Joint Transmission Map Estimation and Dehazing using Deep Networks

He Zhang, *Student Member, IEEE*, Vishwanath A. 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 superior 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 directly from the loss function. Extensive experiments 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 or fog is a natural atmospheric phenomenon caused by the absorption or reflection of light by floating particles in the air. In the presence of haze, the captured images tend to suffer from low contrast and faint color conditions. A sample hazy image is shown on the left side of Figure 1. It can be clearly observed that haze greatly obscures the content in the image. The problem of estimating a clear image from a single hazy input image (known as dehazing) has attracted a significant interest in the computer vision and image processing communities in recent years [1], [2], [3], [4], [5], [6], [7].

The image degradation due to the presence of haze is mathematically formulated as:

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

where $\mathbf{x}$ is the pixel coordinates, $\mathbf{I}$ represents the observed hazy image, $\mathbf{J}$ is the true scene radiance (image before degradation), $\mathbf{A}$ is the global atmospheric light, and $t(\mathbf{x})$ is the transmission map [8]. The transmission map is a distance-dependent factor that affects the fraction of light that reaches the camera sensor. 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 $\mathbf{A}$. When the atmospheric light $\mathbf{A}$ is homogeneous, the transmission map can be expressed as

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

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

where β represents the attenuation coefficient of the atmosphere and $d(\mathbf{x})$ is the scene depth.

Since multiple solutions can be found to the same input hazy image, haze removal is a highly ill-posed problem. Many previous methods overcome this issue by including more information 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, some prior information has to be included into the optimization framework such as dark-channel prior [5], contrast color-lines [9] and haze-line prior [4]. 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]. 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 recover the dehazed image 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, they consider the task of transmission map estimation and dehazing as two separate tasks. By doing so,

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

they are unable to accurately capture the transformation between the transmission map and the dehazed image. Motivated by this observation, we relax the constant atmospheric light assumption [10], [11] and propose to jointly learn the transmission map and dehazed image from an input hazy image using a deep CNN-based network. Relaxing the constancy assumption 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), however, after relaxing the assumption of constant atmospheric light. 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 poor quality dehazed images. To tackle this issue, we incorporate adversarial loss to generate the transmission map.

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 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.
- Extensive experiments on synthetic and real image datasets are conducted to demonstrate the effectiveness of the proposed method.

II. Related Work

In this section, we review recent related works on single image dehazing and some commonly used losses in various CNN-based image reconstruction tasks.

A. Single Image Dehazing

One of the earliest methods for single image dehazing was proposed in [12], where the authors aim to maximize the contrast per-patch based on the observation that haze or fog reduces the contrast of the color images. Then, Kratz and Nishino [13] proposed a factorial MRF model to estimate the albedo and depths filed. He. *et.al* in [5] made a surprising observation that the images captured in the outdoor environments contain a lot of dark pixels. Motivated by this observation, they proposed a dark-channel prior-based model to estimate the transmission map. Meng *et al.* [14] extended this dark-channel prior model by imposing an inherent boundary constraint on the transmission function to better estimate the transmission map. Tang [3] argued that a single prior is not sufficient to tackle the ill-posed nature of the image dehazing problem and they proposed a learning-based methods to combine different types of features to generate better transmission maps. More recently, Berman *et al.* [4] proposed a non-local patch prior method based on the assumption that the colors of a haze-free image are well approximated by a few hundred distinct colors that form tight clusters in the RGB space.

The success of CNNs for high-level vision tasks such as object recognition has inspired researchers to explore CNN-based algorithms for low-level vision tasks such as image restoration and super-resolution [15], [16], [17], [1], [18]. Unlike previous methods that use different priors to estimate the transmission map, Cai *et al.* [2] train an end-to-end CNN network for estimating the transmission map given an input hazy image. Most recently, Ren *et al.* [1] proposed a multi-scale deep neural network to learn the mapping between hazy images and their corresponding transmission maps. 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, in this paper 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.

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 [19], [20]. Since these losses operate at pixel level, their ability to capture high level perceptual/contextual details is 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.

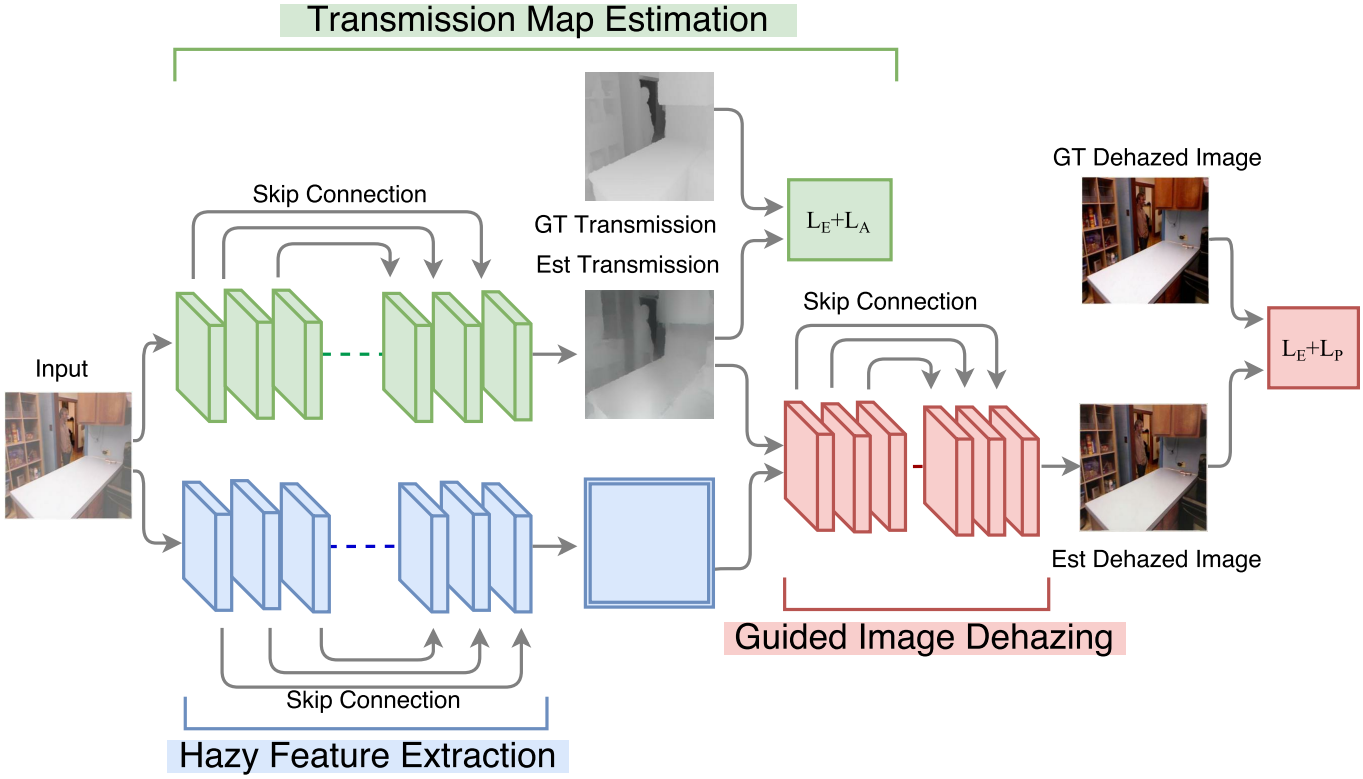


Fig. 2: Overview of the proposed multi-task method for image dehazing. The proposed method consists of three modules: (a) Hazy feature extraction, (b) Transmission map estimation, and (c) Guided image dehazing. First, the transmission map is estimated from the input hazy image and it is concatenated 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.

Adversarial loss: Adversarial loss is based on the recently introduced GAN framework that was proposed by Goodfellow *et al.* in [21] to synthesize realistic images by effectively learning the distribution of training images. The authors adopted a game theoretic min-max optimization framework to simultaneously train two models: a generative model, G, and a discriminative model, D. The success of GANs in synthesizing realistic images has led to researchers exploring the adversarial loss for numerous low-level vision applications such as style transfer [22], image in-painting [23], image to image translation [24], image super-resolution [25] and image de-raining [26]. Inspired by the success of these methods, we propose to use the adversarial loss to learn the distribution of transmission maps for their accurate estimation.

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 [27], [28]. The perceptual function is usually defined using high-level features extracted from a 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 [22], [29], [30]. In this work, a VGG-16 architecture [31] 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 guided by the transmission map and feature maps from the first two modules. 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 [32]. Similar to earlier works on

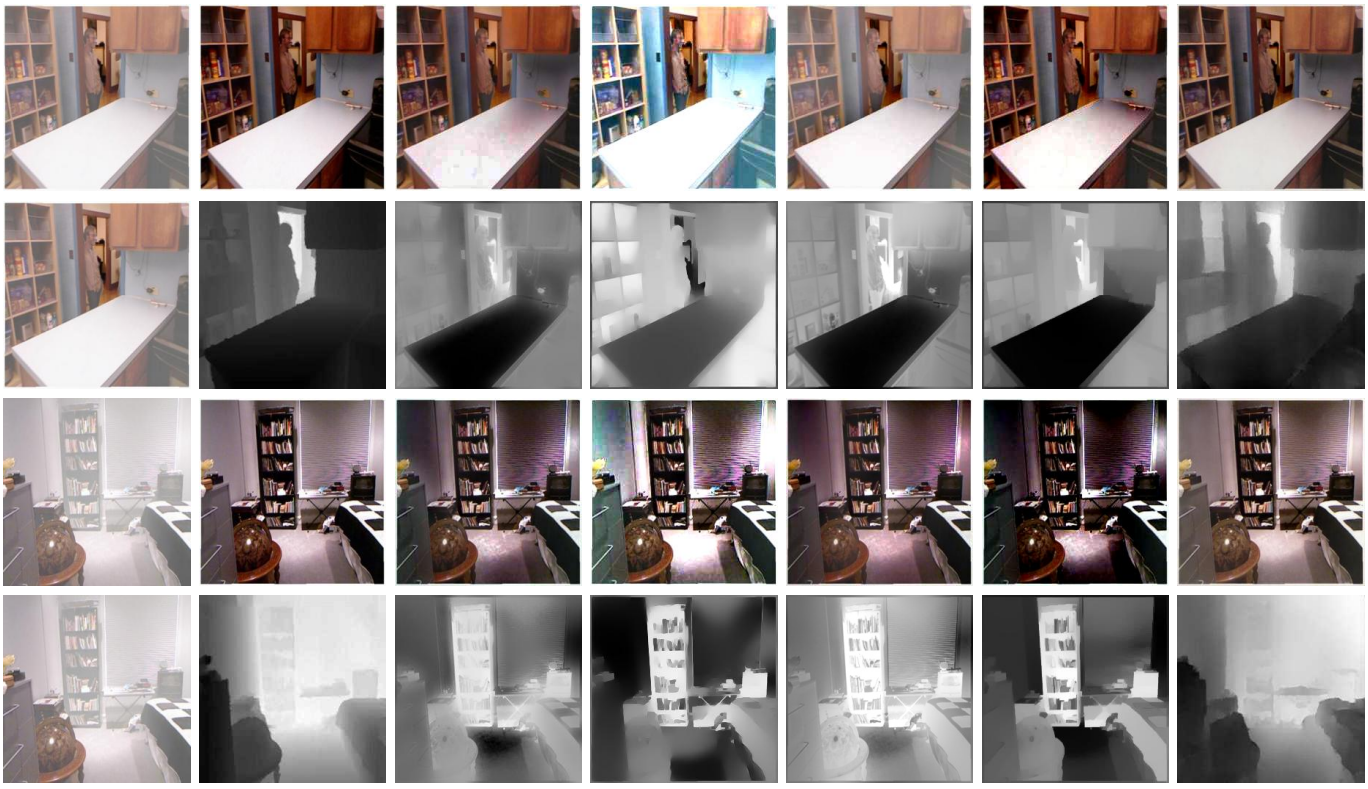


Fig. 3: Qualitative results of the transmission map estimation and dehazing evaluated on the synthetic test dataset.

GANs for image reconstruction tasks [26], [33], [25], 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 [34], [16], [33], [35]. In this work, we adopt a 'U-Net'-based structure [34] as the generator for the transmission map estimation. Rather than concatenating the symmetric layers during training, shortcut connections [36] 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 covolutional layer, TC represents transpose convolution layer, P indicates Prelu [37] and B indicates batch-normalization [38]. The number in the bracket represents the number of output feature maps of the corresponding layer.

The structure of the proposed discriminator network is as follows:

CB(48)-CBP(96)-CBP(192)-CBP(384)-CBP(384)-C(1)-Sigmoid.

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 [33], while learning a mapping from label-like image to a RGB color image, it is difficult to maintain its original color information and hence it is not possible to accurately predict the dehazed output even though the learned mapping can generate visually reasonable outputs.

To ensure that the dehazed image maintains color constancy as compared with the input image, we leverage the input hazy image by using it to guide the transmission map. Inspired by *guided filtering* [39], [40], [41], where a guidance image is used to filter the transmission map,a set of convolutional layers with symmetric skip connections are stacked in the front and they serve as a hazy image feature extractor. These feature maps extracted from the input image are concatenated with the estimated transmission map and are 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 [31] 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.

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 and an adversarial error. The transmission map loss is defined as

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

where λ_a is a weighting factor, L^t_E is the pixel-wise Euclidean loss and L^t_A is the adversarial loss and are respectively defined as follows

$$L^t_E = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\|(\phi_G(\mathbf{I}))^{w,h} - \mathbf{y}_t^{w,h}\|_2, \tag{5}$$

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

where $\mathbf{I}$ is a C-channel input hazy image, $\mathbf{y}_t$ is the ground truth transmission map, $W \times H$ is 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.

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

$$L^d = L^d_E + \lambda_p L^t_P, \tag{7}$$

where λ_p is a weighting factor, L^d_E is the pixel-wise Euclidean loss and L^t_P is the perceptual loss and are respectively defined as

$$L^d_E = \frac{1}{CWH}\sum_{c=1}^{C}\sum_{w=1}^{W}\sum_{h=1}^{H}\|\phi_E(\mathbf{I})^{c,w,h} - \mathbf{J}^{c,w,h}\|_2, \tag{8}$$

$$L^d_P = \frac{1}{C_iW_iH_i}\sum_{c=1}^{C_i}\sum_{w=1}^{W_i}\sum_{h=1}^{H_i}\|V(\phi_E(\mathbf{I}))^{c,w,h} - V(\mathbf{J})^{c,w,h}\|_2, \tag{9}$$

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 [28], 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 [31].[3] Note that the dehazing loss L^d is also to be propagated to the transmission estimation part.

IV. Experiments

In this section, we demonstrate the effectiveness of the proposed approach by conducting various experiments on synthetic and real datasets that contain a variety of hazy conditions. First we describe the datasets used in our experiments. We then discuss the details of the training procedure. Next, we discuss the results of the 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.

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 [42]. 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 [42] and hence depths of all the images

[2]Note that our network is quite different from the network proposed in [39] in the sense that the proposed network is a multi-task learning network with a single input while the network in [39] is a single-task network with two inputs.

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

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 around 30 hazy images downloaded from the Internet.

B. Training Details

The entire network is trained on a Nvidia Titan-X GPU using the torch framework [43]. We choose $\lambda_a = 0.003$ 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 [44] 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.

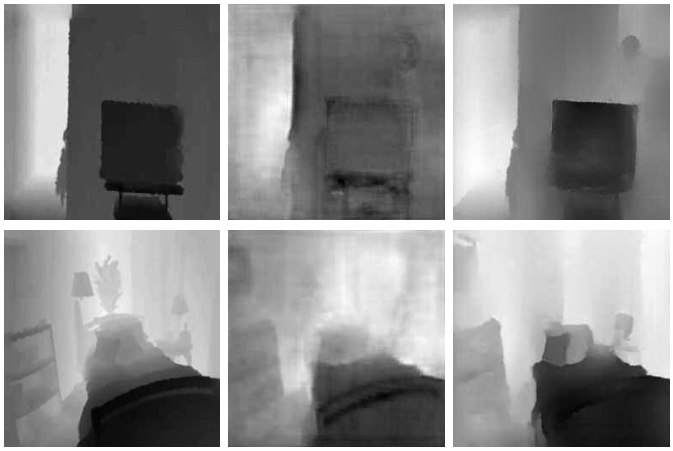
C. Ablation Study

In order to better demonstrate the improvements obtained by different modules in the proposed network, we perform an ablation study involving the following experiments: 1) *Transmission map estimation with and without adversarial loss*, 2) *Image dehazing with and without perceptual loss*, 3) *Image dehazing with and without Euclidean loss*, and 4) *Image dehazing with and without transmission map*.

Transmission map estimation without adversarial loss: In this experiment, the transmission map estimation module is trained in two different settings: 1) *Without adversarial loss*: The network is trained using only pixel-wise Euclidean loss given by (5), and 2) *With adversarial loss*: The network is trained using a combination of pixel-wise Euclidean loss and adversarial loss given by (4). The results of the network trained in these two settings for two sample images from the test dataset are shown in Figure 4. It can be clearly observed from this figure that the transmission maps obtained by using a weighted combination of Euclidean loss and adversarial loss are much sharper and more visually pleasing as compared to those obtained using only Euclidean loss.

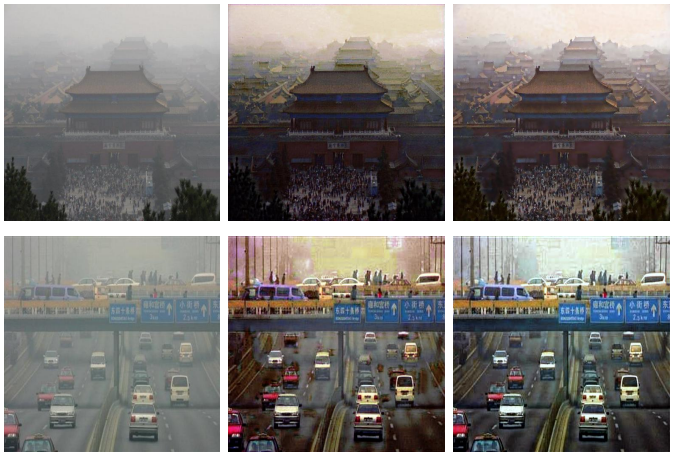
Image dehazing without perceptual loss: In this experiment, the effect of using perceptual loss with the help of VGG network is studied. As before, the entire network is trained in two settings: 1) *Without perceptual loss*: The network is trained using only the Euclidean loss given by (8), and 2) *With perceptual loss*: The network is trained using a weighted combination of Euclidean loss and perceptual loss given by (7). The results from these two settings on a sample image from the test dataset are shown in Figure 5. It can be observed clearly that minimizing only the Euclidean loss results in loss of details and make the dehazed result visually unappealing.

Image dehazing without Euclidean loss: In this experiment, the effect of using perceptual plus Euclidean loss is studied. As before, the entire network is trained in two settings: 1) *Without Euclidean loss*: The network is trained using only the perceptual loss given by (9), and 2) *With Euclidean loss*: The network is trained using a weighted combination of Euclidean loss and perceptual loss given by (7). The results from these two settings on a sample image from the test dataset are shown in Figure 6. It can be clearly observed that minimizing only perceptual loss tends to change the color of the dehazed result (It can be better seen from the recovered sky).



(a) (b) (c)

Fig. 4: Transmission map estimation with/without adversarial loss. (a) Ground truth. (b) Results without adversarial loss. (c) Results with adversarial loss.



(a) (b) (c)

Fig. 5: Image dehazing with/without perceptual loss. (a) Input image. (b) Results without perceptual loss. (c) Results with perceptual loss.

Image dehazing without transmission map: As the proposed method involves multi-task learning of transmission map estimation and dehazing, we evaluate the importance of enforcing the network to learn and use transmission map for guided dehazing by learning the network in two settings: 1) *Without transmission map estimation*: The network is trained to perform an end-to-end image dehazing without transmission map estimation, and 2) *With transmission map estimation*. The results corresponding to these two settings for two sample images from the test dataset are shown in Figure 7. It can be clearly observed that in the first setting, the network is unable

	Input	He. *et al.* [5]	Meng. *et.al* [14]	Cai, *et al.* [2]	Berman. *et al.* [4]	Ren. *et al.* [1]	Our
PSNR (dB)	13.60	18.01	16.56	16.45	18.12	18.09	**18.68**
SSIM	0.7276	0.8053	0.6896	0.6982	0.8159	0.8135	**0.8255**

TABLE I: Quantitative results on the synthetic test dataset consisting of 300 images.

(a) (b) (c)

Fig. 6: Image dehazing with/without Euclidean loss. (a) Input image. (b) Result without Euclidean loss. (c) Result with Euclidean loss.

(a) (b) (c)

Fig. 7: Image dehazing with/without transmission map. (a) Input image. (b) Results without transmission map. (c) Result with transmission map.

to remove the haze completely as compared to the second network.

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 on synthetic and real datasets.

Evaluation on synthetic dataset: The proposed network is trained and evaluated using the synthetic dataset described in Section 4.1. Since the dataset is synthesized, the ground truth dehazed images are available, hence making it possible for us perform qualitative as well as quantitative evaluations.

Results for the proposed method and four recent state-of-the-art methods [5], [14], [4], [2] on two sample images from the test dataset are shown in Figure 3. It can be observed from these results that the dark-channel prior-based methods [5], [14] tend to overestimate the dark-pixels resulting in lower color contrast in the output images. This is especially evident when one looks at the recovered carpet in the third row. Even though, [4] is able to achieve good performance in the presence of moderate haze, its performance reduces in the presence of heavy haze as shown in the fourth row in Figure 3. 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], [14], [1], [4], [2] using two standard metrics: Peak Signal to Noise Ratio (PSNR) and SSIM [45]. The quantitative results are tabulated in Table I. It can be observed from this table that the proposed method achieves the best performance in terms of both metrics. 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] cannot estimate the predicted transmission map, the result for the transmission estimation corresponding to [1] is not included in the discussion.

Evaluation on real dataset: To better demonstrate the generalization ability of the proposed method in dealing with real-world images, we evaluated the proposed method on several real-world hazy images provided by previous methods and also some challenging hazy images downloaded from the Internet. Note that the network is trained on the synthetic training dataset, which is generated by the dataset discussed above and tested on the real-world dataset. The dehazing results are compared against five state-of-art methods [5], [14], [4], [2], [1].

Comparison of results on four sample images used in earlier methods compared with various approaches is shown in Figure 9. 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], [2] underestimate the thickness of haze resulting in partial dehazing. Even though Berman *et al.* [4] 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

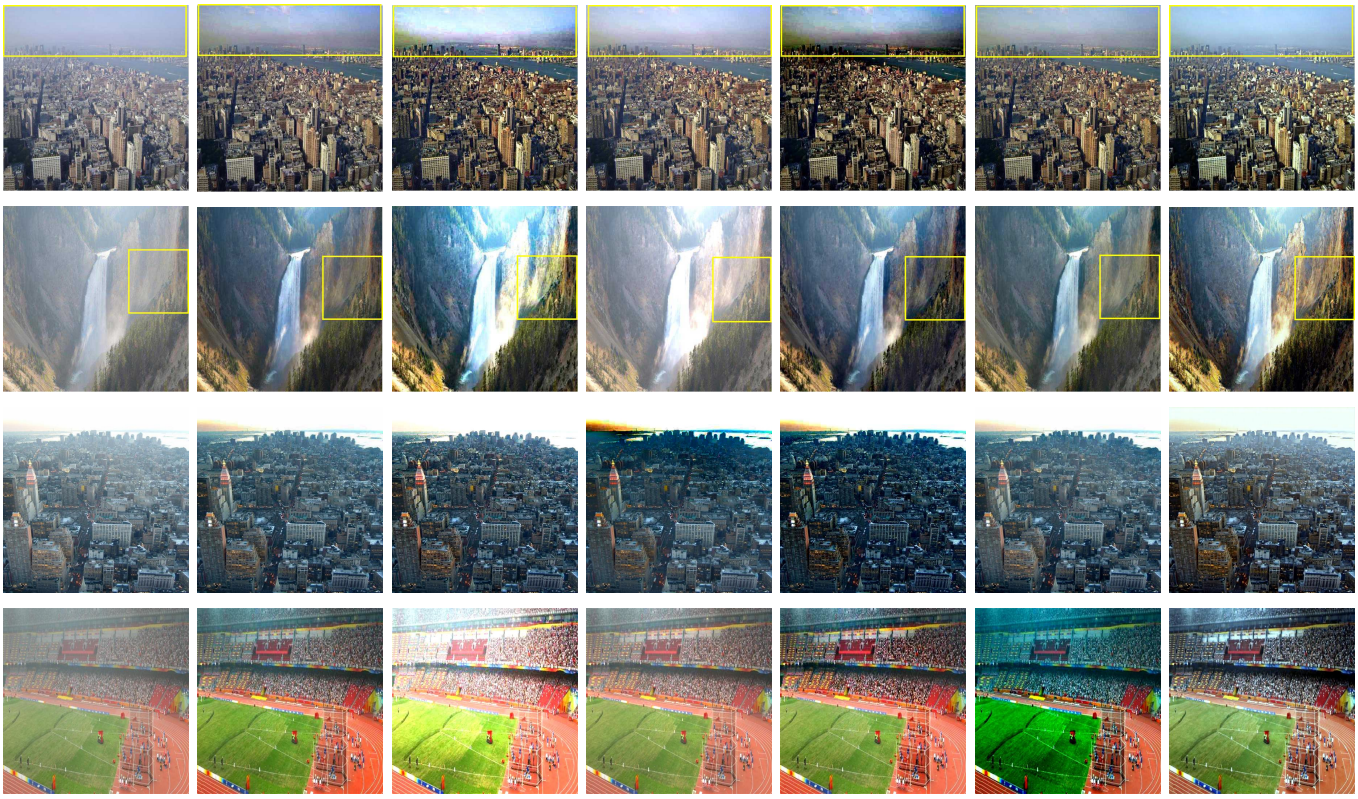


Fig. 8: Qualitative comparison of dehazing on real-world dataset that is presented in previous dehazing papers. Top row: Results on a sample image from the real-world dataset provided by previous methods.

methods. He. *et al* [5], Cai. *et al* [2] 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 Meng. *et al* and Berman *et al* 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 10, 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.

Finally, the running time of our method is also evaluated. On average, our method can processes 512×512 images at 18 frames per second (fps), making it feasible to run in real-time.

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 evaluated on synthetic and real datasets demonstrated that the proposed method is able to achieve significantly better results as compared to the recent state-of-the-art methods. In addition, an ablation study was performed to demonstrate the improvements obtained by different modules in the proposed method.

Aknowledgement

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.

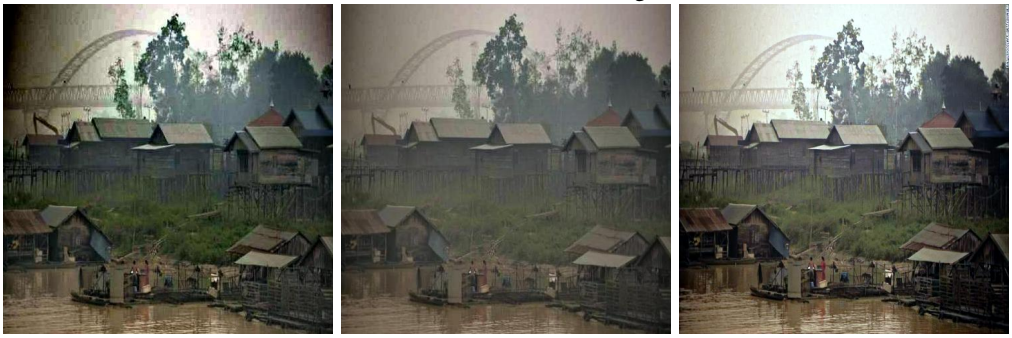
[3] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *CVPR*, 2014, pp. 2995–3000.

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

Input He. *et al.* [5] Meng. *et al.* [14] Cai. *et al.* [2]



Berman. *et al.* [4] Ren. *et al.* [1] Our

Input He. *et al.* [5] Meng. *et al.* [14] Cai. *et al.* [2]

Berman. *et al.* [4] Ren. *et al.* [1] Our

Fig. 9: Qualitative comparison of dehazing on real-world dataset. Results on two sample images from a set of images downloaded from the Internet.

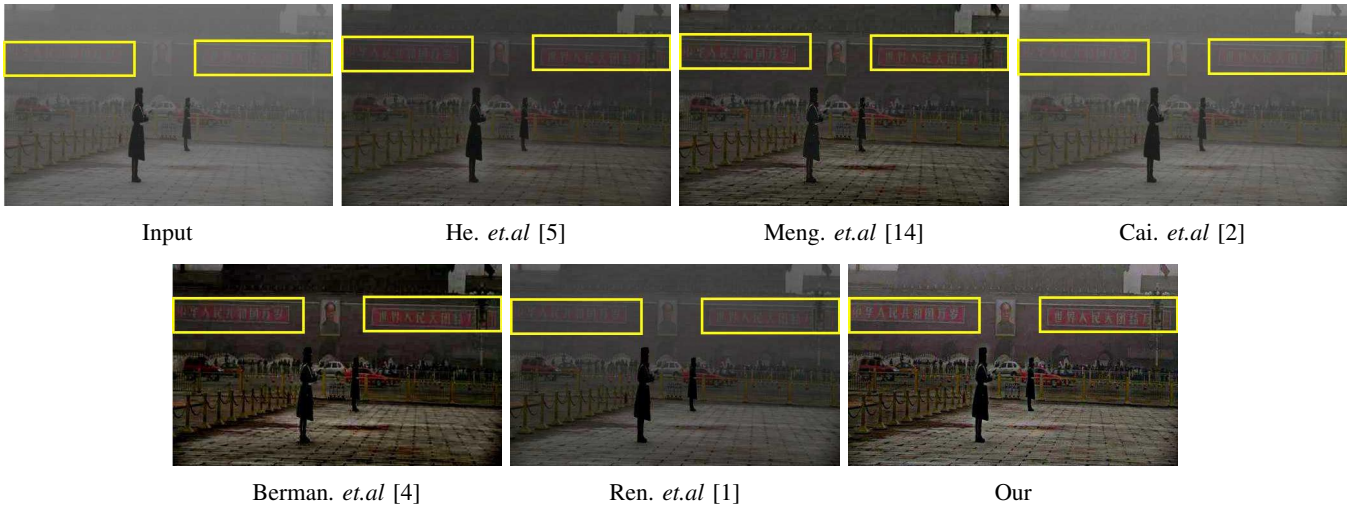


Fig. 10: Qualitative comparison of dehazing on real-world dataset. Top row: Results on a sample image from the real-world dataset provided by previous methods. Bottom two rows: Results on two sample images from a set of images downloaded from the Internet.

[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] 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

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

[10] 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.

[11] J.-P. Tarel, N. Hautiere, L. Caraffa, A. Cord, H. Halmaoui, and D. Gruyer, "Vision enhancement in homogeneous and heterogeneous fog," *IEEE Intelligent Transportation Systems Magazine*, vol. 4, no. 2, pp. 6–20, 2012.

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

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

[14] 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.

[15] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV*. Springer, 2016, pp. 391–407.

[16] 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.

[17] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Joint rain detection and removal via iterative region dependent multi-task learning," *CoRR*, vol. abs/1609.07769, 2016. [Online]. Available: http://arxiv.org/abs/1609.07769

[18] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

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

[20] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*, 2015, pp. 2650–2658.

[21] 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.

[22] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *ECCV*, 2016, pp. 702–716.

[23] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.

[24] 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.

[25] 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.

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

[27] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666.

[28] 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.

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

[30] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," *arXiv preprint arXiv:1703.06953*, 2017.

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

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

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

[34] 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.

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

[36] 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.

[37] ——, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[38] 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*, 2015, pp. 448–456.

[39] 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.

[40] X. Shen, C. Zhou, L. Xu, and J. Jia, "Mutual-structure for joint filtering," in *ICCV*, 2015, pp. 3406–3414.

[41] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *ICCV*, 2013, pp. 993–1000.

[42] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.
[43] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.
[44] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[45] 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.