Towards Perceptual Image Dehazing by Physics-Based Disentanglement and Adversarial Training

Xitong Yang, Zheng Xu
Department of Computer Science
University of Maryland
College Park, MD 20740
{xyang35, xuzh}@cs.umd.edu

Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, NY 14627
jluo@cs.rochester.edu

Abstract

Single image dehazing is a challenging under-constrained problem because of the ambiguities of unknown scene radiance and transmission. Previous methods solve this problem using various hand-designed priors or by supervised training on synthetic hazy image pairs. In practice, however, the predefined priors are easily violated and the paired image data is unavailable for supervised training. In this work, we propose *Disentangled Dehazing Network*, an end-to-end model that generates realistic haze-free images using only unpaired supervision. Our approach alleviates the paired training constraint by introducing a physical-model based disentanglement and reconstruction mechanism. A multi-scale adversarial training is employed to generate perceptually haze-free images. Experimental results on synthetic datasets demonstrate our superior performance compared with the state-of-the-art methods in terms of PSNR, SSIM and CIEDE2000. Through training on purely natural haze-free and hazy images from our collected HazyCity dataset, our model can generate more perceptually appealing dehazing results.

Introduction

Images with clear visibility is desirable for most current computer vision applications.However, images taken in outdoor scenes usually suffer from visual degradation due to the presence of aerosols in the atmosphere. These small particles, as the constituents of haze, attenuate and scatter the light in the atmosphere and affect the visibility of the image (Narasimhan and Nayar 2002). As a result, image dehazing, especially single image dehazing, is highly desirable and the problem has been extensively studied over the past ten years (Fattal 2008; He, Sun, and Tang 2011; Tang, Yang, and Wang 2014; Zhu, Mai, and Shao 2015; Cai et al. 2016; Berman, Avidan, and others 2016; Li et al. 2017a).

Many of the successful approaches rely on strong priors or assumptions to estimate the medium transmission map, for example, the well-known *Dark-Channel Prior* (He, Sun, and Tang 2011) and the more recent *Non-local Color Prior* (Berman, Avidan, and others 2016). However, these priors can be easily violated in practice, especially when the scene is complex or contains irregular illumination. For instance, the dark-channel prior does not hold for areas that are similar to the atmospheric light. This usually leads to unsatisfied dehazing quality for sky regions or bright objects. The non-local color prior, on the other hand, may fail in scenes where the airlight is significantly brighter than the scene.

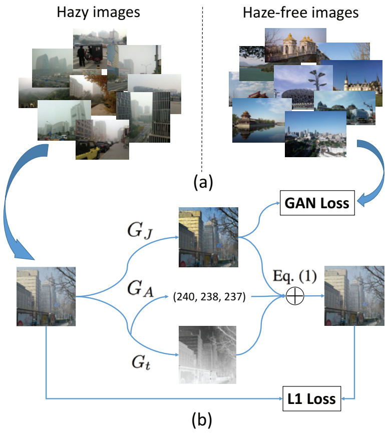

Figure 1: (a) Unpaired dataset with natural hazy images and haze-free images. (b) Overall architecture of our Disentangled Dehazing Network. G_J, G_t, G_A indicate the generators for the scene radiance, the medium transmission and the global atmosphere light, respectively.

Recent works propose to use convolutional neural networks to estimate the transmission map or even the scene radiance directly (Cai et al. 2016; Ren et al. 2016; Li et al. 2017a). Although these methods can obtain encouraging results by virtue of the modeling power of CNN, they are restricted in practical application because the model training requires a large amount of "paired" data for supervision. To create the image pairs for training, (Cai et al. 2016) and (Ren et al. 2016) synthesize hazy local patches (16×16) under the assumption that image content is independent of the locally constant transmission within a patch. This strategy is sub-optimal as the training process only considers very local information and loses the global correlation. (Li et al. 2017a) synthesizes hazy images using the depth meta-data from an indoor dataset. However, it is almost impossible to obtain the ground-true depth information in most real-world scenarios. It is also arguable that using synthetic hazy images for training can lead to perceptually unsatisfactory solutions because the synthesized haze cannot faithfully represent the true haze distribution in real cases.

An effective and practical dehazing model should be able to learn the mapping from hazy images to haze-free images *without* using paired supervision. Moreover, the dehazed images should be perceptually consistent with the haze-free images perceived as such by humans. In this paper, we proposed *Disentangled Dehazing Network*, a novel weakly-supervised model that satisfies the above criteria. As shown in Figure 1, our model introduces a physical-model based disentanglement and reconstruction mechanism: the hazy image input is first disentangled into three hidden factors, the scene radiance, the medium transmission and the atmosphere light, by three generator networks; these factors are then combined to reconstruct the original input using the physical model (Eq. 1). The hidden factors are also constrained by adversarial loss and regularizations. The whole framework, sharing similar ideas with CycleGAN (Scharstein et al. 2014) and the recent AIGN (Tung et al. 2017), relieves the constraint of paired training by utilizing the feedback signal from a backward/rendering process. In contrast to their approaches, our disentanglement mechanism enables us to introduce separate constraints on different hidden factors and learn a physically valid model.

Our *Disentangled Dehazing Network* provides a new viewpoint for image dehazing in realistic scenes, which we call *perceptual dehazing*. Unlike previous methods that view haze removal as an image restoration process and try to fully recover the original scene radiance, our objective is to generate perceptually haze-free images that are visually pleasing to humans. In fact, it is not only challenging but also unnecessary to restore the true scene radiance in most practical scenarios. First, images of outdoor scenes can contain heterogeneous atmosphere, complex scenes and irregular illumination (see examples in Figure 4), which makes the estimation of true medium transmission unreliable. Second, removing the haze thoroughly can cause unnatural image as the presence of haze is a cue for human to perceive depth (He, Sun, and Tang 2011). As a result, we aims to generate perceptually pleasing dehazing results that fit the distribution of human-perceived haze-free images.

We make the following contributions in this paper:

- We propose a novel image dehazing approach based on *Disentangled Dehazing Network*, which is trained by adversarial process and performs physical-model based disentanglement.
- We collect a challenging dataset for image dehazing research, with more than 800 natural hazy images and 1000 haze-free images of outdoor scenes.
- We evaluate perceptual image dehazing through extensive experiments on both synthetic and real image datasets.

Related Work

Single image dehazing is a challenging while important problem as the existence of haze dramatically degrades the visibility of the original scenes and hinders most high-level computer vision tasks. The success of many previous methods rely on using strong priors or assumptions. (Tan 2008) improves the image visibility by maximizing the local contrast, with the observation that haze-free images must have higher contrast. (Fattal 2008) proposes to infer the transmission by estimating the albedo of the scene, under the assumption that the transmission and surface shading are locally uncorrelated. The seminal work (He, Sun, and Tang 2011) proposes the dark-channel prior which can estimate the transmission map effectively in general cases. (Zhu, Mai, and Shao 2015) proposes a color attenuation prior and develops a linear model to estimate the scene depth using the brightness and saturation of the image as the input. Instead of using local priors, (Berman, Avidan, and others 2016) proposed a non-local color prior, with the observation that a haze-free image can be faithfully represented with only a few hundreds of distinct colors. These image processing methods are designed by researchers based on priors and assumptions, which require no training from labeled data.

Data-driven dehazing models recently become popular due to the success of machine learning in various vision applications (Tang, Yang, and Wang 2014). Particularly, neural networks have been trained to estimate the transmission map from hazy input images (Cai et al. 2016; Ren et al. 2016). In a more recent work, (Li et al. 2017a) proposed an end-to-end network that directly output dehazed images from hazy inputs. In contrast to our approach, these models have to be trained in a supervised fashion and require a large amount of paired images or depth information.

Generative adversarial networks (GANs) have become one of the most successful methods for image generation and manipulation since (Goodfellow et al. 2014). In GANs, two networks are adversarially trained at the same time, where the discriminator is updated to distinguish the real samples and the output of the generator, and the generator is updated to output fake data to fool the discriminator. Particularly, conditional adversarial networks, in which the generator is conditioned on the input images, have been applied to several image-to-image translation tasks (Isola et al. 2017; Ledig et al. 2017). An image can be translated into an output by the conditional generator trained from paired samples, such as images of a scene in day and night. When unpaired samples are provided, cycleGAN (Zhu et al. 2017), discoGAN (Kim et al. 2017), dualGAN (Yi et al. 2017) and (Liu, Breuel, and Kautz 2017) trained multiple generators and discriminators together to tackle the tasks; AIGN (Tung et al. 2017) requires an implicit alignment between input and output to be defined by the user. WaterGAN (Li et al. 2017b) uses a physical model to generate synthetic underwater image from in-air image. Different from previous works, we disentangle a hazy image into hidden factors based on the physical model, and the adversarial loss is used to regularize the distribution of the disentangled haze-free images.

GANs have also been applied to disentangle latent factors in computer vision, such as pre-defined attributes of images. The disentangled factors are represented by an embedded vector and later used as input of the generative neural network to control the output images (Mathieu et al. 2016;

Chen et al. 2016; Tran, Yin, and Liu 2017; Fu et al. 2017; Donahue et al. 2017). (Shu et al. 2017) decomposes face images according to the intrinsic face properties and performs face editing by traversing the manifold of the disentangled latent spaces. A preliminary draft (Tung and Fragkiadaki 2016) decomposes an image into several components by retrieving training samples from a dataset of all the components. In our approach, the disentanglement is derived from the physical model for the hazy image generation process, and the hidden factors are generated from hazy images by neural networks. Different from (Shu et al. 2017), our approach does not rely on any external algorithm or paired data for extra supervision, and the disentangled components are only constrained by adversarial loss and priors.

Our Model

In this section, we first review the physical model of hazy images. We then introduce our *Disentangled Dehazing Network* and adversarial training in detail. The network architectures and implementation details used in our experiments are also described.

Physical Model

In computer vision and computer graphics, the *atmosphere scattering model* has been widely used as the description for the hazy image generation process (Narasimhan and Nayar 2000; Fattal 2008; He, Sun, and Tang 2011):

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

where $I(x)$ is the observed hazy image, $J(x)$ is the scene radiance (haze-free image). $t(x)$ is the medium transmission map, and A is the global atmospheric light. The goal of image dehazing is to recover $J(x)$, $t(x)$ and A from $I(x)$.

The medium transmission map $t(x)$ describes the portion of the light that is not scattered and reaches the camera. When the atmosphere is homogeneous, $t(x)$ can be expressed as a function of the scene depth $d(x)$ and the scattering coefficient β of the atmosphere :

$$t(x) = \exp^{-\beta d(x)}. \tag{2}$$

Image dehazing is an under-constrained problem if the input is only a single haze image because we need to estimate all the three components at the same time. In other words, the presence of haze introduces *ambiguities* for inferring the radiance and the depth of the scene.

Estimating the medium transmission map is known as the key to achieving haze removal (He, Sun, and Tang 2011; Tang, Yang, and Wang 2014; Cai et al. 2016). However, it is very challenging to estimate the true $t(x)$ in many realistic outdoor scenes because of the heterogeneous atmosphere, complex scenes and irregular illumination. In our model, we propose to generate the haze-free image and the medium transmission map *simultaneously* via disentanglement and reconstruction, through which each estimation can benefit from the other.

Disentangled Dehazing Network

We cast the perceptual image dehazing problem as a *unpaired image-to-image translation* problem, in which images in the source domain (hazy) are mapped to the target domain (haze-free) without any paired information. The problem is challenging because without paired supervision, a model can learn arbitrary mapping to the target domain and cannot guarantee to map an individual input to its desired output. Previous works solve this problem by introducing an extra backward generator to generate the original input (Zhu et al. 2017; Yi et al. 2017; Kim et al. 2017). Although these approaches can be applied to our task, we found that they fail to tackle the ambiguities introduced by haze, i.e., they cannot distinguish the effect of light attenuation from the original scene radiance. To better model the formation of hazy images, we propose to solve the unpaired image-to-image problem by introducing the physical-model based disentanglement and reconstruction.

Figure 1 (b) shows the overall framework of our approach. Given two sets of unpaired images (hazy and haze-free) as (weak) supervision, our goal is to learn a model that can disentangle the hazy input into hidden factors, under the constraint of the physical model. The hidden factors are further constrained by an adversarial training procedure and prior-based regularizations. Our approach facilitates unpaired training for the following reasons: 1) It enables separate constraints/priors on different disentangled factors. 2) Different generators can be optimized jointly to achieve the best disentanglement. 3) The reconstruction process is physically valid and provides a harder constraint on the generation procedure.

Inspired by the atmosphere scattering model, we disentangle the input hazy image into three hidden factors: the scene radiance, the transmission map and the global atmosphere light. These three components are then combined using Eq. (1) to reconstruct the original hazy image. Formally, let us denote $\mathcal{I} = \{I_i\}_{i=1}^N$ and $\mathcal{J} = \{J_j\}_{j=1}^M$ as the two sets of training samples corresponding to hazy images and haze-free images, respectively. Our model first performs disentanglement using three generators: $\hat{J} = G_J(I)$, $\hat{t} = G_t(I)$ and $\hat{A} = G_A(I)$. The three components are then composited to reconstruct the original input: $\hat{I} = \hat{J} \odot \hat{t} + \hat{A} \odot (1 - \hat{t})$, where $\odot$ indicates element-wise multiplication.

Our objective function contains three terms: a reconstruction loss, an adversarial loss and a regularization loss. We use the traditional L_1 losses as the reconstruction loss to encourage both pixel-level consistency and less blurring (compared with L_2 loss):

$$\mathcal{L}_{recon}(G_J, G_t, G_A) = \mathbb{E}_{I \sim \mathcal{I}} \|I - \hat{I}\|_1. \tag{3}$$

In order to generate both perceptually pleasing and haze-free images, we introduce a multi-scale adversarial training procedure for the intermediate output $\hat{J}_x$. Specifically, while the multi-scale discriminator D is trained to detect whether an image is "real" or "fake", the generator G_J is adversarially trained to "fool" the discriminator. Same as the setting in Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), here "real" data means the images sampled from the target domain (haze-free images) and "fake" data means the images generated from samples of the source domain (hazy images). The classical GAN loss can be described as:

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{J\sim\mathcal{J}}\left[\log D(J)\right] + \mathbb{E}_{I\sim\mathcal{I}}\left[\log(1 - D(G(I)))\right].$$

Inspired by (Isola et al. 2017) and (Zhu et al. 2017), we use a patch-level discriminator to distinguish the real and fake images. Different from their approaches which choose a compromised receptive field size (RFS) for balancing the trade-off between the sharpness and artifacts of the result, we proposed to use a multi-scale discriminator that combines a local discriminator (small RFS) and a global discriminator (large RFS). While the local discriminator focuses on modeling high-frequency structure that is beneficial for texture/style recognition, the global discriminator can incorporate more global information and alleviate the tiling artifacts. Our multi-scale discriminator combines the best of the two worlds, as shown in Figure 2.

As a result, our multi-scale adversarial loss is:

$$\mathcal{L}_{adv}(G_J, D) = \frac{1}{2}(\mathcal{L}_{GAN}(G_J, D^{loc}) + \mathcal{L}_{GAN}(G_J, D^{glo})).$$

The generation of disentangled haze-free image is regularized by the previous adversarial loss. For the disentangled transmission map, we introduce priors as regularization. Among the various known priors, we study the simple while effective choice: the smoothness of the medium transmission map (Tan 2008; Berman, Avidan, and others 2016). Mathematically, we use the traditional total variation of $\hat{t}$ as the regularization loss:

$$\mathcal{L}_{reg}(G_t) = TV(t) = \sum_{i,j} |t_{i+1,j} - t_{i,j}| + |t_{i,j+1} - t_{i,j}|. \quad (4)$$

Our final objective function is:

$$\begin{aligned}\mathcal{L}(G_J, G_t, G_A, D) &= \mathcal{L}_{adv}(G_J, D) \\ &+ \lambda\mathcal{L}_{recon}(G_J, G_t, G_A) + \gamma\mathcal{L}_{reg}(G_t).\end{aligned} \quad (5)$$

And we optimize the objective by

$$G_J^*, G_t^*, G_A^* = \arg\min_{G_J, G_t, G_A}\max_{D} \mathcal{L}(G_J, G_t, G_A, D). \quad (6)$$

Recovering the Haze-free Image

With the trained model, we can disentangled a hazy image into three corresponding components, and obtain two recovered scene radiances.

The first one is directly obtain from the output of the generator G_J, which we denote as $\hat{J}$. The second one, denoted as $\hat{J}^t$, can be obtained using the estimated transmission map $\hat{t}$ and atmosphere light $\hat{A}$ following the reformulation of Equation (1):

$$\hat{J}^t = \frac{I - \hat{A}}{\hat{t}} + \hat{A} \quad (7)$$

Following (Cai et al. 2016), we apply guided image filtering (He, Sun, and Tang 2013) on the estimated transmission map $\hat{t}$ during the recovery to introduce further smoothness shaper edge.

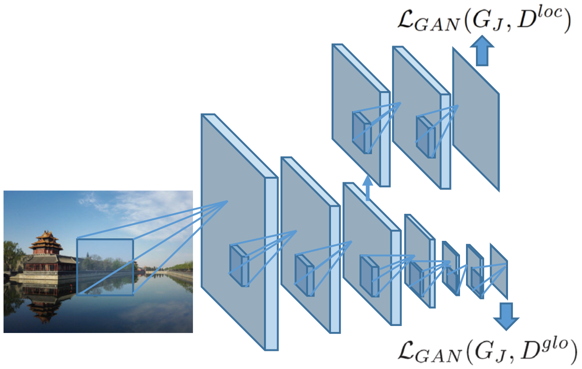
$\mathcal{L}_{GAN}(G_J, D^{loc})$
$\mathcal{L}_{GAN}(G_J, D^{glo})$

Figure 2: The architecture of the multi-scale discriminator.

Our disentangled generators can provide different viewpoints of the dehazing process. Specifically, the generator G_J tends to generate images with more texture details and perceptually more haze-clear because it is trained to learn the mapping to the haze-free images. However, the outputs of G_J can contain undesirable artifacts due to upsampling and unstable adversarial training, and are prone to affected by noise in the hazy regions. The $\hat{J}^t$ derived from the output of the generator G_t, on the other, are more smooth and visually pleasing because of the use of guided image filtering. But this can result in under-estimation of the haze level of the image.

With these two recovered scene radiances from different aspects, we generate our output haze-free image by blending the two recovered images

$$\hat{J}^{com} = \hat{J} \odot \hat{t} + \hat{J}^t \odot (1 - \hat{t}). \quad (8)$$

The blending retain more details for the regions with less haze, and ensure the smoothness within the regions with heavier haze. Different choices of recovery are analyzed in the ablation study. We report the results of the recovered $\hat{J}^{com}$ in all experiments, unless otherwise stated.

Network Architectures and Implementation Details

We adapt our generator and discriminator architectures from those in (Zhu et al. 2017). More details on network architectures and training procedures are presented in the appendix.

The generators G_I and G_t employ the same network architecture except using different numbers of filter channels. The generator G_A shares the same network with G_t until the upsampling layer. Specifically, the output of the last ResNet block in G_A is connected to a global max-pooling layer followed by a fully connected layer.

The multi-scale discriminator is implemented in a fully convolution fashion and introduces minor computational overhead compared with the original patch-level discriminator. As shown in Figure 2, we first design a K-layer discriminator that has a global receptive field size at the last layer. Then we extract the activation from one of the low-level layers (say, the k-th layer) and map it to an output. In our experiment, the local discriminator and global discriminator has a receptive field size of 70×70 and 256×256, respectively.

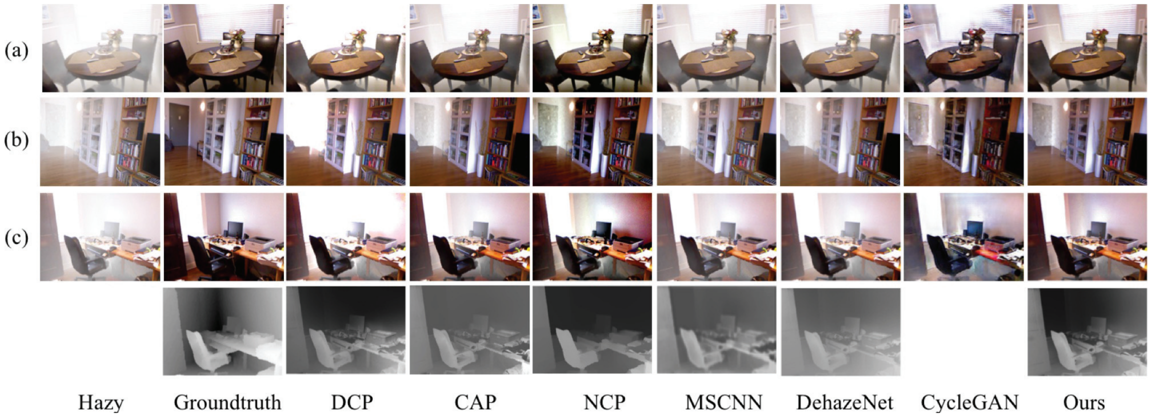

Figure 3: Qualitative results on NYU-Depth dataset (best viewed in color). CycleGAN cannot generate the corresponding transmission map and the result is left white.

Experiments

To verify the effectiveness of our *Disentangled Dehazing Network*, we perform extensive experiments on both synthetic and natural hazy image datasets. For synthetic dataset, we quantitatively evaluate our results using the ground-truth haze-free images. For natural hazy image dataset, we provide qualitative results to illustrate our superior performance on generating perceptually pleasing and haze-free images.

Synthetic Dataset

We use *D-HAZY* dataset (Ancuti, Ancuti, and De Vleeschouwer 2016), a public dataset built on the Middlebury (Scharstein et al. 2014) and NYU-Depth (Silberman et al. 2012) datasets. These two datasets provide images of various scenes and their corresponding depth maps, which can be used to create hazy images. Specifically, the hazy scenes in the dataset are synthesized using the aforementioned atmosphere scattering model (Eq. (1)). The medium transmission is estimated based on Eq. (2) using the depth map and the scattering coefficient $\beta = 1$. The global atmosphere light is set to be pure white. The synthesized hazy Middlebury and NYU-Depth datasets contain 23 and 1449 image pairs, respectively.

HazyCity Dataset

We collect a real image dataset, HazyCity dataset, for investigating the perceptual image dehazing problem in realistic scenes. Our dataset is built on PM25 dataset (Li, Huang, and Luo 2015), which is used for haze-level estimation in previous work. Images in our dataset are crawled from a tourist website and are photos of various attraction sites and street scenes in Beijing. We ask the annotators to label each image as “perceptually” hazy or not. In other word, an image is annotated as hazy if the haze can be visually perceived by human. Each image is annotated by 3 different annotators, and we only include the images of consistent labels from all annotators in our dataset. Finally, the HazyCity dataset contains 845 natural hazy images and 1891 haze-free images. Figure 4 illustrates some example images in our dataset.

The HazyCity dataset differs from all previous datasets for image dehazing studies in three aspects: 1) A large scale real image dataset with natural hazy and haze-free images. 2) Images are taken from outdoor scenes, which is more practical than previous datasets with mostly indoor scenes (D-HAZY dataset). 3) The dataset is much more challenging because natural hazy images can contain complex scenes, heterogeneous atmosphere, irregular illumination and heavy haze, etc. Our HazyCity dataset can be used as a new testbed for evaluating image dehazing algorithms in practical outdoor scenes. See Figure 4 for some examples of hazy images in the dataset.

Comparison Methods and Evaluation Metrics

We compare our proposed model with several state-of-the-art dehazing methods. The baseline methods can be separated in to two groups: prior-based and learning-based. For prior-based methods, we compare with the classical Dar-Channel Prior (**DCP**)(He, Sun, and Tang 2011), and the more recent Color Attenuation Prior (**CAP**) (Zhu, Mai, and Shao 2015) and Non-local Color Prior (**NCP**) (Berman, Avidan, and others 2016). For learning-based methods, we compare with two recent works **DehazeNet** (Cai et al. 2016) and **MSCNN** (Ren et al. 2016). We also compare with **CycleGAN**, a general framework for unpaired image-to-image translation using GAN. We test the baseline methods using their released codes and train the model with the recommended parameters.

These baselines are very strong and can obtain decent visual result for synthetic data. For fair and comprehensive comparison, we use three different quantitative evaluation

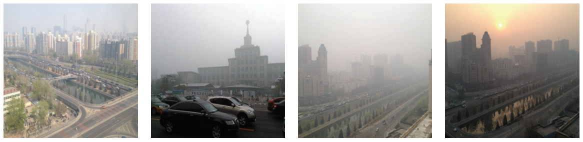

Figure 4: Example hazy images in our HazyCity dataset.

Metrics	DCP	CAP	NCP	MSCNN	DehazeNet	CycleGAN	Ours
PSNR	10.9803	12.7844	13.0461	12.2669	12.8426	13.3879	15.5456
SSIM	0.6458	0.7095	0.6678	0.7000	0.7175	0.5223	0.7726
CIEDE2000	18.9781	16.0327	16.1845	17.4497	15.8782	17.6113	11.8414

Table 1: Average PSNR, SSIM and CIEDE2000 results on NYU-Depth dataset. Numbers in blue indicate the second best results. Our approach consistently outperforms other methods by a large margin.

Metrics	DCP	CAP	NCP	MSCNN	DehazeNet	CycleGAN	Ours
PSNR	12.0234	14.1601	14.1827	13.5501	13.5959	11.3037	**14.9539**
SSIM	0.6902	0.7621	0.7123	0.7365	0.7502	0.3367	**0.7741**
CIEDE2000	18.0229	14.3317	15.6075	16.1304	15.4261	26.3181	**13.4826**

Table 2: Average PSNR, SSIM and CIEDE2000 results on the cross-domain Middlebury dataset.

metrics: PSNR, SSIM and CIEDE2000. PSNR provides a pixel-wise evaluation and is capable of indicating the effectiveness of haze removal. On the other hand, SSIM is consistent with human perception, and CIEDE2000 measures the color difference (smaller values indicate better color preservation).

Results on Synthetic Dataset

We train our model using the indoor images in NYU-Depth database of D-HAZY dataset. Although the dataset contains paired hazy and haze-free images, we do not use this information to train our approach and randomly shuffle to simulate the unpaired supervision. We perform standard data augmentation techniques including rescaling, random cropping and normalization (Zhu et al. 2017). Baseline methods are also tested on the resized and cropped images for fair comparison. To reduce the effect of randomness during training, we train our model and CycleGAN for five times and report the median results.

Table 1 shows the quantitative results of our model on NYU-Depth dataset. Our model outperforms the state-of-the-art methods consistently for all the three metrics. (The larger values of PSNR, SSIM and the smaller value of CIEDE2000 indicate the better dehazing and perceptual quality.) In particular, our model achieves a much higher PSNR score, indicating that we can generate higher quality images. The recent dehazing approaches (CAP, NCP, MSCNN, DehazeNet) achieve comparable performance with each other and outperform the classical DCP method, which is consistent with the observation in previous work (Li et al. 2017a). Although CycleGAN can get decent PSNR score, it performs relatively worse in terms of SSIM and CIEDE2000, which indicates limited visual quality.

We show the dehazing results on the NYU-Depth dataset in Figure 3. We observe that our model can generate better details than the baseline methods, especially in the regions with heavy haze. DCP often fails in bright and heavily hazy regions and produces unrealistically bright result since the dark-channel prior is violated. Although CAP, MSCNN and DehazeNet can generate visually decent outputs, they tend to under-estimate the haze level and the outputs still look fairly hazy. NCP, on the other hand, tends to produce over-saturated images with obvious color difference. CycleGAN fails to generate desired outputs with consistent content, even with the use of adversarial learning and backward mapping. This is due to the ambiguity of the mapping with the presence of haze. Our approach can alleviate this problem by physical-model based disentanglement and presents reliable restoration of haze-free images. Overall, our model can generate both haze-free and perceptually pleasing images.

We also show the estimated transmission maps of the example image. Without any direct guidance of the transmission map, our model can still generate decent results comparable with prior-based methods or supervised training methods. Moreover, our estimated transmission map achieves higher contrast than that of CAP, MSCNN and DehazeNet. This explains how we generate haze-clear images of higher quality. Although all of these approaches tend to predict low transmission value for white/bright region (indicating hazy region), our approach is more robust than that of DCP and NCP, and does not produce images with unrealistically bright regions and sever color difference.

To evaluate the generalization performance, we apply the model trained on NYU-depth on the cross-domain dataset, Middlebury. Middlebury contains quite different scenes compared with the NYU-Depth. Table 2 reports the quantitative results on the Middlebury dataset and the visual results are provided in the appendix. We observe that our model generalize well and obtain the best performance. Notably, our model generalize much better than CycleGAN. Our disentanglement mechanism can benefit the model generalization because it is trained to learn the underlying generation process of hazy images.

Ablation Study

Analysis of the Multi-scale discriminator We test the effectiveness of our proposed multi-scale discriminator by comparing with the single local discriminator (70 × 70) and global discriminator (256 × 256). Table 3 shows the quantitative results. Note that the evaluation is based on the output of generator G_J, i.e., the $\hat{J}$, because it is directly affected by the choice of discriminators.

The local discriminator tends to produce sharper results and therefore achieves higher SSIM scores. On the other

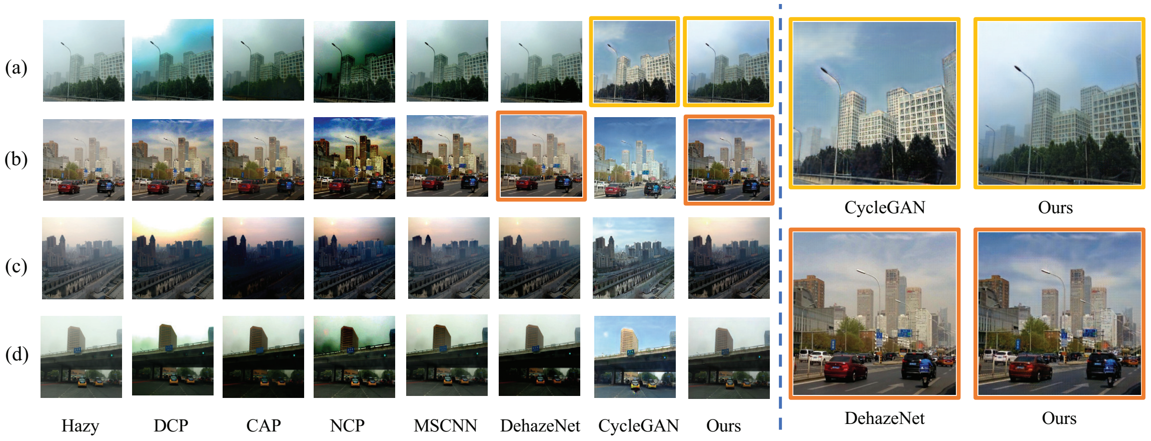

Figure 5: Qualitative results on HazyCity dataset (best viewed in color).

Metric	Local	Global	Multi-scale
PSNR	13.7518	*15.3304*	**15.6440**
SSIM	*0.6909*	0.6779	**0.7432**
CIEDE	15.1497	*13.7075*	**12.3797**

Table 3: Comparison between our proposed multi-scale discriminator and the local / global discriminator.

hand, the global discriminator can better estimate the haze level of the image and generate more haze-clear results, corresponding to higher PSNR scores. Our multi-scale discriminator combine the best of the two worlds and can achieve both sharp and haze-clear results (See Table 3).

Analysis of recovering haze-free images We show the effectiveness of blending two generated haze-free images $\hat{J}$ and $\hat{J}^t$ in Table 4.

Metric	$\hat{J}$	$\hat{J}^t$	$\hat{J}^{com}$
PSNR	*15.644*	15.1486	**15.5456**
SSIM	0.7432	*0.7548*	**0.7726**
CIEDE	12.3797	*12.2036*	**11.8414**

Table 4: Comparison of different methods for recovering the haze-free image.

Analysis of model learning We analyze how the presence of corresponding hazy / haze-free scenes in the training samples can effect the model training. We first randomly split the dataset into two halves (split 1 and 2). We use different combinations of the images for model training, and then test the model on the hazy images in split 2. The different settings and corresponding results are shown in Table 5.

Our results show that the best dehazing results can be obtain when both corresponding hazy and haze-free scenes are used for training, even though without paired supervision. However, other settings can still achieve comparable performance. This implies that it is not necessary to have paired hazy and haze-free scenes during training (see setting 2 and 3). And our model has decent generalization performance even when no corresponding scenes presented in training data (see setting 4).

Setting	Hazy	Haze-free	PSNR	SSIM	CIEDE
1	Split 2	Split 2	15.2269	0.7510	12.1974
2	Split 1	Split 2	14.8033	0.7246	13.0321
3	Split 2	Split 1	14.7089	0.7466	13.0126
4	Split 1	Split 1	14.7460	0.7368	12.9533

Table 5: Different settings for model training. All settings are tested on hazy images in split 2.

Results on HazyCity Dataset

We evaluate our *Disentangled Dehazing Network* on the real image dataset. Figure 5 shows the visual results of our model and the comparison with other dehazing algorithms.

These hazy examples, though very common in outdoor scenes, turn out to be very challenging for most of current dehazing algorithms. Most of the prior-based approaches fail to generate visually pleasing results because the priors and assumptions used in their algorithms are easily violated. The data-driven methods (MSCNN, DehazeNet and our model) tend to be more robust without the limitation of the haze-relevant priors or heuristic cues. Comparing with DehazeNet, our model can generate more haze-clear and vivid results (see the result in row (c)). Moreover, DehazeNet may misestimate the transmission on some haze-clear regions and generate undesirably dark results (see the results in row (a) and (d)). CycleGAN can generate haze-free style images, thanks to the power of adversarial training. However, it is not faithful to the original scene radiance and fails to retain content details. Although our model cannot generate fully haze-free images, it provides a practical approach to enhance the visibility of hazy images, using only unpaired data.

Conclusion

In this paper, we propose Disentangled Dehazing Network, a novel image dehazing model that learns to perform physical-model based disentanglement by adversarial training. The model can be trained using only unpaired supervision and is able to generate perceptually appealing dehazing results. Extensive experiments on both synthetic and real image datasets verify the effectiveness and generalization ability of our approach.

Although we focus on image dehazing in this paper, the proposed method can be generalized to many other applications where the layered image models (Wang and Adelson 1994; Yan et al. 2016) can be applied, such as image deraining and image matting. We intend to investigate more general applications of our disentangled network in the future.

References

Ancuti, C.; Ancuti, C. O.; and De Vleeschouwer, C. 2016. D-hazy: A dataset to evaluate quantitatively dehazing algorithms. In *Image Processing (ICIP), 2016 IEEE International Conference on*, 2226–2230. IEEE.

Berman, D.; Avidan, S.; et al. 2016. Non-local image dehazing. In *CVPR*.

Cai, B.; Xu, X.; Jia, K.; Qing, C.; and Tao, D. 2016. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing* 25(11):5187–5198.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 2172–2180.

Donahue, C.; Balsubramani, A.; McAuley, J.; and Lipton, Z. C. 2017. Semantically decomposing the latent spaces of generative adversarial networks. *NIPS Workshop on Learning Disentangled Representations (2017)*.

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

Fu, T.-C.; Liu, Y.-C.; Chiu, W.-C.; Wang, S.-D.; and Wang, Y.-C. F. 2017. Learning cross-domain disentangled deep representation with supervision from a single domain. *arXiv preprint arXiv:1705.01314*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS)*, 2672–2680.

He, K.; Sun, J.; and Tang, X. 2011. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(12):2341–2353.

He, K.; Sun, J.; and Tang, X. 2013. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(6):1397–1409.

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

Kim, T.; Cha, M.; Kim, H.; Lee, J.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. *ICML*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

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

Li, J.; Skinner, K. A.; Eustice, R. M.; and Johnson-Roberson, M. 2017b. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation Letters (RA-L)*.

Li, Y.; Huang, J.; and Luo, J. 2015. Using user generated online photos to estimate and monitor air pollution in major cities. In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, 79. ACM.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems 30 (NIPS)*. 700–708.

Mathieu, M. F.; Zhao, J. J.; Zhao, J.; Ramesh, A.; Sprechmann, P.; and LeCun, Y. 2016. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 5040–5048.

Narasimhan, S. G., and Nayar, S. K. 2000. Chromatic framework for vision in bad weather. In *CVPR*.

Narasimhan, S. G., and Nayar, S. K. 2002. Vision and the atmosphere. *International Journal of Computer Vision* 48(3):233–254.

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

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

Shu, Z.; Yumer, E.; Hadap, S.; Sunkavalli, K.; Shechtman, E.; and Samaras, D. 2017. Neural face editing with intrinsic image disentangling. In *CVPR*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. *ECCV* 746–760.

Tan, R. T. 2008. Visibility in bad weather from a single image. In *CVPR*.

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

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*.

Tung, H.-Y. F., and Fragkiadaki, K. 2016. Inverse problems in computer vision using adversarial imagination priors. In *ICCV*.

Tung, H.-Y. F.; Harley, A.; Seto, W.; and Fragkiadaki, K. 2017. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *ICCV*.

Wang, J. Y., and Adelson, E. H. 1994. Representing moving images with layers. *IEEE Transactions on Image Processing* 3(5):625–638.

Yan, X.; Yang, J.; Sohn, K.; and Lee, H. 2016. Attribute2image: Conditional image generation from visual attributes. In *ECCV*. Springer.

Yi, Z.; Zhang, H.; Gong, P. T.; et al. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

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