Joint Transmission Map Estimation and Dehazing using Deep Networks

He Zhang, *Student Member, IEEE*, Vishwanath Sindagi, *Student Member, IEEE*
Vishal M. Patel, *Senior Member, IEEE*

***Abstract*—Single image haze removal is an extremely challenging problem due to its inherent ill-posed nature. Several prior-based and learning-based methods have been proposed in the literature to solve this problem and they have achieved visually appealing results. However, most of the existing methods assume constant atmospheric light model and tend to follow a two-step procedure involving prior-based methods for estimating transmission map followed by calculation of dehazed image using the closed form solution. In this paper, we relax the constant atmospheric light assumption and propose a novel unified single image dehazing network that jointly estimates the transmission map and performs dehazing. In other words, our new approach provides an end-to-end learning framework, where the inherent transmission map and dehazed result are learned jointly from the loss function. Extensive experiments evaluated on synthetic and real datasets with challenging hazy images demonstrate that the proposed method achieves significant improvements over the state-of-the-art methods.**

I. Introduction

Haze is the obscuration of lower atmosphere, typically caused by the presence of suspended particles in the air such as dust, smoke and other dry particulates. The presence of haze usually reduces the visibility range, thus affecting quality of images captured by camera sensors that will be processed by computer vision systems. A sample hazy image is shown on the left side of Figure 1. It can be clearly observed that the existence of haze in an image greatly obscures the background scene. The problem of estimating a clear image from a single hazy input image is commonly referred to as dehazing. Image dehazing has attracted a significant interest in the computer vision and image processing communities in recent years [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16].

The deterioration of image quality is captured by the following mathematical model [17]:

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

where $\mathbf{x}$ is the location in the image co-ordinates, $\mathbf{I}$ represents the observed hazy image, $\mathbf{J}$ is the image before degradation, $\mathbf{A}$ is the global atmospheric light, and $t(\mathbf{x})$ is the transmission map. Transmission map contains the per-pixel attenuation information that affects the light reaching the camera sensor and it is a factor of depth as shown below:

$$t(\mathbf{x}) = e^{-\beta d(\mathbf{x})}, \tag{2}$$

where β is attenuation coefficient of the atmosphere and $d(\mathbf{x})$ is the depth map. One can view (1) as the superposition of two components: 1. *Direct attenuation* $(\mathbf{J}(\mathbf{x})t(\mathbf{x}))$, and 2. *Airlight* $(\mathbf{A}(\mathbf{x})(1 - t(\mathbf{x}))$. Direct attenuation represents the effect of scattering of light and the eventual decay of light before it reaches the camera sensor. Airlight is a phenomenon that results from the scattering of environmental light causing a shift in the apparent brightness of the scene. Note that Airlight is a function of scene depth and the global atmospheric light $\mathbf{A}$. As it can be observed from Eq. 1, image dehazing is an inherently ill-posed problem which has been addressed in different ways. Many previous methods overcome this issue by including extra prior assumption such as multiple images of the same scene [7] or depth information [6] to determine a solution. However, no extra information such as depth or multiple images is available for the problem of single image dehazing. To tackle this issue, different prior information has to be considered into the optimization framework such as dark-channel prior [5], contrast color-lines [18]] and hazeline prior [4]. For example, based on the observation that there always exists one channel that is significant dark in the captured outdoor images, dark-channel prior [5] is leveraged in the optimization framework to guarantee dehazed images are "dark-channel". Different from dark-channel prior, [4] leverage the haze-line prior in the framework, based on the observation that color cluster in the clear image can be approximated as the haze-line in RGB space. More recently, several learning-based methods have also been proposed, where different learning algorithms such as random forest regression and Convolutional Neural Networks (CNNs) are trained for predicting the transmission map [3], [1], [2], [8]. Many existing methods make an important assumption of constant atmospheric light [1] in the image degradation model (1) and tend to follow a two-step procedure. First, they learn the mapping from input hazy image to its corresponding transmission map and then using the estimated transmission map they calculate the clear image by reformulating Eq. 1 as

$$\mathbf{J}(\mathbf{x}) = \frac{\mathbf{I}(\mathbf{x}) - \mathbf{A}(\mathbf{x})(1 - t(\mathbf{x}))}{t(\mathbf{x})}. \tag{3}$$

As a result, most of the previous methods consider the task of transmission map estimation and dehazing as two separate tasks, except the Li *et al.* [8]. By doing so, they are unable to accurately capture the transformation between the transmission map and the dehazed image. Motivated by this observation,

He Zhang was with the Department of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ USA. email: he.zhang92@rutgers.edu
Vishwanath Sindagi and Vishal M. Patel are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA. Email: {vsindag1, vpatel36}@jhu.edu

[1]Meaning that the intensity of atmosphere light $\mathbf{A}$ is independent from its spatial location $\mathbf{x}$.

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

we relax the constant atmospheric light assumption [19] and propose to jointly learn the transmission map and dehazed image from an input hazy image using a deep CNN-based network. Relaxed constant atmospheric light hypothesis within a certain adjustable limit not only allows us to exploit the benefits of multi-task learning but it also enables us to regress on losses defined in the image space. By enforcing the network to learn the transmission map, we still follow the popular image degradation model (1). This joint learning enables the network to implicitly learn the atmospheric light and hence avoiding the need for manual calculation. On the other hand, previous learning-based CNN methods [1], [2] utilize Euclidean loss in generating the corresponding transmission map, which may result in blurry effect and hence poor quality dehazed images [20]. To tackle this issue, we incorporate the gradient loss combined with the adversarial loss to generate better transmission map with sharper edges.

Figure 2 gives an overview of the proposed single image dehazing method. Our network consists of three parts: 1. Transmission map estimation, 2. Hazy image feature extraction, and 3. Dehazing network guided by transmission map and hazy image features. The transmission map estimation is learned using a combination of adversarial loss, gradient loss and pixel-wise Euclidean loss. The transmission maps from this module are concatenated with the output of hazy image feature extraction module and processed by the dehazing network. Hence, the transmission maps are also involved in the dehazing procedure via the concatenation operator. The dehazing network is learned by optimizing a weighted combination of perceptual loss and pixel-wise Euclidean loss to generate perceptually better results. Shown in Figure 1 is a sample dehazed image using the proposed method.

This paper makes the following contributions:

- A novel joint transmission map estimation and image dehazing using deep networks is proposed. This is enabled by relaxing the constant atmospheric light assumption, thus allowing the network to implicitly learn the transformation from input hazy image to transmission map and transmission map to dehazed image.
- We propose to use the recently introduced Generative Adversarial Network (GAN) framework for learning the transmission map.
- By performing a joint learning of transmission map and image dehazing, we are able to minimize losses defined in the image space such as perceptual loss and pixel-wise Euclidean loss, thereby generating perceptually better results with high quality details.
- Extensive experiments on synthetic and real image datasets are conducted to demonstrate the effectiveness of the proposed method.

II. Related Work

We briefly review recent works on image dehazing and some commonly used losses in various CNN-based image reconstruction tasks.

A. Single Image Dehazing

Early methods tend to address the dehazing problem via including certain prior assumption. For example, the authors in [21] tend to recover the contrast for each patch relying on the assumption that that haze greatly decrease the contrast of the color images. Then, Kratz and Nishino [22] proposed to model the image with a factorial Markov random field in which the scene albedo and depth are two statistically independent latent layers. He. *et.al* in [5] proposed a dark-channel prior based on the surprising observation that RGB images from outdoor scene tend to have one channel that in significantly dark. Built on dark channel prior, Meng *et al.* [23] imposing a specific boundary constraint during the estimation of transmission map. More recently, Berman *et al.* [4] proposed a non-local prior method based on the observation that the colors of a haze-free image can be well represented by a few hundred different colors that fall into several tight clusters in the RGB space.

The success of CNNs in modeling the non-learning mapping between input and output has also inspired researchers to explore CNN-based algorithms for low-level vision tasks such as image dehazing [1], [2], [8]. Unlike previous prior-based methods in the estimation of transmission map, Cai *et al.* [2] train an end-to-end CNN network to directly estimate the transmission map from the input haze image. More recently, Ren *et al.* [1] proposed a multi-scale deep architecture to directly regress the transmission maps via a course to fine fashion. However, the method of both Ren *et al.* [1] and Cai *et al.* [2] still leveraged a two-step procedure and hence the whole algorithm is not end-to-end optimized. Most recently, Li *et al* proposed an all-in-one dehazing network, where a linear embedding is leveraged to encode the transmission map and the atmospheric light into a single variable. Though these CNN-based learning methods achieve superior performance over the recent state-of-the-art methods, they limit their capabilities by learning a mapping only between the input hazy image and the transmission map. This is mainly due to the fact that these methods are based on the popular image degradation model given by (1) which assumes a constant atmospheric light. In contrast, we relax this assumption and thus enable the network to learn a transformation from the input hazy image to transmission map and transmission map to dehazed image. By doing this, we are also able to use losses defined in the image domain to learn the network. In the following sub-sections, two different losses that we use to improve the performance of the proposed network are reviewed.

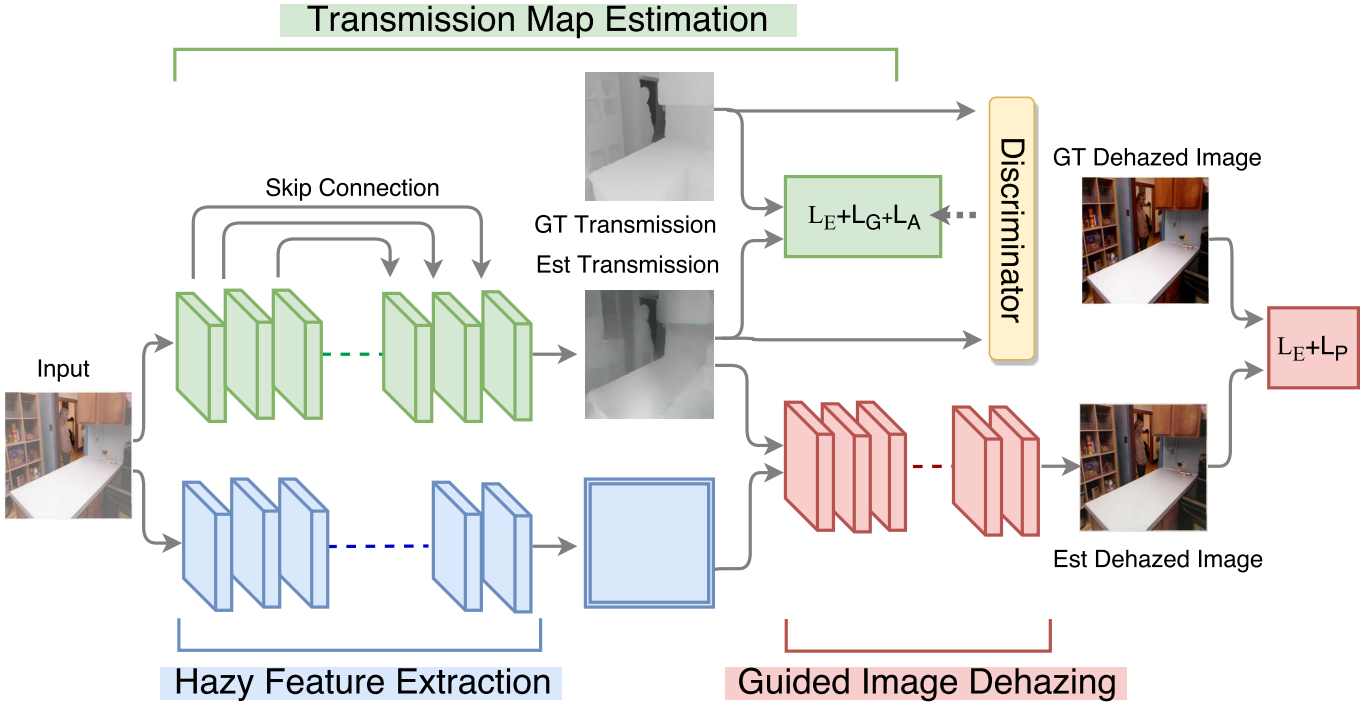

Fig. 2: Overview of the proposed multi-task method for image dehazing. The proposed method consists of three modules: (a) Hazy feature extraction, (b) Transmission map estimation, and (c) Guided image dehazing. First, the transmission map is estimated from the input hazy image and it is concatenated with high dimensional feature map. These concatenated maps are fed into the guided dehazing module to estimate the dehazed image. The transmission map estimation module is trained using a GAN framework. The image dehazing module is trained by minimizing a combination of perceptual loss and Euclidean loss.

B. Loss Functions

Loss functions form an important and integral part of a learning process, especially in CNN-based reconstruction tasks. Initial work on CNN-based image regression tasks optimized over pixel-wise L2-norm (Euclidean loss) or L1-norm between the predicted and ground truth images [24], [25]. Since these losses operate at per-pixel level, their ability to capture high level perceptual/contextual details such as sharp edges and complicated contour are limited and they tend to produce blurred results. In order to overcome this issue, we use two different loss functions: adversarial loss and perceptual loss for learning the transmission map and dehazed image, respectively.

1) Adversarial loss: The adversarial loss, formulated in the Generative Adversarial Networks(GAN) work by Goodfellow *et al.* [26], has been widely used in generating realistic images. GAN consists of a generator and a discriminator that are jointly optimized. While the generator's goal is to synthesize images that are similar in distribution of the training images, the discriminator's job is to identify if the images fed to it are real or synthesized (fake). After the success of this method in generating realistic images, this concept has been explored as different formulations in various applications such as data augmentation [27], paired and unpaired 2d/3d image to image translation [28], [29], [30], [31], image super-resolution [32], image inpainting [33], [34] and image de-raining [35]. In our work, we propose to use the GAN framework as an additional loss function to guide the learning of transmission map, which when optimized appropriately, will generated realistic transmission maps.

2) Perceptual loss: Many researchers have argued and demonstrated through their results that it would be better to optimize a perceptual loss function in various applications [36], [37], [38]. The perceptual function is usually defined using high-level features extracted from a pre-trained convolutional network. The aim is to minimize the perceptual difference between the reconstructed image and the ground truth image. Perceptually superior results were obtained for both super-resolution and artistic style-transfer [39], [15], [40]. In this work, a VGG-16 architecture [41] based perceptual loss is used to train the network for performing dehazing.

III. Proposed Method

The proposed network is illustrated in Figure 2 which consists of the following modules: 1. *Transmission map estimation*, 2. *Hazy image feature extraction*, and 3. *Transmission guided image dehazing*, where the first module learns to estimate transmission maps from corresponding input hazy images, the second module extracts haze relevant features from the input hazy image and the third module learns to perform image dehazing by combining the feature information extracted from the hazy image with the estimation transmission map. In what follows, we explain these modules in detail.

A. Transmission Map Estimation

The task of predicting transmission map from a given input hazy image is considered as a pixel-level image regression

task. In other words, the aim is to learn a pixel-wise non-linear mapping from a given input image to the corresponding transmission map by minimizing the loss between them. In contrast to the method used by Ren *et al.* in [1], our method uses adversarial loss in addition to pixel-wise Euclidean loss to learn better quality transmission maps. Also, the network architecture used in this work is very different from the one used in [1].

For incorporating the adversarial loss, the transmission map estimation is learned in the Conditional Generative Adversarial Network (CGAN) framework [42]. Similar to earlier works on GANs for image reconstruction tasks [35], [43], [32], the proposed network for learning the transmission map consists of two sub-networks: Generator G and Discriminator D. The goal of GAN is to train G to produce samples from training distribution such that the synthesized samples are indistinguishable from the actual distribution by the discriminator D. The sub-network G is motivated by the success of encoder-decoder structure in pixel-wise image reconstruction [44], [45], [43]. In this work, we adopt a 'U-Net'-based structure [44] as the generator for the transmission map estimation. Rather than concatenating the symmetric layers during training, shortcut connections [46] are used to connect the symmetric layers with the aim of addressing the vanishing gradient problem for deep networks. To better capture the semantic information and make the generated transmission map indistinguishable from the ground truth transmission map, a CNN-based differentiable discriminator is used as a 'guidance' to guide the generator in generating better transmission maps. The proposed generator network is as follows (the shortcut connection is neglected here):
CP(15)-CBP(30)-CBP(60)-CBP(120)-CBP(120)-CBP(120)-CBP(120)-CBP(120)-TCBR(120)-TCBR(120)-TCBR(120)-TCBR(120)-TCBR(60)-TCBR(30)-TCBR(15)-TC(1)-TanH,
where C represents the convolutional layer, TC represents transpose convolution layer, P indicates Prelu and B indicates batch-normalization The number in the bracket represents the number of output feature maps of the corresponding layer.

To ensure that the estimated transmission map is indistinguishable from the ground truth image, a learned discriminator sub-network is designed to classify if each input image is real or fake. Inspired by the success of patch-discriminator in distinguish real from fake, we also adopt a $70\times$ 70 patch discriminator, where 70×70 indicates the receptive field of the discriminator, to generate visually pleasing and sharper results. [47] also explores other ways to make the images sharper. The structure of the discriminator is defined as follows:
CB(48)-CBP(96)-CBP(192)-CBP(384)-CBP(384)-C(1)-Sigmoid.

Furthermore, we propose to employ gradient-based loss function in order to enforce consistency in the gradients between the estimated and target transmission map. The use of gradient loss function is inspired by its success in several other tasks such as depth estimation.

B. Hazy Feature Extraction and Guided Image Dehazing

A possible solution to image dehazing is to directly learn an end-to-end non-linear mapping between the estimated transmission map and the desired dehazed output. However, as shown in [43], while learning a mapping from transmission map-like to an RGB color image is possible, one may loose some information due to the absence of the albedo and the lighting information.

To generate better dehazed image and enable the whole process (estimation of the transmission map and the dehazed image) end-to-end, we propose a deep transmission guided network for single image dehazing via relaxing the assumption of constant atmospheric light. Inspired by *guided filtering*, where a guidance image is leveraged to guided the generation of high-quality results (eg. depth map), a set of convolutional layers with symmetric skip connections are stacked in the front and they serve as a hazy image feature extractor. Basically, the hazy feature extraction part extract deep features from the input hazy image. Then, These feature maps are concatenated with the estimated transmission map. Then the concatenation is fed into the guided image dehazing module. This module consists of another set of CNN layers with non-linearities and it essentially acts as a fusion CNN whose task is to learn a mapping from transmission map and high-dimensional feature maps to dehazed image. [2] To learn this network, a perceptual loss function based on VGG-16 architecture [41] is used in addition to pixel-wise Euclidean loss. The use of perceptual loss greatly enhances the visual appeal of the results. Details of the network structure for the hazy feature extraction and guided image dehazing module are as follows:

CP(20)-CBP(40)-CBP(80)-C(1)-Conca(2)-CP(80)-CBP(40)-CBP(20)-C(3)-TanH,
where $Conca$ indicates concatenation.

In summary, a non-linear mapping from the input hazy image and transmission map to dehazed image is learned in a multi-task end-to-end fashion. By learning this mapping, we enforce our network to implicitly learn the estimation of atmospheric light, thereby avoiding the "manual" estimation as followed by some of the existing methods.

C. Training Loss

As discussed earlier, the proposed method involves joint learning of two tasks: transmission map estimation and dehazing. Accordingly, to train the network, we define two losses L^t and L^d, respectively for the two tasks.

1) Transmission map loss L^t: To overcome the issue of blurred results due to the minimization of L_2 error, the transmission map estimation network is learned by minimizing a weighted combination of L_2 error, an adversarial error and a gradient loss. The transmission map loss is defined as

$$L^t = L^t_E + \lambda_a L^t_A + \lambda_G L^t_G, \tag{4}$$

[2]Note that our network is quite different from the network proposed in [48] in the sense that the proposed network is a multi-task learning network with a single input while the network in [48] is a single-task network with two inputs.

Fig. 3: Transmission estimation results for **Ablation 1**. It can be observed that gradient loss enable sharper edges and final GAN framework help to preserved better structure information for each object.

where λ_a and λ_G are two weights, L_E^t is the pixel-wise Euclidean loss, L_A^t is the adversarial loss and L_G^t is the two-directional gradient loss. These three losses are defined as follows

$$L_E^t = \sum_{w,h} \|(\phi_G(\mathbf{I}))^{w,h} - \mathbf{y}_t^{w,h}\|_2, \tag{5}$$

$$L_A^t = -\log(\phi_D(\phi_G(\mathbf{I})), \tag{6}$$

where $\mathbf{I}$ is a C-channel input hazy image, $\mathbf{y}_t$ is the ground truth transmission map, $w \times h$ indicates the dimension of the input image and transmission map, ϕ_G is the generator sub-network G for generating the transmission map and ϕ_D is the discriminator sub-network D. The directional gradient loss, which has been discussed in other applications [49], [50], the is defined as:

$$\begin{aligned} L_G^t = &\sum_{w,h} \|(H_x(G_t(\mathbf{I})))^{w,h} - (H_x(t))^{w,h}\|_2 \\ &+ \|(H_y(G_t(\mathbf{I})))^{w,h} - (H_y(t))^{w,h}\|_2, \end{aligned} \tag{7}$$

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.

Traditional techniques for transmission map estimation employ only the Euclidean loss (L_E^t) to learn the network weighs. However, Euclidean loss is known to introduce blur in the generated output. Hence, the use of additional loss functions (adversarial loss and gradient loss) incorporates further constraints into the learning framework. Specifically, the adversarial loss (L_A^t) enforces the network to generate transmission maps that are closer to the input distribution and the gradient loss (L_G^t) ensures consistency between the gradients of the target and estimated transmission map. The weights λ_A, λ_G are set using validation.

2) Dehazing loss L^d: The dehazing network is learned by minimizing a weighted combination of the pixel-wise Euclidean loss and perceptual loss between the ground-truth dehazed image and the network output and is defined as follows

$$L^d = L_E^d + \lambda_p L_P^t, \tag{8}$$

where λ_p is a weighting factor, L_E^d is the pixel-wise Euclidean loss and L_P^t is the perceptual loss and are respectively defined as

$$L_E^d = \sum_{c,w,h} \|\phi_E(\mathbf{I})^{c,w,h} - \mathbf{J}^{c,w,h}\|_2, \tag{9}$$

$$L_P^d = \sum_{c_i,w_i,h_i} \|V(\phi_E(\mathbf{I}))^{c_i,w_i,h_i} - V(\mathbf{J})^{c_i,w_i,h_i}\|_2, \tag{10}$$

where $\mathbf{I}$ is a c-channel input hazy image, $\mathbf{J}$ is the ground truth dehzed image, $w \times h$ is the dimension of the input image and the dehazed image, ϕ_E is the proposed network, V represents a non-linear CNN transformation and C_i, W_i, H_i are the dimensions of a certain high level layer of V. Similar to the idea proposed in [36], we aim to minimize the distance between high-level features along with pixel-wise Euclidean loss. In our method, we compute the feature loss at layer relu3_1 in VGG-16 model [41].[3] Note that the dehazing loss L^d is also to be propagated to the transmission estimation part.

D. Discussion

Relaxing the condition of constant atmospheric light enables the network to be trained in an end-to-end fashion, thus allowing the network to implicitly learn the transformation from input hazy image to transmission map and transmission map to dehazed image. While it allows more flexibility in the learning process, it introduces more complexity on the model. Hence, to efficiently learn the network parameters, the transmission map is considered since it preserve information about the portion of the light that is not scattered that reaches the camera. Furthermore, additional losses such as adversarial loss and gradient loss function introduce strong regularization, thus enabling better estimation of transmission map.

IV. Experiments

In this section, we present the details and results of various experiments conducted on synthetic and real datasets that contain a variety of hazy conditions. First we describe the datasets used in our experiments. Then, we discuss the details of the training procedure. Next, we discuss the results of the

[3]https://github.com/ruimashita/caffe-train/blob/master/vgg.train_val.prototxt

ablation study conducted to understand the improvements obtained by various modules of the proposed method. Finally, we compare the results of the proposed network with recent state-of-the-art methods. Through these experiments, we attempt to demonstrate the superiority of the proposed method and the effectiveness of its' various components.

A. Datasets

Since it is extremely difficult to collect a dataset that contains a large number of hazy/clear/transmission-map image pairs, training and test datasets are synthesized using (1) and following the idea proposed in [3], [2], [1]. All the training and test samples are obtained from the NYU Depth dataset [51]. More specifically, given a haze-free image, we randomly sample four atmosphere light $\mathbf{A}(\mathbf{x}) \in [0.5, 1.2]$ and the scattering coefficient of the atmosphere $\beta \in [0.4, 1.6]$ to generate its corresponding hazy images and transmission maps. An initial set of 600 images are randomly chosen from the NYU dataset. From each image belonging to this initial set, 4 training images are generated by using randomly sampled atmospheric light and scattering coefficient, obtaining a total of 2400 training images. In a similar way, a test dataset consisting of 300 images is obtained. We ensure that none of the training images are in the test set. By varying $\mathbf{A}$ and β, we generate our training data with a variety of different conditions.

As discussed in [1], [3], the image content is independent of its corresponding depth. Even though the training images are from the indoor dataset [51] and hence depths of all the images are relatively shallow, we could modify the value of the attenuation coefficient β to vary the haze concentration to make sure the datasets can also used for outdoor image dehazing. Meanwhile, the experimental results have also demonstrated the effectiveness of discussed training datasets.

To demonstrate the effectiveness of the proposed method on real-world data, we also created a test dataset including 30 hazy images downloaded from the Internet.

B. Training Details

The entire network is trained on a Nvidia Titan-X GPU using the torch [52] framework. We choose $\lambda_a = 0.003$ and $\lambda_G = 1$ for the loss in estimating the transmission map and $\lambda_p = 1.5$ for the loss in single image dehazing. During training, we use ADAM [53] as the optimization algorithm with learning rate of 2×10^{-3} and batch size of 10 images. All the training samples are resized to 256×256. To efficiently train the multi-task network, we leverage the stage-wise training strategy. First, the transmission map estimation module is trained using the loss in Eq. 4 . Then, the entire network is fine-tuned using both Eq. 8 and Eq. 4.

C. Ablation Study

In order to demonstrate the improvements obtained by different modules for both transmission maps and dehazed images, we conduct two ablation studies for estimating transmission maps and dehazed images, separately.

TABLE I: Quantitative SSIM results for **Ablation 1** evaluated on synthetic datasets for transmission map.

	Input	***T-L2***	***T-L2-G***	***T-L2-ALL***	Target
Transmission Map	0.4523	0.9052	0.9257	**0.9388**	1.0000

Ablation 1: This ablation study demonstrates the effectiveness of different modules in the transmission map estimation block and it consists of the following experiments:
1) *Transmission map estimation using only L2 loss (**T-L2**)*,
2) *Transmission map estimation using L2 loss and gradient loss (**T-L2-G**)*, and
3) *Transmission map estimation using L2 loss, gradient loss and adversarial loss (**T-L2-G-GAN**)*.

Sample results are shown in Fig 3. It can be observed that the introduction of gradient loss (***T-L2-G***) eliminates halo-artifacts near complicated edges [20]. Furthermore, the introduction of the discriminator (GAN framework-***T-L2-G-GAN***) effectively refine the local regions and enables sharper reconstructions, thereby preserving the structure for each object. Results of quantitative analysis on synthetic datasets are presented in Table I. The effect of different modules in the proposed network can be clearly observed from this table.

Ablation 2: Similarly, another ablation study is conducted to demonstrate the improvements obtained by different modules for dehazing images. This ablation study involves the following experiments:
1) *Image dehazing using L2 loss without estimation of transmission map (**I-L2-noT**)*,
2) *Image dehazing using L2 loss with estimation of transmission map (**I-L2-T**)*, and
3) *Image dehazing using L2 loss and perceptual loss with estimation of transmission map (**I-L2-Per-T**)*.

Sample results are shown in Fig 4. It can be observed that the method (***I-L2-noT***) is unable to accurately estimate the haze level and depth (both are inherently captured in the transmission map) and hence the dehazed results tend to contain some color distortion. The introduction of the branch for the estimation of transmission map helps to generate better quality images. This can be seen by comparing the second column and the third column in Fig 4. Furthermore, the final involvement of the perceptual loss ***I-L2-Per-T*** is able to generate better dehazed images with high quality details (observed from the zoom-in parts in Fig 4). We also compare the inference running time for each ablation study, as tabulated in Table III. It can be observed that the multi-task learning results in slight increase in complexity of training and inference time. However, it leads to substantial improvements in the dehazing quality. The introduction of different loss functions such as gradient loss and perceptual loss increase the training time, however, it does not affect the inference time.

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

To demonstrate the improvements achieved by the proposed method, it is compared against recent state-of-the-art methods

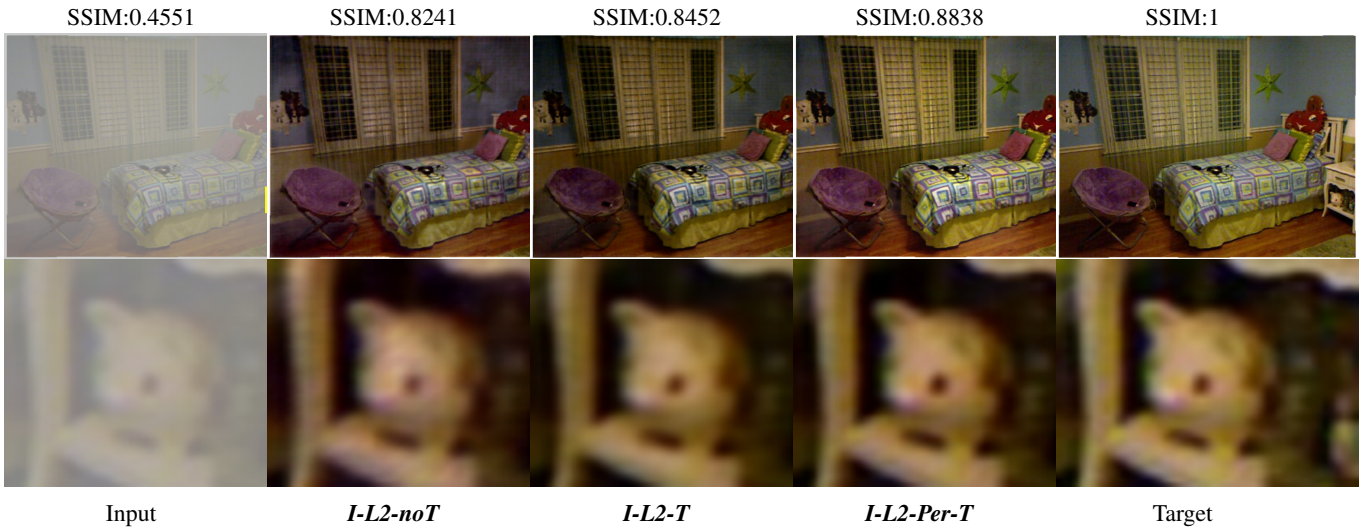

Fig. 4: Dehazed image results and certain zoomed-in parts for **Ablation 2**. It can be observed that the introduction of transmission map reduce the color distortion and the involvement of perceptual loss enable high quality dehazed result.

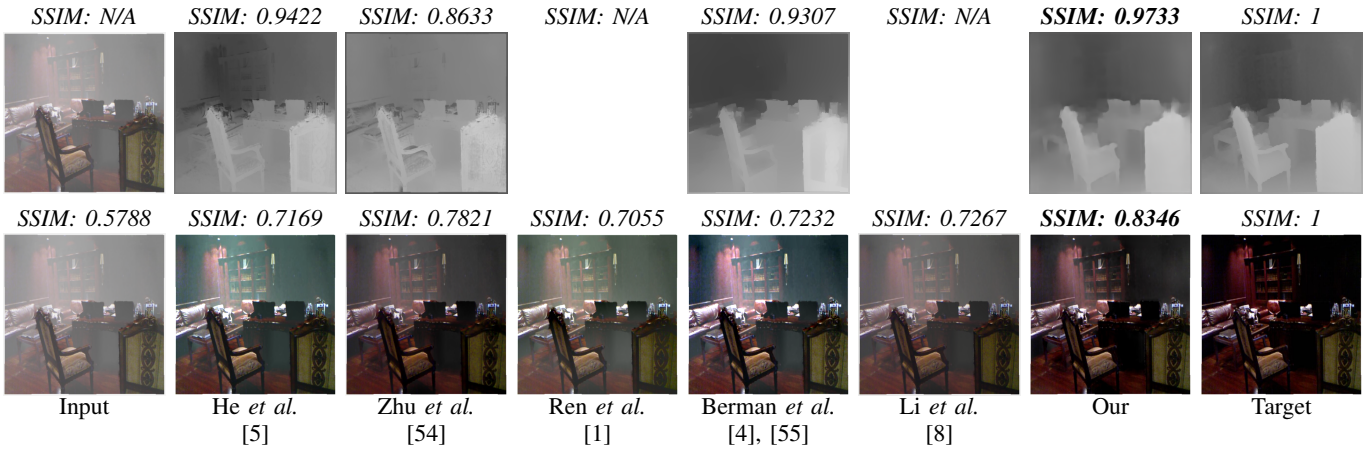

Fig. 5: Dehazing results from our synthetic images, where the first row correspond to the estimated transmission map and the last row corresponds to the dehazed image.

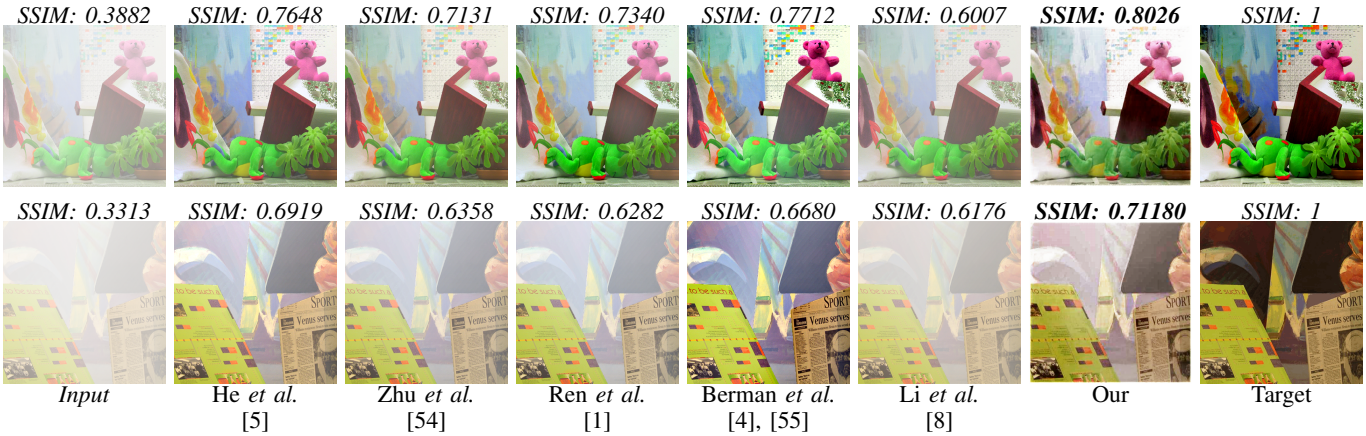

Fig. 6: Dehazed visual comparisons for results of synthetic image used by previous methods [2], [8].

on synthetic and real datasets.

Evaluation on synthetic dataset: Synthetic dataset, as described in Section IV(A), is used for the purpose of training and evaluating the network. Due to the availability of ground-truth images, we conduct both qualitative and quantitative evaluations.

Figure 5 shows results of the proposed method as compared

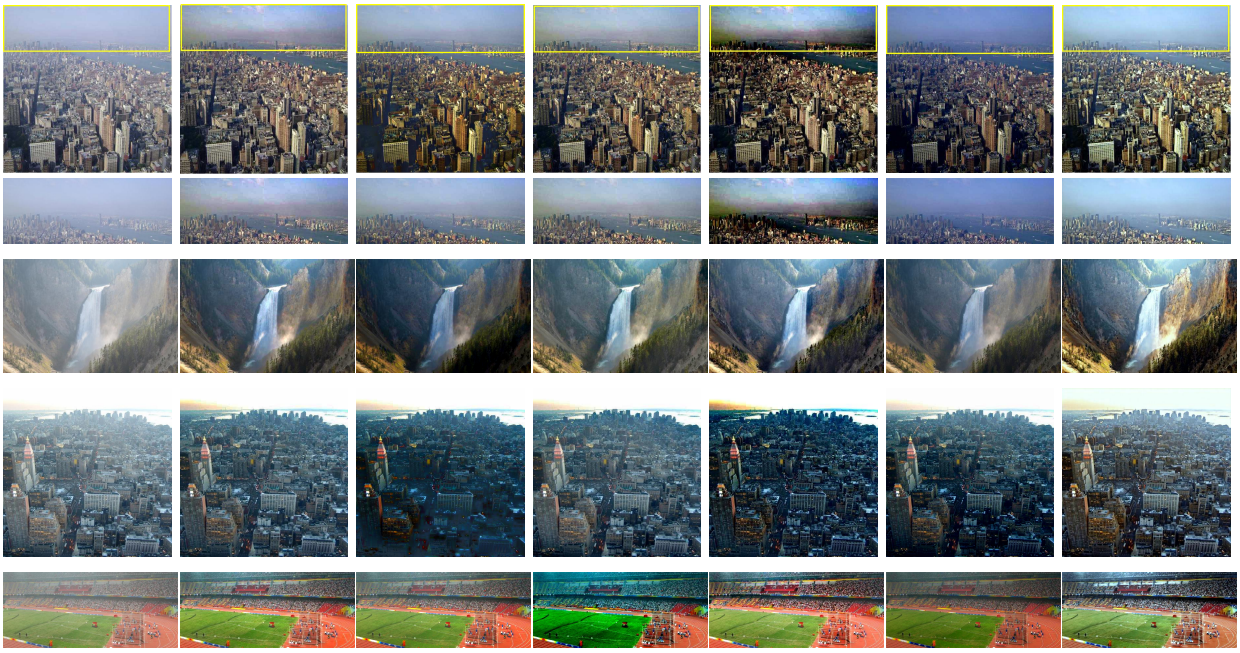

Fig. 7: Qualitative comparison of dehazing on real-world dataset that is presented in previous dehazing papers. It can be observed from the highlighted region that previous methods may result in undesirable effects such as artifacts and color over-saturation in the output images

Fig. 8: Qualitative comparison of dehazing on real-world dataset. Results on two sample images from a set of images downloaded from the Internet.

TABLE II: Quantitative SSIM results for **Ablation 2** evaluated on synthetic datasets for dehazed image.

	Input	*I-L2-noT*	*I-L2-T*	*I-L2-Per-T*	Target
Dehazed Image	0.7041	0.8835	0.9002	**0.9133**	1.0000

TABLE III: Average running time for **Ablation 2** evaluated on synthetic datasets for dehazed image.

	Input	*I-L2-noT*	*I-L2-T*	*I-L2-Per-T*
Time (s)	2.65	3.33	3.33	3.33

with recent state-of-the-art methods ([5], [54], [4], [55], [1], [8]) on a sample image from the test split of the synthetic dataset. After carefully analyzing these results, we observed that the recent best methods resulted in either incomplete removal of haze or over-correction which reduced the visual appeal of the image. Even though, [4] is able to achieve good performance in the presence of moderate haze, its dehazed results tend to contain color shift. In contrast, the proposed method is able to achieve better dehazing for a variety of haze contents. Similar results can be observed regarding the quality of transmission maps estimated by the proposed multi-task method as compared with the existing methods. It can be noted that the previous methods are unable to accurately estimate the relative depth in a given image, resulting in lower quality of dehazed images. In contrast, the proposed method not only estimates high quality transmission maps, but also achieves better quality dehazing.

The quantitative performance of the proposed method is compared against five state-of-the-art methods [5], [54], [1], [4], [8] using SSIM [56]. The quantitative results are tabulated in Table IV. It can be observed from this table that the proposed method achieves the best performance in

TABLE IV: Quantitative SSIM results on the synthetic dataset.

	Input	He. *et al.* [5]	Zhu. *et al.* [54]	Ren. *et al.* [1]	Berman. *et al.* [4], [55]	Li. *et al.* [8]	Our
Transmission	N/A	0.8739	0.8326	N/A	0.8675	N/A	**0.9388**
Image	0.7041	0.8642	0.8567	0.8203	0.7959	0.8842	**0.9133**

terms SSIM. Note that, we have attempted to obtain the best possible results for the other methods by fine-tuning their respective parameters based on the source code released by the authors and kept the parameter consistent for all the experiments. As the code released by [1], [8] cannot estimate the predicted transmission map, the results for the transmission estimation corresponding to [1], [8] is not included in the discussion.

Furthermore, we also evaluate the proposed method on the synthetic images used by previous methods [2], [8]. Results are shown in Fig 6. It can be clearly observed that Berman *et al.* [4], [55] and the proposed methods achieve the best visual performance among all. However, by looking closer at the upper right part of Fig 6, it can be found that method from Berman *et al.* [4], [55] tend to bring in the color-shift and hence degrade the overall performance.

Evaluation on real dataset: In addition to the synthetic dataset, we also conducted evaluation experiments on real dataset which consists of hazy images from the real world, collected from the internet. Since the ground truths are not available for such images, we do not use this dataset for training and we perform only qualitative evaluations.

Comparison of results on four sample images used in earlier methods compared with various approaches is shown in Figure 7. Yellow rectangles are used to highlight the improvements obtained using the proposed method. Though the existing methods seem to achieve good visual performance in the top row, it can be observed from the highlighted region that these methods may result in undesirable effects such as artifacts and color over-saturation in the output images. For the bottom two rows, the existing methods either make the image darker due to overestimation of dark pixels or are unable to perform complete dehazing. For example, leaning-based methods [1], [8] underestimate the thickness of haze resulting in partial dehazing. Even though Berman *et al.* [4], [55] leaves less haze in the output, the resulting image tends to be darker as the haze line is tough to detect under heavy haze conditions. In contrast, the proposed method is able to achieve near-complete dehazing with visually appealing results by avoiding any undesirable effects in the output images.

Furthermore, we also illustrate three qualitative examples of dehazing results on real-world hazy images by different methods. He. *et al* [5], Li. *et al* [8] and Ren. *et al* [1] method perform well but they tend to leave haze in the output leading to loss in color contrast. Even though Berman *et al* [4], [55] perform better, they tend to over-estimate the haze level resulting darker output images. Overall, our proposed method is able to tackle the problems brought by the other methods and achieve the best performance visually.

In Fig 9, we present a very tough hazy image to illustrate the results. The visual comparison here also confirms our findings in the previous experiments. Particularly, from the highlighted yellow rectangle, it can be observed that the method can better recover the Mandarin characters hidden behind the haze.

Through these experiments on real dataset, we are able to demonstrate that the proposed method, although trained on synthetic dataset, is able to generalize well to real world conditions.

Run Time Comparison: The proposed method is evaluated for its computational complexity. On average, our method is able to processes 512×512 images at 18 frames per second (fps), thus providing real-time performance. Further more, the proposed method is compared against several recent methods as shown in Table V. The proposed method is comparable to the Li. *et.al* [8] but with better performance. On average, it takes about 3.3s to de-rain an image of size 512×512.

E. Failure Cases

Although the proposed method is able to generalize well to most of the outdoor cases, it results in saturation of certain region of specific images. For example, as shown in dehazed images in Fig 11, central part of the sky is not recovered appropriately and it looks over-exposed. This is primarily due to the rarity of similar samples during training. This is a common problem in most existing methods.

Though the success of using synthetic samples for avoiding the need of expensive annotations has demonstrated the effectiveness in single image dehazing, the performance gap between the results on synthetic and real-world images illustrates some of the limitations in learning from synthetic data. Hence, it is necessary to explore new possibilities for leveraging synthetic data in order to obtain better generalization across real world images.

V. Conclusion

This paper presented a new multi-task end-to-end CNN-based network that jointly learns to estimate transmission map and performs image dehazing. In contrast to the existing methods that consider the transmission estimation and single image dehazing as two separate tasks, we bridge the gap between them by using multi-task learning. This is achieved by relaxing the constant atmospheric light assumption in the standard image degradation model. In other words, we enforce the network to estimate the transmission map and use it for further dehazing thereby following the standard image degradation model for image dehazing. Experiments were conducted on multiple datasets (synthetic and real) and the results were compared against several recent methods. Further, detailed ablation studies were conducted to understand

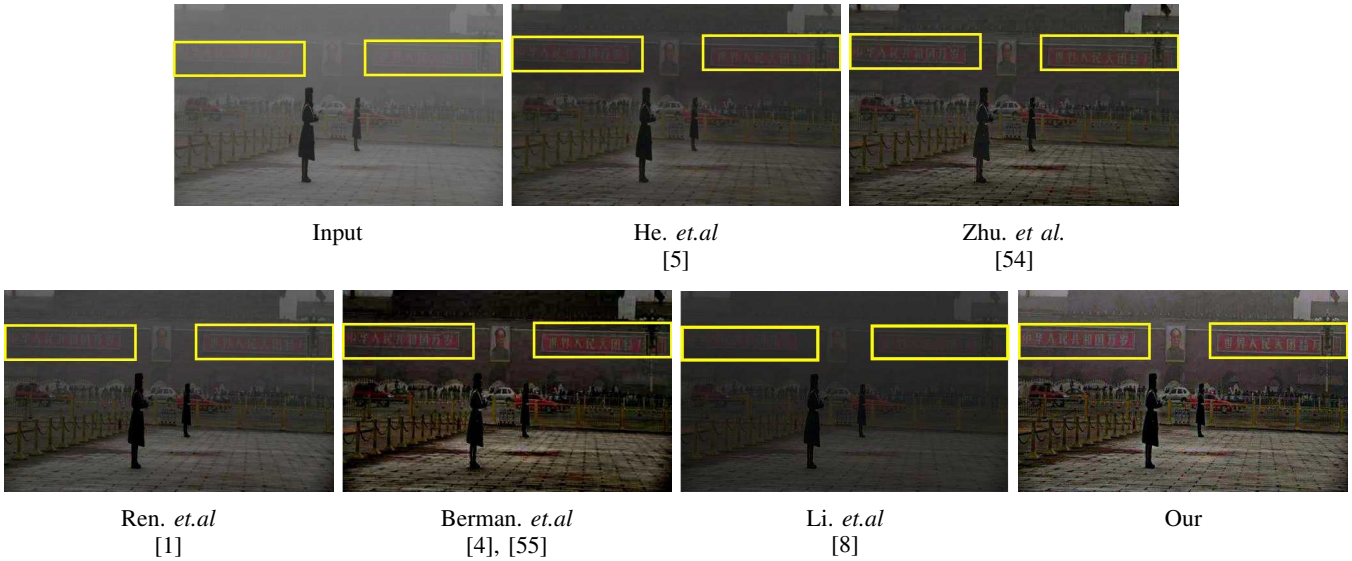

Fig. 9: Qualitative comparison of dehazing on real-world dataset. Top row: Results on a sample image from the real-world dataset provided by previous methods. Bottom two rows: Results on two sample images from a set of images downloaded from the Internet.

TABLE V: Average running time on the synthesized dataset. M: Matlab implementation, P: Python implementation.

	He. *et.al* (M) [5] (M)	Zhu. *et.al* [54] (M)	Ren. *et.al* [1] (M)	Berman. *et.al* [4] (M)	Li. *et.al* [8] (P)	Our (P)
Time (s)	25.08	3.92	3.75	8.41	**3.18**	**3.33**

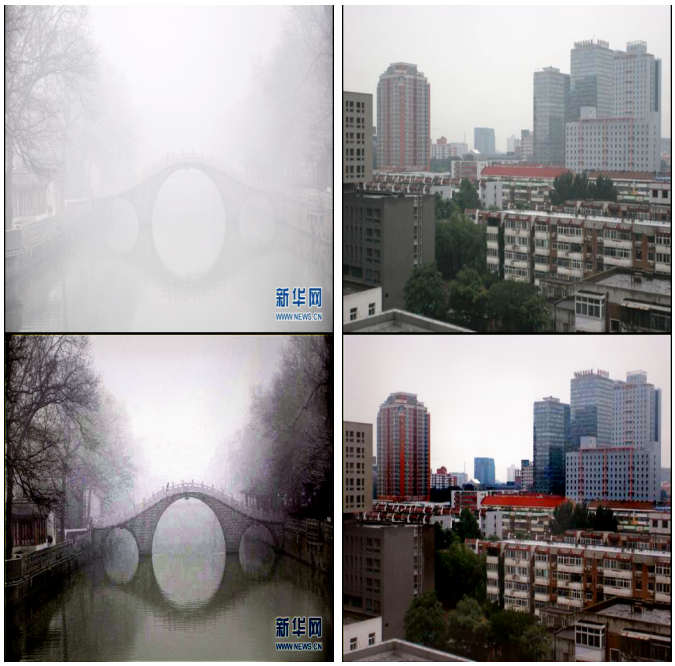

Fig. 10: More dehazing results on the real-world images. The first row show the original hazy image and second row show dehazed images of the proposed method.

Fig. 11: Failure case of the proposed method,

the significance of the different components in the proposed. method.

ACKNOWLEDGEMENT

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

REFERENCES

[1] 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*. Springer, 2016, pp. 154–169.

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

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

[4] D. Berman, S. Avidan *et al.*, "Non-local image dehazing," in *CVPR*, 2016, pp. 1674–1682.

[5] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. on PAMI*, vol. 33, no. 12, pp. 2341–2353, 2011.

[6] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski, "Deep photo: Model-based photo-

graph enhancement and viewing," in *ACM TOG*, vol. 27, no. 5. ACM, 2008, p. 116.

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

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

[9] X. Yang, Z. Xu, and J. Luo, "Towards perceptual image dehazing by physics-based disentanglement and adversarial training," 2018.

[10] J.-H. Kim, W.-D. Jang, J.-Y. Sim, and C.-S. Kim, "Optimized contrast enhancement for real-time image and video dehazing," *Journal of Visual Communication and Image Representation*, vol. 24, no. 3, pp. 410–425, 2013.

[11] A. Galdran, J. Vazquez-Corral, D. Pardo, and M. Bertalmio, "Fusion-based variational image dehazing," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 151–155, 2017.

[12] W. Wang, X. Yuan, X. Wu, and Y. Liu, "Fast image dehazing method based on linear transformation," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1142–1155, 2017.

[13] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3194–3203.

[14] W. Ren, J. Zhang, X. Xu, L. Ma, X. Cao, G. Meng, and W. Liu, "Deep video dehazing with semantic segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1895–1908, 2019.

[15] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 902–911.

[16] Z. Xu, X. Yang, X. Li, X. Sun, and P. Harbin, "Strong baseline for single image dehazing with deep features and instance normalization."

[17] R. Fattal, "Single image dehazing," in *ACM SIGGRAPH 2008 Papers*, ser. SIGGRAPH '08. New York, NY, USA: ACM, 2008, pp. 72:1–72:9. [Online]. Available: http://doi.acm.org/10.1145/1399504.1360671

[18] ——, "Dehazing using color-lines," vol. 34, no. 13. New York, NY, USA: ACM, 2014.

[19] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 226–234.

[20] 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*, vol. 24, no. 10, pp. 1814–1824, 2014.

[21] R. T. Tan, "Visibility in bad weather from a single image," in *CVPR*. IEEE, 2008, pp. 1–8.

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

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

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[25] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2944–2956, 2017.

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

[27] X. Peng, Z. Tang, F. Yang, R. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," *arXiv preprint arXiv:1805.09707*, 2018.

[28] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," *arXiv preprint arXiv:1802.09655*, 2018.

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

[30] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *International Journal of Computer Vision*, pp. 1–18.

[31] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "Siclope: Silhouette-based clothed people," *CoRR*, vol. abs/1901.00049, 2019. [Online]. Available: http://arxiv.org/abs/1901.00049

[32] 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.

[33] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[34] Y. Zhao, W. Chen, J. Xing, X. Li, Z. Bessinger, F. Liu, W. Zuo, and R. Yang, "Identity preserving face completion for large ocular region occlusion," *arXiv preprint arXiv:1807.08772*, 2018.

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

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

[37] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer."

[38] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *CoRR*, vol. abs/1903.10082, 2019. [Online]. Available: http://arxiv.org/abs/1903.10082

[39] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[40] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," *CoRR*, vol. abs/1901.05945, 2019. [Online]. Available: http://arxiv.org/abs/1901.05945

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

[42] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arxiv*, 2016.

[44] 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*. Springer, 2015, pp. 234–241.

[45] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections," *arXiv preprint arXiv:1603.09056*, 2016.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[47] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, and Q. Dai, "A fast uyghur text detector for complex background images," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3389–3398, 2018.

[48] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *European Conference on Computer Vision*. Springer, 2016, pp. 154–169.

[49] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, and Q. Dai, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles," *IEEE transactions on intelligent transportation systems*, vol. 19, no. 1, pp. 284–295, 2018.

[50] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, and Q. Dai, "Effective uyghur language text detection in complex background images for traffic prompt identification," *IEEE transactions on intelligent transportation systems*, vol. 19, no. 1, pp. 220–229, 2018.

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

[52] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

[53] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[54] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522–3533, 2015.

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

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

He Zhang [S'14] received his Ph.D. degree in Electrical and Computer Engineering from Rutgers University, NJ, USA in 2018. He is currently a Research Scientist in Adobe, CA. His research interests include image restoration, image compositing, generative adversarial network, deep learning and sparse and low-rank representation.

Vishwanath Sindagi [S'16] is a PhD student in the Dept. Of Electrical & Computer Engineering at The Johns Hopkins University. Prior to joining Johns Hopkins, he worked for Samsung R&D Institute-Bangalore. He graduated from IIIT-Bangalore with a Master's degree in Information Technology. His research interests include deep learning based crowd analytics, object detection, applications of generative modeling, domain adaptation and low-level vision. Attachments area

Vishal M. Patel [SM'15] is an Assistant Professor in the Department of Electrical and Computer Engineering (ECE) at Johns Hopkins University. Prior to joining Hopkins, he was an A. Walter Tyson Assistant Professor in the Department of ECE at Rutgers University and a member of the research faculty at the University of Maryland Institute for Advanced Computer Studies (UMIACS). He completed his Ph.D. in Electrical Engineering from the University of Maryland, College Park, MD, in 2010. He has received a number of awards including the 2016 ONR Young Investigator Award, the 2016 Jimmy Lin Award for Invention, A. Walter Tyson Assistant Professorship Award, Best Paper Award at IEEE AVSS 2017, Best Paper Award at IEEE BTAS 2015, Honorable Mention Paper Award at IAPR ICB 2018, two Best Student Paper Awards at IAPR ICPR 2018, and Best Poster Awards at BTAS 2015 and 2016. He is an Associate Editor of the IEEE Signal Processing Magazine and serves on the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is serving as the Vice President, Conferences of the IEEE Biometrics Council. He is a member of Eta Kappa Nu, Pi Mu Epsilon, and Phi Beta Kappa.