Self-augmented Unpaired Image Dehazing via Density and Depth Decomposition

Yang Yang[1], Chaoyue Wang[2], Risheng Liu[3], Lin Zhang[4], Xiaojie Guo[1,⋆], Dacheng Tao[2,5]

[1]Tianjin University, Tianjin, China [2]The University of Sydney, Sydney, Australia
[3]Dalian University of Technology, Liaoning, China
[4]Tongji University, Shanghai, China
[5]JD Explore Academy, Beijing, China

yangyangcic@tju.edu.cn, chaoyue.wang@outlook.com, rsliu@dlut.edu.cn
cslinzhang@tongji.edu.cn, {xj.max.guo, dacheng.tao}@gmail.com

Abstract

To overcome the overfitting issue of dehazing models trained on synthetic hazy-clean image pairs, many recent methods attempted to improve models' generalization ability by training on unpaired data. Most of them simply formulate dehazing and rehazing cycles, yet ignore the physical properties of the real-world hazy environment, i.e. the haze varies with density and depth. In this paper, we propose a self-augmented image dehazing framework, termed D^4 (Dehazing via Decomposing transmission map into Density and Depth) for haze generation and removal. Instead of merely estimating transmission maps or clean content, the proposed framework focuses on exploring scattering coefficient and depth information contained in hazy and clean images. With estimated scene depth, our method is capable of re-rendering hazy images with different thicknesses which further benefits the training of the dehazing network. It is worth noting that the whole training process needs only unpaired hazy and clean images, yet succeeded in recovering the scattering coefficient, depth map and clean content from a single hazy image. Comprehensive experiments demonstrate our method outperforms state-of-the-art unpaired dehazing methods with much fewer parameters and FLOPs. Our code is available at https://github.com/YaN9-Y/D4.

1. Introduction

Haze is a kind of natural phenomenon caused by the scattering effect of aerosol particles in the atmosphere. It can cause severe disclarity to visual content, which brings trouble to both human observers and computer vision systems. Dehazing methods aim to remove the haze and improve the visual quality of real-world hazy images, which can benefit computer vision tasks like image segmentation [4, 38], object detection [15, 34] on the hazy weather.

The degradation of haze effect can be formulated by Koschmieder's law [29, 30]:

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

where $\mathbf{I}(z)$ indicates the z-th pixel of observed hazy image, $\mathbf{J}(z)$ and $\mathbf{A}$ are the scene radiance and global atmosphere light, respectively. Transmission map $\mathbf{t}(z) = e^{-\beta \mathbf{d}(z)}$ is defined by the scene depth $\mathbf{d}(z)$ and the scattering coefficient β that reflects the haze density.

With the great learning capability of deep neural networks [12, 13], plenty of methods were proposed to solve the image restoration tasks [3, 25, 35, 46, 48, 49], as well as image dehazing [7, 27, 31, 36], in a supervised manner. Through training on a large amount of synthetic hazy-clean image pairs, supervised deep dehazing methods achieved impressive results on specific test sets.

However, there exists a relatively large domain gap between synthetic and real-world hazy images. Dehazing models that are solely trained on paired synthetic images are easy to over-fitting, and generalize poorly to real-world hazy conditions.

Since the desired real-world hazy&clean image pairs are nearly unreachable, in recent years, many unpaired deep learning methods were proposed to explore the dehazing cues from unpaired training data. Among them, constructing the dehazing cycle and rehazing cycle is widely adopted [8, 9, 17, 26, 45, 50] since it provides a simple and effective scheme for keeping content consistency while performing domain transformation. If the hazy and clean image domains can be accurately modeled, the cycle framework is expected to gain promising performance on unpaired dehaz-

[⋆] Corresponding author.

This work was supported by the National Natural Science Foundation of China under Grant no. 62072327, and TSTC under Grant no. 20JC-QNJC01510. Dr Chaoyue Wang is supported by ARC FL-170100117.

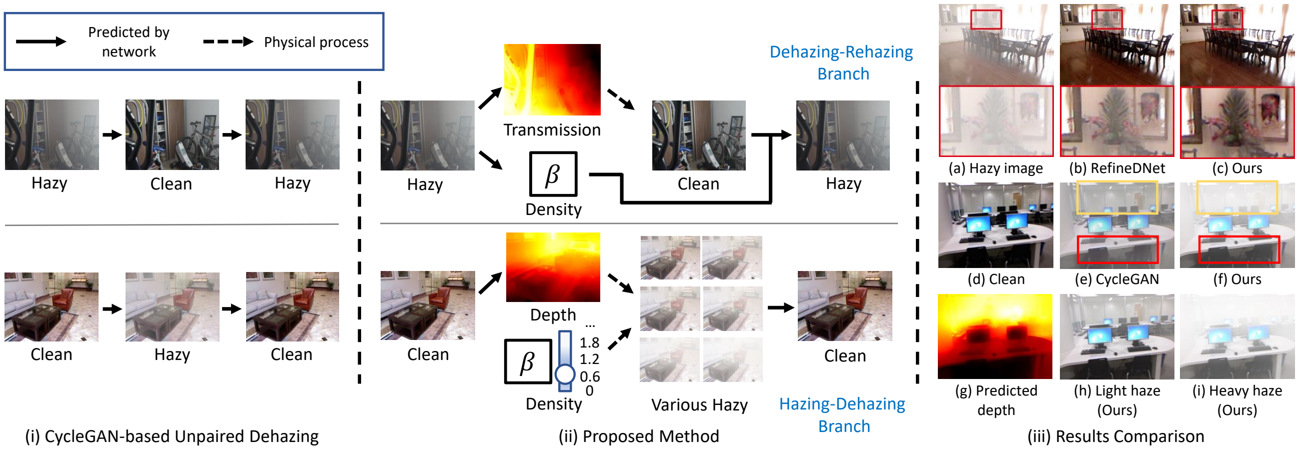


Figure 1. Illustration of (i) previous CycleGAN-based dehazing methods, (ii) our proposed method and, (iii) the results comparison. Compared to RefineDNet [51], our method can better remove the haze. The CycleGAN-based methods can only generate haze with fixed density for certain clean image, and the generated haze is not consistent with the depth.

ing. However, we argue that simply inheriting the CycleGAN [52] framework from unpaired image-to-image translation methods would fail to handle unpaired image dehazing tasks. Existing cycle-based dehazing methods ignore the physical properties of real-world hazy environments, *i.e*. the real-world haze varies with density and depth. As shown in Fig. 1 (i) and (iii)-(e), CycleGAN-based methods easily collapse to synthesizing haze with fixed density, and may incorrectly model the haze effect, *e.g*. the haze should be thicker accompanied with the increasing of scene depth.

In this paper, we propose a novel dehazing framework termed D^4, *i.e*. Dehazing via Decomposing transmission map into Density and Depth, for unpaired haze synthesis and removal. Following the hazy image formation process, we explicitly model the scattering coefficient β and depth map $\mathbf{d}(z)$ of the target scene. As shown in Fig. 1 (ii), on the Dehzing-Rehazing branch, our model is trained to directly estimate both transmission map and scattering coefficient from a hazy image. According to the physics process expressed in Eq. (1), scene depth and clean content then can be derived directly. On the Hazing-Dehazing branch, our model aims to estimate the depth information of the input clean image, then synthesize hazy images with different densities, *i.e*. scattering coefficients. Considering the fact that 'spatial-variant haze thickness provides an additional cue for perceiving scene depth', depth maps estimated from hazy images act as pseudo ground truth of the depth of clean images. Similarly, in the Hazing-Dehazing branch, the randomly sampled scattering coefficients β in the Hazing step act as pseudo ground-truth of the density predicted in the Dehazing step.

Finally, with our novel unpaired dehazing framework, we can (i) estimate depth maps from clean images (Fig. 1 (g)); (ii) synthesize realistic hazy images with various densities (data augmentation) (Fig. 1 (h),(i)); and (iii) achieve better dehazing performance than state-of-the-arts unpaired dehazing methods with less parameters and FLOPs.

Overall, our contributions can be summarized as follows:

- We propose a novel unpaired dehazing framework, which explicitly models the scattering coefficient (*i.e*. density) and the depth map of hazy scenes. The proposed physics-based framework largely alleviates the ill-posed problem that existed in existing unpaired dehazing methods.
- Inspired by the intuition: 'spatial-variant haze thickness reflect scene depth', our model learns to predict depth information from hazy images. Then, with only unpaired hazy and clean images, our model are trained to predict depth information from clean images.
- With estimated scene depth, our model is able to generate hazy images with different thickness by altering the scattering coefficient. Such characteristic acts as a self-argumentation strategy for better training the dehazing network.
- Extensive experiments on both synthetic and real images are conducted to reveal the effectiveness of our design. The proposed D^4 framework shows clear advantages on generalization ability over state-of-the-arts methods.

2. Related Work

This section briefly reviews previous dehazing works closely related to ours, which are grouped into prior-based, supervised and unsupervised learning-based methods.

Prior-based approaches. To allviate the ill-posedness of single image dehazing, early attempts concentrated on discovering priors on haze-free images from statistic analysis or observation. Among them, He *et al.* [14] proposed the dark channel prior, which assumes that the locally lowest intensity in RGB channels should be close to zero in haze-free natural images. Zhu *et al.* introduced the color attenuation prior [53], they point that the difference between the value and saturation of pixels should be positively correlated to the depth of scene in a linear model. In [10], Fattal discovered that the values of small patches distribute largely along a 1-D line in the RGB space. As a serious drawback, these methods are all built upon handcrafted priors, which are not always in line with the complex real environments.

Supervised learning approaches. Benfiting from the success of CNNs and the development of large-scale synthetic datasets, deep learning-based supervised methods relaxed the limitation of handcrafted priors, and have occupied the dominant position. For instance, [2] is the pioneer to estimate the transmission map from hazy images in an end-to-end manner by constructing a convolutional neural network. Ren *et al.* [36] proposed a multi-scale convolutional network to predict the transmission map from coarse to fine. However, these methods may suffer from the cumulative error when separately estimating the transmission and atmospheric light. To deal with this issue, Li *et al.* [22] reformulated Eq. (1) to simultaneously estimate the transmission map and atmospheric light together. Moreover, a number of methods [24, 27, 31, 32] directly estimate the clean image from hazy inputs without explicitly modeling the atmosphere scattering model. Recently, several attempts introduce the domain adaption methods into image dehazing tasks, they aim to shrink the domain gap between synthesized and real data [39, 41, 42]. Overall, although supervised methods have achieved remarkable performance on synthetic datasets, they are easy to overfit to the provided training data and generalizing poorly to other hazy images, specifically for real-world haze.

Unsupervised learning approaches. Comparing to the supervised methods, the unsupervised learning methods do not rely on paired supervision. Some unsupervised methods can be directly trained on hazy images. For instance, [20, 21] perform dehazing in a zero-shot manner, which disentangle the hazy image into the clean image and other components. [11] trains a network by minimizing a DCP-based [14] loss on hazy images. In these methods, since the clean images are not involved in training, the intrinsic property of the clean image domain is not efficiently considered, thus limit their performance.

The unpaired dehazing methods learn the dehazing mapping from unpaired clean and hazy images. Besides of a few methods [47, 51] that try to disentangle the clean component from the hazy image under GAN supervision, most unpaired dehazing methods are based on CycleGAN with other specific designs. For example, the CDNet [8] introduces the optical model into the CycleGAN. [50] adopts double-discriminator to stabilize the cyclic training. CycleDehaze [9] applies a Laplacian pyramid network to deal with high-resolution images and proposes a cycle perceptual loss for better structure preserving. [26] uses a two-stage mapping strategy in each branch of CycleGAN to enhance the effectiveness of fog removal. However, these methods usually ignore the depth information and the variousness of density when generating hazy images. The absence of these factors leads to unrealistic haze generation, which will further affect the dehazing performance. To deal with these issues, in our proposed D^4 framework, we focus on exploring both depth information and scattering coefficient contained in hazy and clean images.

3. Proposed Method

CycleGAN [52] is a broadly adopted framework for unpaired image-to-image translation. On one hand, the GAN loss is employed to enforce images translating between two domains. On the other hand, the cycle reconstruction loss works well to maintain the content consistency. For image dehazing, the CycleGAN-based methods [8, 9, 26] usually contain a dehazing network and a rehazing network, which predict the clean images and hazy images from their counterparts [9]. Here, we argue that such practice may be questionable. Two crucial properties, depth and density, are ignored in those methods. Consequently, the generated haze usually lacks of realism and variousness, which further affects the learning of dehazing network. To address these issues, we proposed a novel unpaired dehazing framework termed D^4 (Dehazing via Decomposing transmission map into Density and Depth). The overall framework and training procedures are detailed as follows.

3.1. The Overall Framework

Given a clean image set $\mathcal{X}_C = \{\mathbf{C}\}_{i=1}^{N_1}$ and a hazy image set $\mathcal{X}_H = \{\mathbf{H}\}_{i=1}^{N_2}$, where N_1 and N_2 stand for the cardinal numbers of the two sets. Unlike the synthetic hazy-clean image datasets, there exists no paired information between two sets. As shown in Fig. 2, our D^4 framework consits of three networks: the dehazing network $\mathcal{G}_D$, the depth estimation network $\mathcal{G}_E$, and the refine network $\mathcal{G}_R$.

The dehazing network $\mathcal{G}_D$ is trained to estimate the transmission map $\hat{\mathbf{t}}$ and the scattering coefficient $\hat{\beta}$ from a hazy image $\mathbf{H}$, which can be formulated as:

$$(\hat{\beta}, \hat{\mathbf{t}}) = \mathcal{G}_D(\mathbf{H}). \tag{2}$$

According to Eq. 1, the depth map $\mathbf{d}$ of the hazy image can be calculated from the estimated transmission $\hat{\mathbf{t}}$ and scattering coefficient $\hat{\beta}$ by:

$$\mathbf{d}(z) = \frac{\ln \hat{\mathbf{t}}(z)}{-\hat{\beta}}. \tag{3}$$

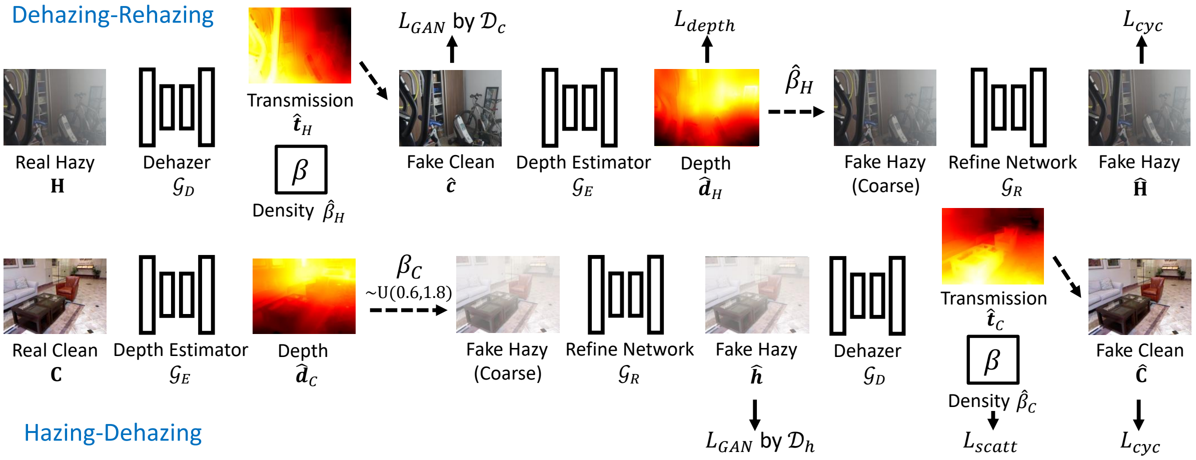


Figure 2. The architecture of our proposed network D^4. The whole network comprises of the Dehazing-Rehazing branch and the Hazing-Dehazing branch. Both depth and density information are considered in haze generation. Two pairs of pseudo supervision are introduced to confirm the accuracy of estimated depth and scattering coefficient.

The depth estimation network $\mathcal{G}_E$ aims to estimate the depth $\hat{\mathbf{d}}$ from a clean image $\mathbf{C}$, which is formulated as

$$\hat{\mathbf{d}} = \mathcal{G}_E(\mathbf{C}). \tag{4}$$

Note that our depth estimation network $\mathcal{G}_E$ shares the same function as other single image depth estimation networks [19,33], yet we do not employ any pretrained weights from existing depth estimators or ground truth depth supervision during training. In our D^4 framework, network $\mathcal{G}_E$ is trained using the pseudo supervision from the dehazing network $\mathcal{G}_D$, more details are introduced in Sec. 3.3.

The refine network $\mathcal{G}_R$**.** Different from previous CycleGAN-based methods that directly synthesize a hazy image from an input clean image, the proposed D^4 models the rehazing process by considering two physical properties (density and depth). Specifically, we first derive a coarse fake hazy image by combining a clean image, its estimated depth and a scattering coefficient. Then, the refine network $\mathcal{G}_R$ works as an image-to-image translation network that maps coarse fake hazy images to hazy images that follow the distribution of real hazy images, *i.e.*

$$\hat{\mathbf{H}} = \mathcal{G}_R(\hat{\mathbf{H}}_{coarse}). \tag{5}$$

In other words, the proposed refine network can be regarded as performing a conditional hazy image generation. Given the depth and density information, the refine network aims to generate visually realistic hazy images.

In our proposed D^4 framework, both the dehazing network $\mathcal{G}_D$ and the depth estimation network $\mathcal{G}_E$ are based on the structure of EfficientNet-lite3 [43], and the refine network $\mathcal{G}_R$ has a UNet [37] structure. The detailed network architectures are provided in our supplementary material.

3.2. Training Procedure

As shown in Fig. 2, the training of our D^4 contains two branches: (i) *the Dehazing-Rehazing branch* and (ii) *the Hazing-Dehazing branch*.

The Dehazing-Rehazing branch. By feeding a hazy image $\mathbf{H}$ into the dehazing network $\mathcal{G}_D$, we can obtain the estimated transmission map $\hat{\mathbf{t}}$, estimated scattering coefficient $\hat{\beta}_H$ and the calculated depth $\mathbf{d}_H$. Meanwhile, the dehazed result $\hat{\boldsymbol{c}}$ can be calculated by:

$$\hat{\boldsymbol{c}}(z) = \frac{\mathbf{H}(z) - \hat{A}}{\hat{\mathbf{t}}(z)} + \hat{A}, \tag{6}$$

where $\hat{A}$ is the atmospheric light estimated by the prior.

With the dehazed image $\hat{\boldsymbol{c}}$, the depth estimator $\mathcal{G}_E$ predicts the depth $\hat{\mathbf{d}}_H$ from it. Then, we rehaze the dehazed image $\hat{\boldsymbol{c}}$ with previously estimated scattering coefficient $\hat{\beta}_H$ and the estimated $\hat{\mathbf{d}}_H$. Specifically, we first derive a coarse hazy image $\hat{\mathbf{H}}_{coarse}$ with the haze formation process:

$$\hat{\mathbf{H}}_{coarse}(z) = \hat{\boldsymbol{c}}(z)e^{-\hat{\beta}_H\hat{\mathbf{d}}_H(z)} + A(\mathbf{1} - e^{-\hat{\beta}_H\hat{\mathbf{d}}_H(z)}), \tag{7}$$

where we adopt the brightest pixel as the atmospheric light A for haze generation. Then, the coarse hazy image $\hat{\mathbf{H}}_{coarse}$ is processed by the refine network to obtain the final rehazing image $\hat{\mathbf{H}} = \mathcal{G}_R(\hat{\mathbf{H}}_{coarse})$.

The Hazing-Dehazing branch. In this branch, we sample a clean image $\mathbf{C}$ from the set $\mathcal{X}_C$. The depth estimation network $\mathcal{G}_E$ is employed to estimate the depth map $\hat{\mathbf{d}}_C$ from the image $\mathbf{C}$. Then, we randomly sample a β_C from a predefined uniform distribution. Following the same physical process presented in Eq. 7, we derive the coarse hazy image $\hat{\boldsymbol{h}}_{coarse}$ with variable haze density, *i.e.*

$$\hat{\boldsymbol{h}}_{coarse}(z) = \mathbf{C}(z)e^{-\beta_C\hat{\mathbf{d}}_C(z)} + A(\mathbf{1} - e^{-\beta_C\hat{\mathbf{d}}_C(z)}). \tag{8}$$

Our fake hazy image $\hat{\boldsymbol{h}}$ is then synthesized by the refine network $\mathcal{G}_R$, and is further processed by the dehazed network $\mathcal{G}_D$ to predict the transmission $\hat{\mathbf{t}}_C$, the scattering coefficient $\hat{\beta_C}$. Finally, we can reconstruct the clean input using the same calculations presented in Eq. 6. It is worth noting that, in this branch, since β_C is sampled from a predefined range, our hazing process can be regarded as a data augmentation operation for the following training of the dehazing network.

3.3. Training Objectives

In the proposed D^4 framework, we train the proposed three networks together to perform the dehazing and rehazing cycles. Similar to CycleGANs, the cycle-consistency loss and the adversarial training loss are employed to penalize the content consistency and data distribution, respectively. Differently, we propose novel pseudo scattering coefficient supervision loss and pseudo depth supervision loss for learning physical properties (density and depth) from unpaired hazy and clean images.

Cycle-consistency loss imposes that an intermediate image transferred from one domain to another should be able to transfer back. In our D^4 framework, the reconstructed clean image $\hat{\mathbf{C}}$ and hazy image $\hat{\mathbf{H}}$ should be consistent with their input counterparts $\mathbf{C}$ and $\mathbf{H}$, respectively. The cycle-consistency loss in our D^4 can be written as follows:

$$L_{cyc} = \mathbb{E}_{\mathbf{C}\sim\mathcal{X}_C}\|\mathbf{C} - \hat{\mathbf{C}}\|_1 + \mathbb{E}_{\mathbf{H}\sim\mathcal{X}_H}\|\mathbf{H} - \hat{\mathbf{H}}\|_1, \quad (9)$$

where $\|\cdot\|_1$ designates the ℓ_1 norm.

Adversarial learning losses evaluate whether a generated image belongs to a specific domain. In other words, it penalizes our dehazed and rehazed images should be visually realistic and following the same distribution as images in training sets $\mathcal{X}_H$ and $\mathcal{X}_C$. We adopt the LSGAN [28] due to its promising stability and visual quality. For the dehazing network $\mathcal{G}_D$ and corresponding discriminator $\mathcal{D}_c$, the adversarial loss can be expressed as follows:

$$\begin{aligned} L_{adv}(\mathcal{D}_c) &= \mathbb{E}[(\mathcal{D}_c(\boldsymbol{c}) - 1)^2] + \mathbb{E}[(\mathcal{D}_c(\hat{\boldsymbol{c}}))^2], \\ L_{adv}(\mathcal{G}_D) &= \mathbb{E}[(\mathcal{D}_c(\hat{\boldsymbol{c}}) - 1)^2], \end{aligned} \quad (10)$$

where $\boldsymbol{c}$ is real clean sample from the clean image set $\mathcal{X}_C$, $\hat{\boldsymbol{c}}$ is the dehazing result from $\mathcal{G}_D$, $\mathcal{D}_c$ is the discriminator judging if the input image belongs to the clean domain. The adversarial loss for the haze refine network $\mathcal{G}_R$ and corresponding discriminator $\mathcal{D}_H$ is in the same form.

Pseudo scattering coefficient supervision loss penalizes the difference between β_C (the randomly sampled scattering coefficient for haze generation in the Hazing-Dehazing branch) and $\hat{\beta}_C$ (the scattering coefficient estimated from the generated hazy image $\hat{\boldsymbol{h}}$),

$$L_{scatt} = (\hat{\beta}_C - \beta_C)^2. \quad (11)$$

For hazy images from the training set $\mathcal{X}_H$, the ground truth scattering coefficient is unavailable. Therefore, we alternatively adopt the randomly sampled scattering coefficients and corresponding generated hazy images to train the proposed dehazing network.

Pseudo depth supervision loss. According to the observation that 'spatial-variant haze thickness provides an additional cue for perceiving scene depth', we employ the depth map $\mathbf{d}_H$ predicted from the hazy image $\mathbf{H}$ as the pseudo ground truth. Then, we train the depth estimation network $\mathcal{G}_E$ to estimate the depth map $\hat{\mathbf{d}}_H$ from the dehazed image $\hat{\boldsymbol{c}}$, *i.e.* $\hat{\mathbf{d}}_H = \mathcal{G}_E(\hat{\boldsymbol{c}})$. Then, we define the training loss,

$$L_{depth} = \|\hat{\mathbf{d}}_H - \mathbf{d}_H\|_1. \quad (12)$$

Overall, the depth estimation network $\mathcal{G}_E$ is optimized alone by the depth loss L_{depth}. The rest modules are jointly optimized with a weighted combination of the cycle loss, adversarial loss and pseudo scattering coefficient loss as:

$$L_{total} = \lambda_{cyc}L_{cyc} + \lambda_{adv}L_{adv} + \lambda_{scatt}L_{scatt}, \quad (13)$$

where λ_{cyc}, λ_{adv} and λ_{scatt} are the weights balancing different terms. In our experiments, empirically setting $\lambda_{cyc} = 1$, $\lambda_{adv} = 0.2$ and $\lambda_{scatt} = 1$ works well.

4. Experiments

4.1. Experimental Configuration

Datasets. In this work, we adopt the RESIDE [23] dataset, I-HAZE [1] dataset and Fattal's dataset [10] to train and evaluate the performance of our model and other candidates.

The RESIDE dataset [23] is a widely-used large-scale dehazing benchmark datasets which comprises of subsets including: (i) ITS/OTS, which contains 13990/313950 synthetic indoor/outdoor hazy images with ground-truth for training. (ii) SOTS-indoor/outdoor, which include 500 synthetic indoor/outdoor hazy images with ground-truth for testing. (iii) RTTS and URHI, which both contain over 4000 real hazy images without ground-truth clean images. The I-HAZE [1] dataset contains 35 image pairs of hazy and corresponding haze-free indoor images. The haze in this dataset is produced by professional haze generators. Fattal's dataset [10] includes 31 real hazy images without ground truth, and it is broadly used for visual comparison.

Competitors & Metrics. We compare our method with several state-of-the-arts dehazing algorithms. Among those methods, some use paired data for training, including EPDN [32], HardGAN [6], FFANet [31], DADehaze [39], and PSD [5]. While other methods are trained without using paired data, including DCP [14], CycleGAN [52], CycleDehaze [9], DisentGAN [47], YOLY [20] and RefineDNet [51]. The metrics including PSNR, SSIM [44] and CIEDE2000 (CIEDE for short) [40] are adopted to quantitatively evaluate the performance.

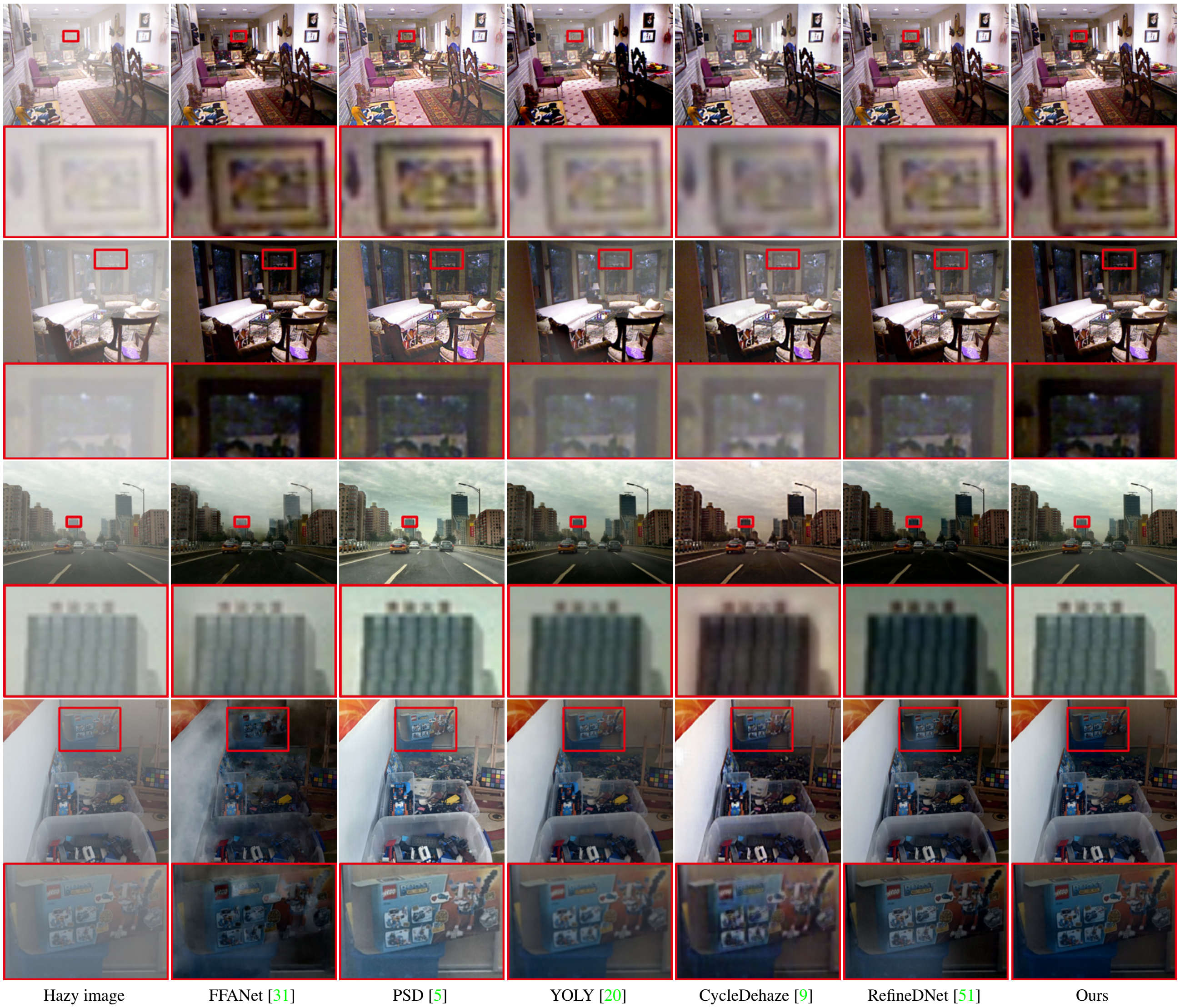


Figure 3. Visual comparison in haze removal on samples from the SOTS-indoor, SOTS-outdoor and I-HAZE dataset. FFANet only performs well on the first and second cases. All results of PSD, YOLY and CycleDehaze are hazy. The first, second and fourth cases of RefineDNet are hazy and the third case is over-dehazed. Our method dehazes well on all cases.

Implementation details. In the training phase, we apply the discriminator proposed in [16] with the patch size of 30×30. The Adam optimizer [18] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate $l_r = 10^{-4}$, and a batch size of 2 is used to optimize the network. All training samples are resized to 256×256, half of which are horizontally flipped for data augmentation. To estimate the atmospheric light, we adopt the method in [14] for outdoor and I-HAZE dataset. While for synthetic indoor scenes, the brightest pixel is regarded as atmospheric light.

4.2. Performance Evaluation

Comparisons on the benchmark dataset. To quantitatively evaluate the performance and the generalization ability of our D^4, we first conduct experiments on the benchmark datasets. Specifically, we train our model and comparison methods on the ITS dataset. For fair comparisons, all supervised methods utilized pair training data within the ITS dataset, yet unpaired methods, including our D^4, abandoned all paired information during training. Considering the SOTS-indoor is synthesised in the same way as ITS, we selected it as one of our test sets. While SOTS-outdoor and I-HAZE various on the scenes and haze types, the results on these two test sets can reflect the dehazing generalization ability of different models. Note that, since the I-HAZE dataset is not used for training, we adopt the entire I-HAZE dataset for testing.

The quantitative comparisons are reported in Tab. 1. On SOTS-indoor test set, the supervised methods HardGAN [6]

Table 1. The quantitative comparison results on SOTS-indoor, SOTS-outdoor and I-HAZE datasets. Higher PSNR, SSIM, and lower CIEDE2000, Params, FLOPs indicate better results. Paired and W/o Paired denote the method used or did not use paired data for training. Since DCP is not based on DNN and YOLY needs to iteratively process on single image, the parameters of DCP and the FLOPS of YOLY, DCP are not available. The best, second best and third best result are denoted in **bold**, underlined and *italic*, respectively.

	Methods	SOTS-Indoor			SOTS-Outdoor			I-HAZE			Efficiency	
		PSNR↑	SSIM↑	CIEDE↓	PSNR↑	SSIM↑	CIEDE↓	PSNR↑	SSIM↑	CIEDE↓	Params (M)	FLOPs (GMac)
Paired	EPDN [32]	25.06	0.931	4.578	20.47	0.896	8.566	15.02	*0.763*	*14.96*	17.38	4.826
	FFANet [31]	**36.36**	**0.993**	**1.108**	20.23	0.905	6.908	12.00	0.592	20.33	4.456	288.3
	HardGAN [6]	35.45	0.991	1.269	23.33	0.937	*5.348*	13.82	0.720	17.54	**2.546**	49.18
	PSD [5]	15.02	0.764	13.80	15.63	0.834	12.60	15.30	**0.800**	14.84	33.11	182.5
W/o Paired	DCP [14]	13.10	0.699	7.404	20.15	*0.919*	7.613	13.10	0.699	19.04	-	-
	CycleGAN [52]	21.34	0.898	7.000	20.55	0.856	9.298	*15.29*	0.756	19.50	11.38	56.89
	CycleDehaze [9]	20.11	0.854	8.761	*21.31*	0.899	9.481	14.69	0.751	19.05	11.38	*49.16*
	DisentGAN [47]	21.51	0.899	7.294	18.45	0.831	12.24	14.48	0.675	16.52	11.48	57.46
	YOLY [20]	15.84	0.819	12.37	14.75	0.857	15.85	14.74	0.688	15.24	32.00	-
	RefineDNet [51]	24.36	*0.939*	4.305	19.84	0.853	8.481	13.60	0.660	17.08	65.80	75.41
	Ours	*25.42*	0.932	*3.670*	**25.83**	**0.956**	**4.295**	**15.61**	0.780	**14.45**	*10.70*	**2.246**

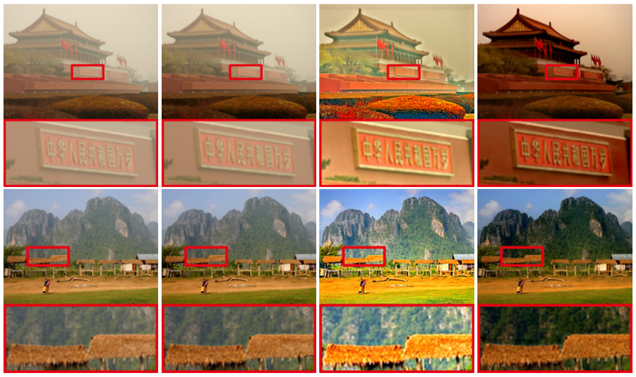


Figure 4. Visual comparison on two real images from Fattal's dataset [10]. FFANet has minor effect on haze removal. PSD causes global color distortion and over-saturation.

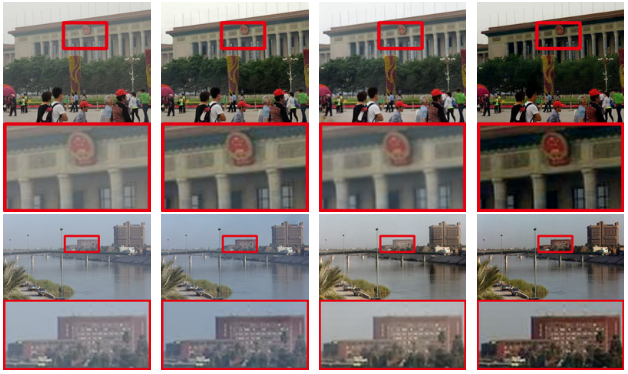
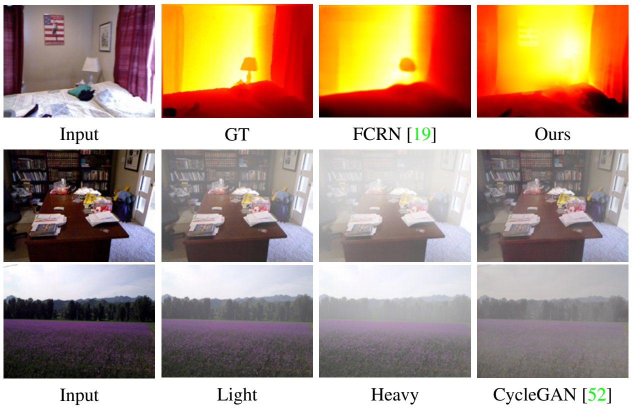


Figure 5. Comparison on two real samples from URHI. HardGAN and DADehaze both leave observable haze on the results.

and FFANet [31] demonstrate their powerful fitting ability and rank the first and second places with absolute advantage. Our D^4 is the best among the methods without using paired training data. In contrast, since the samples in SOTS-outdoor dataset and I-HAZE dataset are inconsistent with training data, the FFANet and HardGAN lose their dominant positions. It partly reveals the overfitting issue of supervised dehazing methods. While our D^4 gets rid of such defect and outperforms other competitors on both datasets. Besides, we also provide several visual comparisons in Fig. 3. For the first and second synthetic indoor cases, both of FFANet and our D^4 can thoroughly remove the haze while other methods remain observable haze on the results. For the last two cases, FFANet remains dense haze on the results. Although PSD produces brighter results, it fails to remove the haze. CycleDehaze suffers from the color distortion and RefineDNet over-dehazes the third case and leaves obvious haze on the last case, while our D^4 successfully removes the haze without leaving obvious artifacts. All these results validate the better generalization ability of our D^4. To summarize, although our result is not comparable with the supervised method in SOTS-indoor, it is outstanding among the methods without using paired data. Considering quantitative, visual results and model efficiency in all three datasets, our method is more appealing.

Comparisons on real-world hazy images. To further evaluate the dehazing performance for real scenes, we conduct experiments on the Fattal's dataset [10] and URHI dataset. We finetune our model on unpaired outdoor clean and hazy images from OTS and RTTS. From indoor to outdoor, we employ an additional hyper-parameter for adjusting transmission estimator when extending to outdoor scene. For DADehaze [39], RefineDNet [51] and PSD [5], we use their released model which is pre-trained on real hazy images. For FFANet [31] and HardGAN [6], we adopt the model trained on paired synthetic images. The visual results on Fattal's dataset and URHI dataset are shown in Fig. 4 and Fig. 5, respectively. From Fig. 4, we can see that FFANet [31] has a very limited effect towards real hazy images. The results produced by PSD [5] suffer severe color distortion. While our method successfully removes the haze in the image. In Fig. 5, results of both HardGAN [6] and DADehaze [39] contain residual haze, while our result is sharp



Input GT FCRN [19] Ours

Input Light Heavy CycleGAN [52]

Figure 6. Example of depth estimation and haze generation. Compared with GT and FCRN, our estimated depth is also visually reasonable. The haze generated by CycleGAN is not consistent with depth, while the haze from ours is more realistic and various.

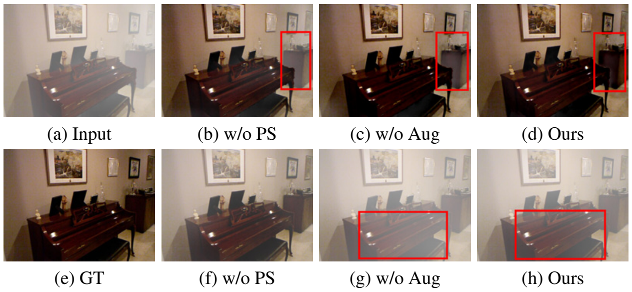
and clear. This part of experiment validates that our method generalizes well on real-world outdoor hazy images,

Results on depth estimation and haze generation. We notice that without paired depth supervision, our method can learn to predict visually reasonable depth maps from unpaired hazy and clean images. Fig. 6 presents a few cases of depth estimation and haze generation, from which we can see that, although a gap between our depth estimation and the supervised FCRN exists, our method is acceptable without ground-truth depth information. Having the predicted depth, our network can produce more realistic and various hazy images on both indoor and outdoor images in comparison with CycleGAN.

The comparison on the efficiency of the method. To measure the efficiency of the networks, we make comparison on the number of parameters, and the FLOPs of the dehazing models. Specifically, only the dehazing part of each model is taken into account. As shown in Tab. 1 , our model has the lowest FLOPs and fewer parameters with a light-weight backbone [43]. With such a light-weight model, our method still outperforms other state-of-the-arts methods, which validates the effectiveness of our design.

4.3. Ablation Study

In this part, we validate the effectiveness of our proposed self-augmentation and pseudo supervision mechanism. For the self-augmentation, we replace the random generated β in the Hazing-Dehazing branch with a fixed value, and keep other settings unchanged (denoted as w/o Aug). For the pseudo supervision, since the pseudo β supervision and pseudo depth supervision are closely tied to each other, we have to add or remove them together. Without these two pseudo supervisions (w/o PS), we remove the function of estimate β of the dehazing network $\mathcal{G}_D$ and change the depth estimation network $\mathcal{G}_E$ to directly estimating the transmission map. Overall, this setting makes our framework degrade to a vanilla-CycleGAN-liked architecture. The quantitative and qualitative results are shown in Tab. 2 and Fig. 7, respectively. From Tab. 2, we can see that any ablation to our network causes an obvious drop in performance. Fig. 7 shows intuitive visual comparisons. From the images on the first row, we can infer that our complete D^4 removes the haze most thoroughly in comparison with the networks without self-augmentation and pseudo supervision mechanism. Besides, as shown in the images on the second row, our complete D^4 generates the most realistic hazy images. Specifically, the red boxed regions are near to the observer, so the haze at that area should be thinner. Such a property is more clearly reflected on the images produced by our complete model than those by the other configurations. The ablation study verifies that both self-augmentation and pseudo supervision are effective.



(a) Input (b) w/o PS (c) w/o Aug (d) Ours

(e) GT (f) w/o PS (g) w/o Aug (h) Ours

Figure 7. The images in (b)-(d) are the dehazed results of (a) by different network configurations, while (f)-(h) are generated hazy images from the clean image (e).

Table 2. Ablation study on the SOTS dataset.

Metrics	SOTS-indoor			SOTS-outdoor		
	PSNR↑	SSIM↑	CIEDE↓	PSNR↑	SSIM↑	CIEDE↓
w/o Aug	23.15	0.9126	4.975	24.08	0.9472	5.041
w/o PS	20.02	0.8509	7.315	22.95	0.9168	5.669
Ours	**25.42**	**0.9321**	**3.670**	**25.83**	**0.9559**	**4.295**

5. Conclusion

This paper has proposed a self-augmented unpaired image dehazing framework termed D^4, which decomposes the estimation of transmission map into predicting the density and the depth map. With the estimated depth, our method is capable of re-rendering hazy images with various densities as self-augmentation to improve the dehazing performance by a large margin. Extensive experiments have validated the clear merits of our method over other state-of-the-arts dehazing methods. However, *our method also has the limitation that*, it usually over-estimates the transmission of extreme bright area, which will mislead the depth estimation network to predict low depth value for over-bright areas. Besides, we found that training data with low quality will make the training unstable. But it is positive that our thought of further decomposing variables in the physical model can be extended to other tasks, like low-light enhancement. We hope our method can innovate future works, especially for unpaired learning in low-level vision tasks.

References

[1] Cosmin Ancuti, Codruta O Ancuti, Radu Timofte, and Christophe De Vleeschouwer. I-haze: a dehazing benchmark with real hazy and haze-free indoor images. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 620–631. Springer, 2018. 5

[2] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE TIP*, 25(11):5187–5198, 2016. 3

[3] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. In *ECCV*, pages 171–187. Springer, 2020. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 1

[5] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *CVPR*, pages 7180–7189, June 2021. 5, 6, 7

[6] Qili Deng, Ziling Huang, Chung-Chi Tsai, and Chia-Wen Lin. Hardgan: A haze-aware representation distillation gan for single image dehazing. In *ECCV*, pages 722–738. Springer, 2020. 5, 6, 7

[7] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2157–2167, 2020. 1

[8] Akshay Dudhane and Subrahmanyam Murala. Cdnet: Single image de-hazing using unpaired adversarial training. In *WACV*, pages 1147–1155, 2019. 1, 3

[9] Deniz Engin, Anil Genç, and Hazim Kemal Ekenel. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *CVPRW*, pages 825–833, 2018. 1, 3, 5, 6, 7

[10] Raanan Fattal. Dehazing using color-lines. *ACM TOG*, 34(1):1–14, 2014. 3, 5, 7

[11] Alona Golts, Daniel Freedman, and Michael Elad. Unsupervised single image dehazing using dark channel prior loss. *IEEE TIP*, 29:2692–2701, 2019. 3

[12] Fengxiang He, Tongliang Liu, and Dacheng Tao. Why resnet works? residuals generalize. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5349–5362, 2020. 1

[13] Fengxiang He and Dacheng Tao. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*, 2020. 1

[14] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE TPAMI*, 33(12):2341–2353, 2010. 3, 5, 6, 7

[15] Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. Dsnet: Joint semantic learning for object detection in inclement weather conditions. *IEEE TPAMI*, pages 1–1, 2020. 1

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017. 6

[17] Yizhou Jin, Guangshuai Gao, Qingjie Liu, and Yunhong Wang. Unsupervised conditional disentangle network for image dehazing. In *ICIP*, pages 963–967. IEEE, 2020. 1

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248, 2016. 4, 8

[20] Boyun Li, Yuanbiao Gou, Shuhang Gu, Jerry Zitao Liu, Joey Tianyi Zhou, and Xi Peng. You only look yourself: Unsupervised and untrained single image dehazing neural network. *IJCV*, 129(5):1754–1767, 2021. 3, 5, 6, 7

[21] Boyun Li, Yuanbiao Gou, Jerry Zitao Liu, Hongyuan Zhu, Joey Tianyi Zhou, and Xi Peng. Zero-shot image dehazing. *IEEE TIP*, 29:8457–8466, 2020. 3

[22] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, pages 4770–4778, 2017. 3

[23] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE TIP*, 28(1):492–505, 2019. 5

[24] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang. Single image dehazing via conditional generative adversarial network. In *CVPR*, pages 8202–8211, 2018. 3

[25] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, pages 254–269, 2018. 1

[26] Wei Liu, Xianxu Hou, Jiang Duan, and Guoping Qiu. End-to-end single image fog removal using enhanced cycle consistent adversarial networks. *IEEE TIP*, 29:7819–7833, 2020. 1, 3

[27] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, pages 7314–7323, 2019. 1, 3

[28] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 5

[29] Srinivasa G Narasimhan and Shree K Nayar. Chromatic framework for vision in bad weather. In *CVPR*, volume 1, pages 598–605, 2000. 1

[30] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. *IJCV*, 48(3):233–254, 2002. 1

[31] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, volume 34, pages 11908–11915, 2020. 1, 3, 5, 6, 7

[32] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *CVPR*, pages 8160–8168, 2019. 3, 5, 7

[33] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 4

[34] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[35] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, pages 3937–3946, 2019. 1

[36] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, pages 154–169, 2016. 1, 3

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 4

[38] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*, pages 707–724, 2018. 1

[39] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *CVPR*, pages 2808–2817, 2020. 3, 5, 7

[40] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005. 5

[41] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Towards domain invariant single image dehazing. In *AAAI*, volume 35, pages 9657–9665, 2021. 3

[42] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Towards domain invariant single image dehazing. *arXiv preprint arXiv:2101.10449*, 2021. 3

[43] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 4, 8

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[45] Pan Wei, Xin Wang, Lei Wang, and Ji Xiang. Sidgan: Single image dehazing without paired supervision. In *ICPR*, pages 2958–2965. IEEE, 2021. 1

[46] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 1

[47] Xitong Yang, Zheng Xu, and Jiebo Luo. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *AAAI*, volume 32, pages 7485–7492, 2018. 3, 5, 7

[48] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 1

[49] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 1

[50] Jingming Zhao, Juan Zhang, Zhi Li, Jenq-Neng Hwang, Yongbin Gao, Zhijun Fang, Xiaoyan Jiang, and Bo Huang. Dd-cyclegan: Unpaired image dehazing via double-discriminator cycle-consistent generative adversarial network. *Engineering Applications of Artificial Intelligence*, 82:263–271, 2019. 1, 3

[51] Shiyu Zhao, Lin Zhang, Ying Shen, and Yicong Zhou. Refinednet: A weakly supervised refinement framework for single image dehazing. *IEEE TIP*, 30:3391–3404, 2021. 2, 3, 5, 6, 7

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2, 3, 5, 7, 8

[53] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE TIP*, 24(11):3522–3533, 2015. 3