a different approach in addressing the end-to-end learning for image dehazing. In particular, we propose a new image dehazing architecture, called Densely Connected Pyramid Dehazing Network (DCPDN), that can be jointly optimized to estimate transmission map, atmospheric light and also image dehazing simultaneously by following the image degradation model Eq. 1 (see Fig. 2). In other words, the end-to-end learning is achieved by embedding Eq. 1 directly into the network via the math operation modules provided by the deep learning framework. However, training such a complex network (with three different tasks) is very challenging. To ease the training process and accelerate the network convergence, we leverage a stage-wise learning technique in which we first progressively optimize each part of the network and then jointly optimize the entire network. To make sure that the estimated transmission map preserves sharp edges and avoids halo artifacts when dehazing, a new edge-preserving loss function is proposed in this paper based on the observation that gradient operators and first several layers of a CNN structure can function as edge extractors. Furthermore, a densely connected encoder-decoder network with multi-level pooling modules is proposed to leverage features from different levels for estimating the transmission map. To exploit the structural relationship between the transmission map and the dehazed image, a joint discriminator-based generative adversarial network (GAN) is proposed. The joint discriminator distinguishes whether a pair of estimated transmission map and dehazed image is a real or fake pair. To guarantee that the atmospheric light can also be optimized within the whole structure, a U-net [35] is adopted to estimate the homogeneous atmospheric light map. Shown in Fig. 1 is a sample dehazed image using the proposed method.

This paper makes the following contributions:

- A novel end-to-end jointly optimizable dehazing network is proposed. This is enabled by embedding Eq. 1 directly into the optimization framework via math operation modules. Thus, it allows the network to estimate the transmission map, atmospheric light and dehazed image jointly. The entire network is trained by a stage-wise learning method.
- An edge-preserving pyramid densely connected encoder-decoder network is proposed for accurately estimating the transmission map. Further, it is optimized via a newly proposed edge-preserving loss function.
- As the structure of the estimated transmission map and the dehazed image are highly correlated, we leverage a joint discriminator within the GAN framework to determine whether the paired samples (i.e. transmission map and dehazed image) are from the data distribution or not.
- Extensive experiments are conducted on two synthetic datasets and one real-world image dataset. In addition, comparisons are performed against several recent state-of-the-art approaches. Furthermore, an ablation study is conducted to demonstrate the improvements obtained by different modules in the proposed network.

2. Related Work

Single Image Dehazing. Single image dehazing is a highly ill-posed problem. Various handcrafted prior-based and learning-based methods have been developed to tackle this problem.

Handcrafted Prior-based: Fattal [9] proposed a physically-grounded method by estimating the albedo of the scene. As the images captured from the hazy conditions always lack color contrast, Tan [41] *et al*. proposed a patch-based contrast-maximization method. In [22], Kratz and Nishino proposed a factorial MRF model to estimate the albedo and depths filed. Inspired by the observations that outdoor objects in clear weather have at least one color channel that is significantly dark, He. *et al.* in [13] proposed a dark-channel model to estimate the transmission map. More recently, Fattal [10] proposed a color-line method based on the observation that small image patches typically exhibit a one-dimensional distribution in the RGB color space. Similarly, Berman *et al.* [3] proposed a non-local patch prior to characterize the clean images.

Learning-based: Unlike some of the above mentioned methods that use different priors to estimate the transmission map, Cai *et al.* [5] introduce an end-to-end CNN network for estimating the transmission with a novel BReLU unit. More recently, Ren *et al.* [33] proposed a multi-scale deep neural network to estimate the transmission map. One of the limitations of these methods is that they limit their capabilities by only considering the transmission map in their CNN frameworks. To address this issue, Li. *et al* [24] proposed an all-in-one dehazing network, where a linear transformation is leveraged to encode the transmission map and the atmospheric light into one variable. Most recently, several benchmark datasets of both synthetic and real-world hazy images for dehazing problems are introduced to the community [53, 25].

Generative Adversarial Networks (GANs). The notion of GAN was first proposed by Goodfellow *et al.* in [12] to synthesize realistic images by effectively learning the distribution of the training images via a game theoretic min-max optimization framework. The success of GANs in synthesizing realistic images has led researchers to explore the adversarial loss for various low-level vision applications such as text-to-image synthesis[32, 52, 54, 6], image-image translation [18, 28, 46] and other applications [23, 50, 55, 58, 45, 38, 44, 31]. Inspired by the success of these methods in generating high-quality images with fine details, we propose a joint discriminator-based GAN to refine the estimated transmission map and dehazed image.

3. Proposed Method

The proposed DCPDN network architecture is illustrated in Fig. 2 which consists of the following four modules: 1) Pyramid densely connected transmission map estimation net, 2) Atmosphere light estimation net, 3) Dehazing via Eq. 2, and 4) Joint discriminator. In what follows, we explain these modules in detail.

Pyramid Densely Connected Transmission Map Estimation Network. Inspired by the previous methods that use multi-level features for estimating the transmission map [33, 5, 42, 1, 24], we propose a densely connected encoder-decoder structure that makes use of the features from multiple layers of a CNN, where the dense block is used as the basic structure. The reason to use dense block lies in that it can maximize the information flow along those features and guarantee better convergence via connecting all layers. In addition, a multi-level pyramid pooling module is adopted to refine the learned features by considering the 'global' structural information into the optimization [56]. To leverage the pre-defined weights of the dense-net [15], we adopt the first *Conv* layer and the first three *Dense-Blocks* with their corresponding down-sampling operations *Transition-Blocks* from a pre-trained dense-net121 as our encoder structure. The feature size at end of the encoding part is 1/32 of the input size. To reconstruct the transmission map into the original resolution, we stack five dense blocks with the refined up-sampling *Transition-Blocks* [19, 59] as the decoding module. In addition, concatenations are employed with the features corresponding to the same dimension.

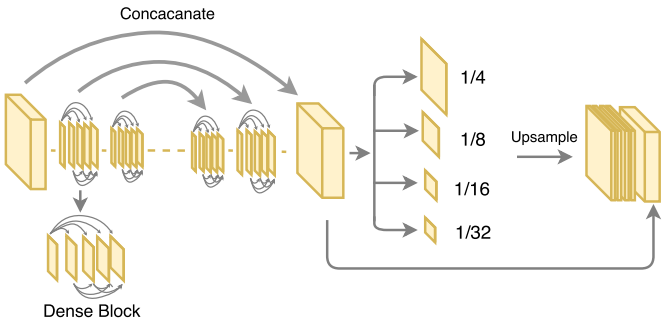


Figure 3: An overview of the proposed pyramid densely connected transmission map estimation network.

Even though the proposed densely connected encoder-decoder structure combines different features within the network, the result from just densely connected structure still lack of the 'global' structural information of objects with different scales. One possible reason is that the features from different scales are not used to directly estimate the final transmission map. To efficiently address this issue, a multi-level pyramid pooling block is adopted to make sure that features from different scales are embedded in the final result. This is inspired by the use of global context information in classification and segmentation tasks [56, 48, 14]. Rather than taking very large pooling size to capture more global context information between different objects [56], more 'local' information to characterize the 'global' structure of each object is needed. Hence, a four-level pooling operation with pooling sizes 1/32, 1/16,

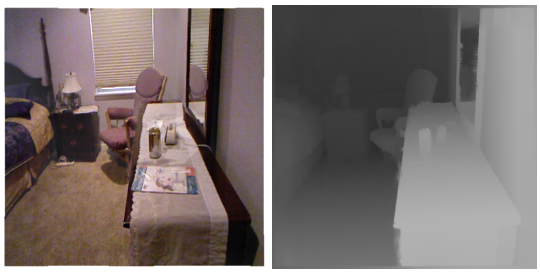


Figure 4: Left: a dehazed image. Right: The transmission map used to produce a hazy image from which the dehazed image on the left was obtained.

1/8 and 1/4 is adopted. Then, all four level features are up-sampling to original feature size and are concatenated back with the original feature before the final estimation. Fig 3 gives an overview of the proposed pyramid densely connected transmission map estimation network.

Atmospheric Light Estimation Network. Following the image degradation model Eq.; 1, we assume that the atmospheric light map A is homogeneous [13, 5]. Similar to previous works, the predicted atmospheric light A is uniform for a given image. In other words, the predicted A is a 2D-map, where each pixel $A(z)$ has the same value (eg. $A(z) = c,\ c$ is a constant). As a result, the ground truth A is of the same feature size as the input image and the pixels in A are filled with the same value. To estimate the atmospheric light, we adopt a 8-block U-net [35] structure, where the encoder is composed of four *Conv-BN-Relu* blocks and the decoder is composed of symmetric *Dconv-BN-Relu* block [1].

Dehazing via Eq. 2. To bridge the relation among the transmission map, the atmospheric light and the dehazed image and to make sure that the whole network structure is jointly optimized for all three tasks, we directly embed (2) into the overall optimization framework. An overview of the entire DCPDN structure is shown in Fig 1.

3.1. Joint Discriminator Learning

Let G_t and G_d denote the networks that generate the transmission map and the dehazed result, respectively. To refine the output and to make sure that the estimated transmission map $G_t(I)$ and the dehazed image $G_d(I)$ are indistinguishable from their corresponding ground truths t and J, respectively, we make use of a GAN [12] with novel joint discriminator.

It can be observed from (1) and also Fig. 4 that the structural information between the estimated transmission map $\hat{t} = G_t(I)$ and the dehazed image $\hat{J}$ are highly correlated. Hence, in order to leverage the dependency in structural information between these two modalities, we introduce a joint discriminator to learn a joint distribution to decide whether the corresponding pairs (*transmission map*, *dehazed image*) are real or fake. By leveraging the joint distribution optimization, the structural correlation between them can be better exploited. Similar to previous works, the predicted air-light A is uniform for a given image. In other words, the predicted air-light A is a 2D-map, where each pixel $A(z)$ has the same value (eg. $A(z) = c,\ c$ is a constant). We propose the following joint-discriminator based optimization

$$\begin{aligned}\min_{G_t,G_d}\max_{D_joint} \quad & \mathbb{E}_{I\sim p_{data(I)}}[\log(1 - D_joint(G_t(I)))]+ \\ & \mathbb{E}_{I\sim p_{data(I)}}[\log(1 - D_joint(G_d(I)))]+ \\ & \mathbb{E}_{t,J\sim p_{data(t,J)}}[\log D_{joint}(t, J))].\end{aligned} \tag{3}$$

In practice, we concatenate the dehazed image with the estimated transmission map as a pair sample and then feed it into the discriminator.

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

Figure 5: Feature visualization for gradient operator and low-level features. (a) Input transmission map. (b) Horizontal gradient output. (c) Vertical gradient output. (d) and (e) are visualization of two feature maps from relu1_2 of VGG-16 [37].

3.2. Edge-preserving Loss

It is commonly acknowledged that the Euclidean loss ($L2$ loss) tends to blur the final result. Hence, inaccurate estimation of the transmission map with just the $L2$ loss may result in the loss of details, leading to the halo artifacts in the dehazed image [16]. To efficiently address this issue, a new edge-preserving loss is proposed, which is motivated by the following two observations. 1) Edges corresponds to the discontinuities in the image intensities, hence it can be characterized by the image gradients. 2) It is known that low-level features such as edges and contours can be captured in the shallow (first several) layers of a CNN structure [47]. In other words, the first few layers function as an edge detector in a deep network. For example, if the transmission map is fed into a pre-defined VGG-16 [37] model and then certain features from the output of layer relu1_2 are visualized, it can be clearly observed that the edge information being preserved in the corresponding feature maps (see Fig. 5).

[1] Con: Convolution, BN: Batch-normalization [17] and Dconv: Deconvolution (transpose convolution).

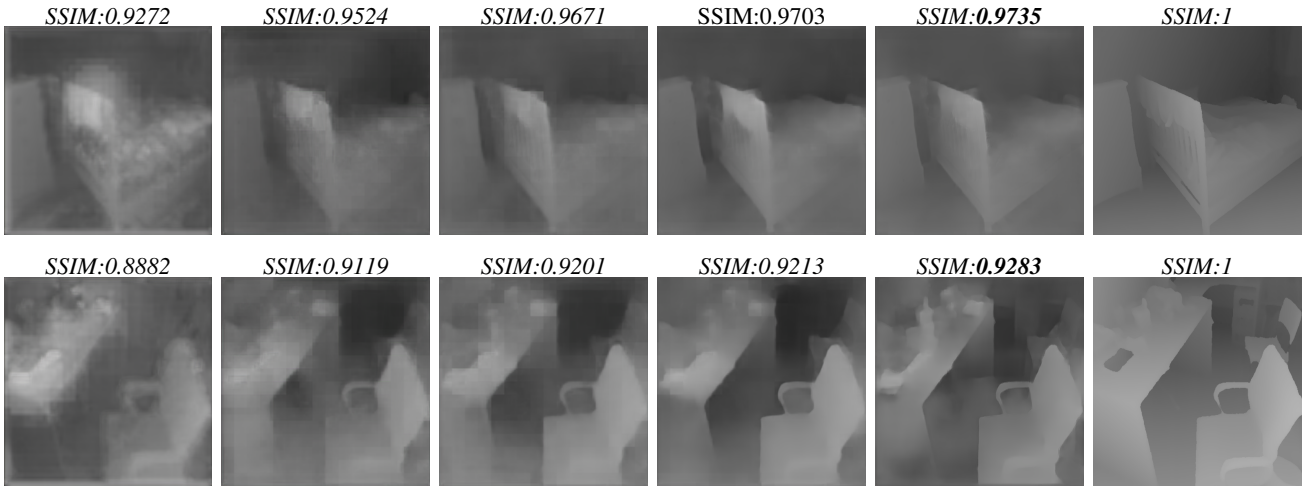


Figure 6: Transmission map estimation results using different modules. (a) DED; (b). DED-MLP; (c).DED-MLP-GRA; (d). DED-MLP-EP; (e). DCPDN; (f) Target. It can be observed that the multi-level pooling module is able to refine better global structure of objects in the image (observed from (a) and (b)), the edge-preserving loss can preserve much sharper edges (comparing (b), (c) and (d)) and the final joint-discriminator can better refine the detail for small objects (comparing (d) and (e)).

Based on these observations and inspired by the gradient loss used in depth estimation [43, 26] as well as the use of perceptual loss in low-level vision tasks [20, 49], we propose a new edge-preserving loss function that is composed of three different parts: $L2$ loss, two-directional gradient loss, and feature edge loss, defined as follows

$$L^E = \lambda_{E,l_2} L_{E,l_2} + \lambda_{E,g} L_{E,g} + \lambda_{E,f} L_{E,f}, \quad (4)$$

where L^E indicates the overall edge-preserving loss, L_{E,l_2} indicates the L2 loss, $L_{E,g}$ indicates the two-directional (horizontal and vertical) gradient loss and $L_{E,f}$ is the feature loss. $L_{E,g}$ is defined as follows

$$\begin{aligned} L_{E,g} = &\sum_{w,h} \|(H_x(G_t(I)))_{w,h} - (H_x(t))_{w,h}\|_2 \\ &+ \|(H_y(G_t(I)))_{w,h} - (H_y(t))_{w,h}\|_2, \end{aligned} \quad (5)$$

where H_x and H_y are operators that compute image gradients along rows (horizontal) and columns (vertical), respectively and $w \times h$ indicates the width and height of the output feature map. The feature loss is defined as

$$\begin{aligned} L_{E,f} = &\sum_{c_1,w_1,h_1} \|(V_1(G_t(I)))_{c_1,w_1,h_1} - (V_1(t))_{c_1,w_1,h_1}\|_2 \\ &+ \sum_{c_2,w_2,h_2} \|(V_2(G_t(I)))_{c_2,w_2,h_2} - (V_2(t))_{c_2,w_2,h_2}\|_2, \end{aligned} \quad (6)$$

where V_i represents a CNN structure and c_i, w_i, h_i are the dimensions of the corresponding low-level feature in V_i. As the edge information is preserved in the low-level features, we adopt the layers before relu1-1 and relu2-1 of VGG-16 [37] as the edge extractors V_1 and V_2, respectively. Here, $\lambda_{E,l_2}, \lambda_{E,g}$, and $\lambda_{E,f}$ are weights to balance the loss function.

3.3. Overall Loss Function

The proposed DCPDN architecture is trained using the following four loss functions

$$L = L^t + L^a + L^d + \lambda_j L^j, \quad (7)$$

where L^t is composed of the edge-preserving loss L^E, L^a is composed of the traditional $L2$ loss in predicting the atmospheric light and L^d represents the dehazing loss, which is also composed of the $L2$ loss only. L^j, which is denoted as the joint discriminator loss [2], is defined as follows

$$L^j = -\log(D_{joint}(G_t(I)) - \log(D_{joint}(G_d(I)). \quad (8)$$

Here λ_j is a constant.

3.4. Stage-wise Learning

During experiments, we found that directly training the whole network from scratch with the complex loss Eq. 7 is difficult and the network converges very slowly. A possible reason may be due to the gradient diffusion caused by different tasks. For example, gradients from the de-hazed image loss may 'distract' the gradients from the loss of the transmission map initially, resulting in the slower convergence. To address this issue and to speed up the training, a stage-wise learning strategy is introduced, which has been

[2] To address the vanishing gradients problem for the generator, we also minimize (8) rather than the first two rows in (3) [12, 11].

Figure 7: Dehazing results from the synthetic test datasets **TestA** (first row) and **TestB** (second row).

used in different applications such as multi-model recognition [7] and feature learning [2]. Hence, the information in the training data is presented to the network gradually. In other words, different tasks are learned progressively. Firstly, we optimize each task separately by not updating the other task simultaneously. After the 'initialization' for each task, we fine-tune the whole network all together by optimizing all three tasks jointly.

4. Experimental Results

In this section, we demonstrate the effectiveness of the proposed approach by conducting various experiments on two synthetic datasets and a real-world dataset. All the results are compared with five state-of-the-art methods: He *et al.* (CVPR'09) [13], Zhu *et al* (TIP'15) [57], Ren *et al.* [33] (ECCV'16), Berman *et al.* [3, 4] (CVPR'16 and ICCP'17) and Li *et al.* [24] (ICCV'17). In addition, we conduct an ablation study to demonstrate the effectiveness of each module of our network.

4.1. Datasets

Similar to the existing deep learning-based dehazing methods [33, 5, 24, 51], we synthesize the training samples $\{$*Hazy /Clean /Transmission Map /Atmosphere Light*$\}$ based on (1). During synthesis, four atmospheric light conditions $A \in [0.5, 1]$ and the scattering coefficient $\beta \in [0.4, 1.6]$ are randomly sampled to generate their corresponding hazy images, transmission maps and atmospheric light maps. A random set of 1000 images are selected from the NYU-depth2 dataset [30] to generate the training set. Hence, there are in total 4000 training images, denoted as **TrainA**. Similarly, a test dataset **TestA** consisting of 400 (100×4) images also from the NYU-depth2 are obtained. We ensure that none of the testing images are in the training set. To demonstrate the generalization ability of our network to other datasets, we synthesize 200 $\{$*Hazy /Clean /Transmission Map /Atmosphere Light*$\}$ images from both the Middlebury stereo database (40) [36] and also the Sun3D dataset (160) [39] as the **TestB** set.

Table 1: Quantitative SSIM results for ablation study evaluated on synthetic **TestA** and **TestB** datasets.

	DED	DED-MLP	DED-MLP-GRA	DED-MLP-EP	DCPDN
TestA					
Transmission	0.9555	0.9652	0.9687	0.9732	0.9776
Image	0.9252	0.9402	0.9489	0.9530	0.9560
TestB					
Transmission	0.9033	0.9109	0.9239	0.9276	0.9352
Image	0.8474	0.8503	0.8582	0.8652	0.8746

Table 2: Quantitative SSIM results on the synthetic **TestA** dataset.

	Input	He. *et al.* [13] (CVPR'09)	Zhu. *et al.* [57] (TIP'15)	Ren. *et al.* [33] (ECCV'16)	Berman. *et al.* [3, 4] (CVPR'16)	Li. *et al.* [24] (ICCV'17)	DCPDN
Transmission	N/A	0.8739	0.8326	N/A	0.8675	N/A	**0.9776**
Image	0.7041	0.8642	0.8567	0.8203	0.7959	0.8842	**0.9560**

Table 3: Quantitative SSIM results on the synthetic **TestB** dataset.

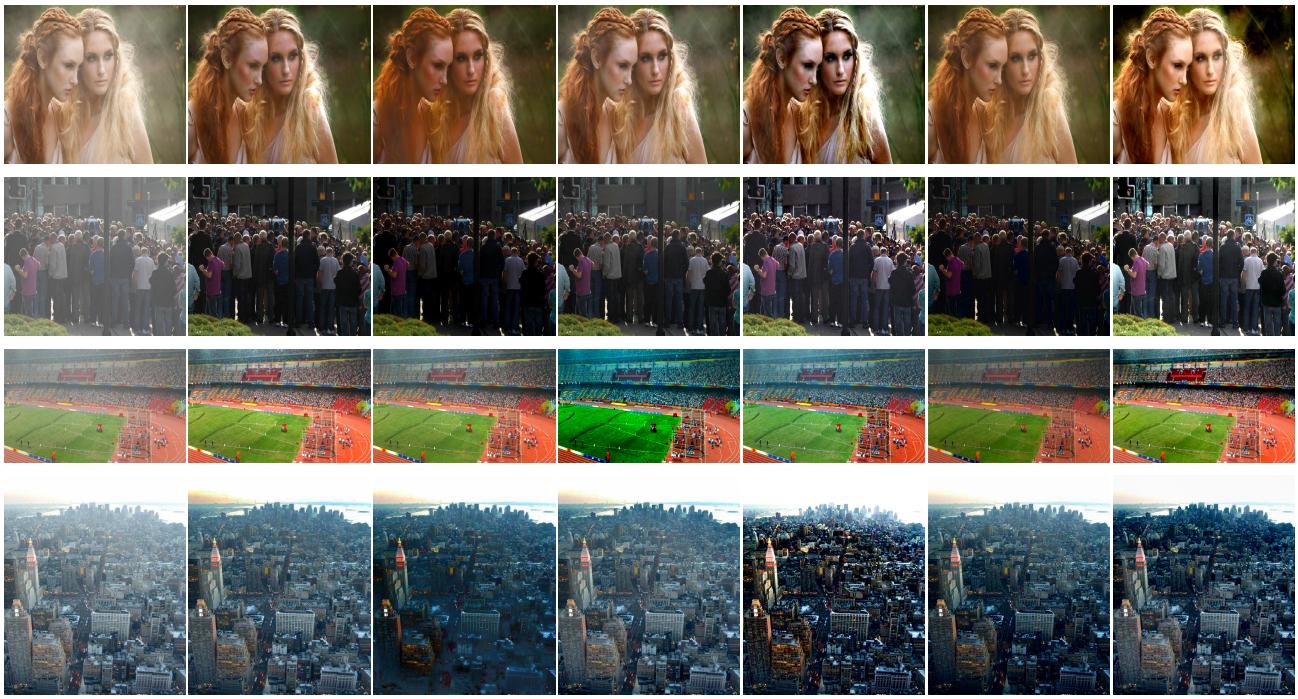
	Input	He. *et al.* [13] (CVPR'09)	Zhu. *et al.* [57] (TIP'15)	Ren. *et al.* [33] (ECCV'16)	Berman. *et al.* [3, 4] (CVPR'16)	Li. *et al.* [24] (ICCV'17)	DCPDN
Transmission	N/A	0.8593	0.8454	N/A	0.8769	N/A	**0.9352**
Image	0.6593	0.7890	0.8253	0.7724	0.7597	0.8325	**0.8746**

4.2. Training Details

We choose $\lambda_{E,l_2} = 1$, $\lambda_{E,g} = 0.5$, $\lambda_{E,f} = 0.8$ for the loss in estimating the transmission map and $\lambda_j = 0.25$ for optimizing the joint discriminator. During training, we use ADAM as the optimization algorithm with learning rate of 2×10^{-3} for both generator and discriminator and batch size of 1. All the training samples are resized to 512×512. We trained the network for 400000 iterations. All the parameters are chosen via cross-validation.

4.3. Ablation Study

In order to demonstrate the improvements obtained by each module introduced in the proposed network, we perform an ablation study involving the following five exper-



Input | He. *et al.* (CVPR'09) [13] | Zhu. *et al.* (TIP'15) [57] | Ren. *et al.* (ECCV'16)[33] | Berman. *et al.* (CVPR'16) [3, 4] | Li. *et al.* (ICCV'17) [24] | DCPDN

Figure 8: Dehazing results evaluated on real-world images released by the authors of previous methods.

iments: 1) Densely connected encoder decoder structure (**DED**), 2) Densely connected encoder decoder structure with multi-level pyramid pooling (**DED-MLP**), 3) Densely connected encoder decoder structure with multi-level pyramid pooling using L2 loss and gradient loss (**DED-MLP-GRA**), 4) Densely connected encoder decoder structure with multi-level pyramid pooling using edge-preserving loss (**DED-MLP-EP**), 5) The proposed DCPDN that is composed of densely connected encoder decoder structure with multi-level pyramid pooling using edge-preserving loss and joint discriminator (**DCPDN**). [3]

The evaluation is performed on the synthesized **TestA** and **TestB** datasets. The SSIM results averaged on both estimated transmission maps and dehazed images for the various configurations are tabulated in Table 1. Visual comparisons are shown in the Fig 6. From Fig 6, we make the following observations: 1) The proposed multi-level pooling module is able to better preserve the 'global' structural for objects with relatively larger scale, compared with (a) and (b). 2) The use of edge-preserving loss is able to better refine the edges in the estimated transmission map, compared with (b), (c) and (d). 3) The final joint-discriminator can further enhance the estimated transmission map by ensuring that the fine structural details are captured in the results, such as details of the small objects on the table shown in the second row in (e). The quantitative performance evaluated on both **TestA** and **TestB** also demonstrate the effectiveness of each module.

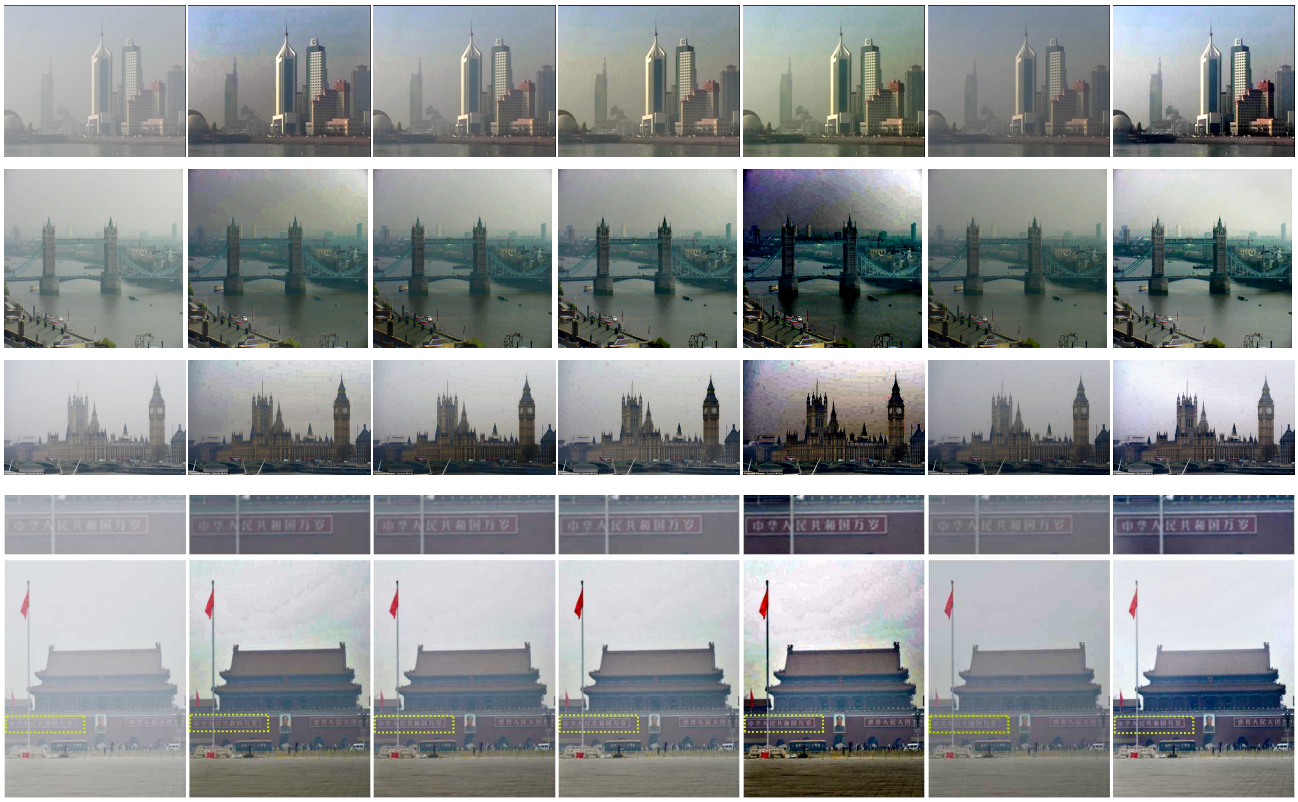
4.4. Comparison with state-of-the-art Methods

To demonstrate the improvements achieved by the proposed method, it is compared against the recent state-of-the-art methods [13, 57, 33, 3, 4, 24]. on both synthetic and real datasets.

Evaluation on synthetic dataset: The proposed network is evaluated on two synthetic datasets **TestA** and **TestB**. Since the datasets are synthesized, the ground truth images and the transmission maps are available, enabling us to evaluate the performance qualitatively as well as quantitatively. Sample results for the proposed method and five recent state-of-the-art methods, on two sample images from the test datasets are shown in Fig. 7. It can be observed that even though previous methods are able to remove haze from the input image, they tend to either over dehaze or under dehaze the image making the result darker or leaving some haze in the result. In contrast, it can be observed from our results that they preserve sharper contours with less color distortion and are more visually closer to the ground-truth. The quantitative results, tabulated in Table 2 and Table 3 [4], evaluated on both **TestA** and **TestB** also demonstrate the effectiveness of the proposed method.

[3] The configuration 1) **DED** and 2) **DED-MLP** are optimized only with $L2$ loss.

[4] N/A: Code released is unable to estimate the transmission map.



Input | He. *et al.* (CVPR'09) [13] | Zhu. *et al.* (TIP'15) [57] | Ren. *et al.* (ECCV'16) [33] | Berman *et al.* (CVPR'16) [3, 4] | Li. *et al.* (ICCV'17) [24] | DCPDN

Figure 9: Dehazing results evaluated on real-world images downloaded from the Internet.

Evaluation on a real dataset: To demonstrate the generalization ability of the proposed method, we evaluate the proposed method on several real-world hazy images provided by previous methods and other challenging hazy images downloaded from the Internet.

Results for four sample images obtained from the previous methods [33, 5, 10] are shown in Fig. 8. As revealed in Fig. 8, methods of He *et al.* [13] and Ren *et al.* [33] (observed on the fourth row) tend to leave haze in the results and methods of Zhu *et al.* [57] and Li *et al.* [24](shown on the second row) tend to darken some regions (notice the background wall). Methods from Berman *et al.* [3, 4] and our method have the most competitive visual results. However, by looking closer, we observe that Berman *et al.* [3, 4] produce unrealistic color shifts such as the building color in the fourth row. In contrast, our method is able to generate realistic colors while better removing haze. This can be seen by comparing the first and the second row.

We also evaluate on several hazy images downloaded from the Internet. The dehazed results are shown in Fig. 9. It can be seen from these results that outputs from He *et al.* [13] and Berman *et al.* [3, 4] suffer from color distortions, as shown in the second and third rows. In contrast, our method is able to achieve better dehazing with visually appealing results.

5. Conclusion

We presented a new end-to-end deep learning-based dehazing method that can jointly optimize transmission map, atmospheric light and dehazed image. This is achieved via directly embedding the atmospheric image degradation model into the overall optimization framework. To efficiently estimate the transmission map, a novel densely connected encoder-decoder structure with multi-level pooling module is proposed and this network is optimized by a new edge-preserving loss. In addition, to refine the details and to leverage the mutual structural correlation between the dehazed image and the estimated transmission map, a joint-discriminator based GAN framework is introduced in the proposed method. Various experiments were conducted to show the significance of the proposed method.

Acknowledgement

This work was supported by an ARO grant W911NF-16-1-0126.

References

[1] C. O. Ancuti and C. Ancuti. Single image dehazing by multi-scale fusion. *IEEE Transactions on Image Processing*, 22(8):3271–3282, 2013.

[2] E. Barshan and P. Fieguth. Stage-wise training: An improved feature learning strategy for deep models. In *Feature Extraction: Modern Questions and Challenges*, pages 49–59, 2015.

[3] D. Berman, S. Avidan, et al. Non-local image dehazing. In *CVPR*, pages 1674–1682, 2016.

[4] D. Berman, T. Treibitz, and S. Avidan. Air-light estimation using haze-lines. In *Computational Photography (ICCP), 2017 IEEE International Conference on*, pages 1–9. IEEE, 2017.

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

[6] X. Di and V. M. Patel. Face Synthesis from Visual Attributes via Sketch using Conditional VAEs and GANs. *ArXiv e-prints*, Dec. 2018.

[7] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 681–687. IEEE, 2015.

[8] R. Fattal. Single image dehazing. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, pages 72:1–72:9, New York, NY, USA, 2008. ACM.

[9] R. Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 27(3):72, 2008.

[10] R. Fattal. Dehazing using color-lines. volume 34, New York, NY, USA, 2014. ACM.

[11] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

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

[14] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[15] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *CVPR*, 2017.

[16] S.-C. Huang, B.-H. Chen, and W.-J. Wang. Visibility restoration of single hazy images captured in real-world weather conditions. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(10):1814–1824, 2014.

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 448–456, 2015.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[19] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.

[20] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[21] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski. Deep photo: Model-based photograph enhancement and viewing. In *ACM TOG*, volume 27, page 116. ACM, 2008.

[22] L. Kratz and K. Nishino. Factorizing scene albedo and depth from a single foggy image. In *ICCV*, pages 1701–1708. IEEE, 2009.

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[24] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. An all-in-one network for dehazing and beyond. *ICCV*, 2017.

[25] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. RESIDE: A Benchmark for Single Image Dehazing. *ArXiv e-prints*, Dec. 2017.

[26] J. Li, R. Klein, and A. Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[27] Z. Li, P. Tan, R. T. Tan, D. Zou, S. Zhiying Zhou, and L.-F. Cheong. Simultaneous video defogging and stereo reconstruction. In *CVPR*, pages 4988–4997, 2015.

[28] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, pages 469–477, 2016.

[29] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan. Efficient image dehazing with boundary constraint and contextual regularization. In *ICCV*, pages 617–624, 2013.

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

[31] X. Peng, Z. Tang, Y. Fei, R. S. Feris, and D. Metaxas. Jointly optimize data and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[33] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, pages 154–169. Springer, 2016.

[34] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[35] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[36] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] V. A. Sindagi and V. M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1879–1888, Oct 2017.

[39] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[40] M. Sulami, I. Glatzer, R. Fattal, and M. Werman. Automatic recovery of the atmospheric light in hazy images. In *Computational Photography (ICCP), 2014 IEEE International Conference on*, pages 1–11. IEEE, 2014.

[41] R. T. Tan. Visibility in bad weather from a single image. In *CVPR*, pages 1–8. IEEE, 2008.

[42] K. Tang, J. Yang, and J. Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *CVPR*, pages 2995–3000, 2014.

[43] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. *CVPR*, 2017.

[44] J. Wang, X. Li, and J. Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[45] L. Wang, V. A. Sindagi, and V. M. Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *FG*, 2018.

[46] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.

[47] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[48] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[49] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017.

[51] H. Zhang, V. Sindagi, and V. M. Patel. Joint transmission map estimation and dehazing using deep networks. *arXiv preprint arXiv:1708.00581*, 2017.

[52] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *ICCV*, 2017.

[53] Y. Zhang, L. Ding, and G. Sharma. Hazerd: an outdoor scene dataset and benchmark for single image dehazing. In *Proc. IEEE Intl. Conf. Image Proc.*, pages 3205–3209, 2017.

[54] Z. Zhang, Y. Xie, and L. Yang. Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[55] Z. Zhang, Y. Xie, and L. Yang. Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[56] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016.

[57] Q. Zhu, J. Mai, and L. Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing*, 24(11):3522–3533, 2015.

[58] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018.

[59] Y. Zhu and S. Newsam. Densenet for dense flow. In *ICIP*, 2017.