Ensemble Dehazing Networks for Non-homogeneous Haze

Mingzhao Yu, Venkateswararao Cherukuri, Tiantong Guo, Vishal Monga
The Pennsylvania State University, The Department of Electrical Engineering, University Park, PA, USA
ethanyu@psu.edu, vmc5164@psu.edu,tiantong@ieee.org, vmonga@engr.psu.edu

Abstract

Image dehazing is one of the most challenging imaging inverse problems. Although deep learning methods produce compelling results, one of the most crucial practical challenge is that of non-homogeneous haze, which remains an open problem. To address this challenge, we propose 3 models that are inspired by ensemble techniques. First, we propose a DenseNet based single-encoder four-decoders structure denoted as EDN-3J, wherein among the four decoders, three of them output estimates of dehazed images ($\mathbf{J}_1$*,* $\mathbf{J}_2$*,* $\mathbf{J}_3$*) that are then weighted and combined via weight maps learned by the fourth decoder. In our second model called EDN-AT, the single-encoder four-decoders structure is maintained while three decoders are transformed to jointly estimate two physical inverse haze models that share a common transmission map* $\mathbf{t}$ *with two distinct ambient light maps (*$\mathbf{A}_1, \mathbf{A}_2$*). The two inverse haze models are then weighted and combined for the final dehazed image. To endow two sub-models flexibility and to induce capability of modeling non-homogeneous haze, we apply attention masks to ambient lights. Both the weight maps and attention maps are generated from the fourth decoder. Finally, in contrast to the above two ensemble models, we propose an encoder-decoder-U-net structure called EDN-EDU, which is a sequential hierarchical ensemble of two different dehazing networks with different modeling capacities. Experiments performed on challenging benchmark image datasets of NTIRE'20 and NTIRE'19 demonstrate that the proposed models outperform many state-of-the-art methods and this fact is particularly demonstrated in the NTIRE-2020 contest where the EDN-AT model achieves the best result in the sense of the perceptual quality metric LPIPS.*

1. Introduction

Dehazing is an important image processing task, which aims at recovering the scene information from images that are corrupted by dust, mist, smoke and other atmospheric particles that cause deflection of light from the objects. Due to the presence of haze, the visual quality of images is degraded drastically and the scene information will be lost. With the scene information being effected, crucial computer vision tasks such as object detection and recognition [1] that are critical to many emerging real-world applications such as autonomous driving, navigation systems, will be severely impacted. Hence, the demand for robust dehazing algorithms has been boosted in recent years.

Over the past decades, image dehazing has been an active research field [2, 3, 4, 5, 6, 4, 7, 8, 9] where the majority of work can be categorized into two classes: multi-image dehazing and single image dehazing. Constrained by the expressive capacity of models, many of early research work focused on adopting different kinds of fusion techniques to combine information from multiple images [10, 11]. However, in many scenarios, multiple images of the same scene under various environmental conditions are not available. Hence, single image dehazing has gradually become the more desirable option.

Most of the single image dehazing algorithms are governed by the physical haze model [12] shown below:

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

where $\cdot$ represents element-wise multiplication, $\mathbf{I}$ is the hazy image, $\mathbf{A}$ is the ambient light intensity and $\mathbf{t}$ is transmission map. $\mathbf{t}$ indicates a fraction of radiance of true scene $\mathbf{J}$ transmitted to the camera sensor. Thus, it is physically constrained within [0,1]. Estimation from the inverse of the haze model has shown advantages since it reflects the physical formation of the haze scene. The main challenge to single image dehazing is the great demand for modeling capacity since it is a heavily ill-posed inverse problem. Deep Learning (DL) techniques owing to their rich modeling capacity – have shown compelling performance across a wide array of imaging and vision problems such as image super-resolution [13, 14, 15], deblurring [16] and inpainting [17]. However, a downside of its tremendous increased flexibility is that its parameters are sensitive to the specifics of training data since they are mostly trained via stochastic training algorithms like SGD [18]. Hence, adequate amounts of training pairs $\{\mathbf{I}, \mathbf{J}\}$ are required to ensure the training dataset can represent the true distribution of general data. Otherwise, the learned models exhibit high variance and may not be able to generalize well.

Furthermore, in many practical scenarios, the haze is not uniform across a given image, thereby induces more challenges to the problem. In such cases, a single haze model might not be capable enough for modeling different haze densities in an image. These challenges combined with the already existing ill-posed nature of the problem may further hinder the performance of the existing deep-learning models by increasing the variance of the model. The NTIRE-2020 Dehazing Challenge [19] aims to tackle these practical issues by providing a very challenging dataset that has images with varying haze density.

It is a well-known fact that ensemble learning [20] has proven to be effective in reducing the variance of the neural networks. The performance of an ensemble model can be better than the performance of the best single network used in isolation [21]. Inspired by this fact, we propose the Ensemble Dehazing Networks to tackle the challenge of non-homogeneous haze. As a starting point, we first propose a simple model called EDN-3J, which consists of one shared Dense encoder and four Dense decoders. Among decoders, three of them output distinct $\mathbf{J}_i$, $i \in \{1, 2, 3\}$, which are learned using different reconstruction loss functions. A crucial step in combining these outputs is to effectively weigh/combine each $\mathbf{J}_i$. Moreover, since the images are corrupted by non-homogeneous haze, each pixel of each $\mathbf{J}_i$ should be assigned a learnable weight. Keeping this fact in mind, we use the fourth decoder to generate the weight maps and to combine the 3 outputs effectively.

Although EDN-3J addresses the non-homogeneous challenge to a certain extent, it does not utilize the physical haze model which is crucial to obtain reliable images. Therefore, we propose another ensemble network based on the physical haze model, denoted as EDN-AT. Similar to EDN-3J, it consists of a common Dense encoder and 4 Dense decoders. Out of these 4 decoders, 2 of them output different ambient light maps $\mathbf{A}_i$, $i \in \{1, 2\}$ and the third decoder outputs a common transmission map $\mathbf{t}$. The motivation behind this design is that the haze is primarily influenced by the atmospheric light $\mathbf{A}$. Hence we restrict the transmission map to be the same for both haze models and naturally drive two ambient light maps to focus on haze of different properties such as dense haze and light haze. To further facilitate each $\mathbf{A}_i$ to represent the ambient light in different haze scene, an attention mask $\mathbf{m}$ of the same size of ambient light map's is generated by the fourth decoder. $\mathbf{m}$ and $\mathbf{1} - \mathbf{m}$ are multiplied to $\mathbf{A}_1, \mathbf{A}_2$ accordingly. Eventually, $\mathbf{J}_1$ and $\mathbf{J}_2$ are weighted by weight map $\mathbf{w}$ from another output channel in the decoder of the attention mask. Note that a single decoder is used to generate $\mathbf{m}$ and $\mathbf{w}$.

In contrast to the previous models where we combine different dehazing networks in a parallel fashion, we develop a sequential ensemble of a Dense encoder, a Dense decoder and a U-net (which also consists of an encoder-decoder architecture but with a lesser model complexity), denoted as EDN-EDU. By cascading encoder-decoder dehazing networks having different modeling capacities, EDN-EDU model shows great ability of extracting and recovering scene information for images corrupted by complex haze.

Note that, unlike standard ensemble techniques where different models are independently learned and combined explicitly, in our approaches, different models share some common properties and also independent properties. Jointly learning with the shared encoder can boost performance of individual decoder while diversity of decoders are crucial to success of ensemble scheme. To obtain diversities of submodels, customized regularization in terms of fidelity and perceptual quality are adopted. Through training with the image pairs $\{\mathbf{I}, \mathbf{J}\}$, three models have shown good performance on the experiments performed on challenging benchmark image datasets of NTIRE-2020 [22] and NTIRE-2019 [23]. Based on PSNR, EDN-EDU model ranks 7^{th} in the NTIRE-2020 Dehazing Challenge, while being 7^{th} in the SSIM metric. More importantly our EDN-AT ranks 1^{st} and EDN-EDU ranks 4^{th} on LPIPS, a metric which is shown to be more consistent with human perception as compared to PSNR and SSIM [24].

2. Related Work

Deep learning based methods have shown tremendous promise for recovering clean images from very dense haze. Recently, in [25], the authors proposed a deep network to estimate individual color channels followed by a subsequent refinement block to enhance the final synthesized RGB image. Various methods have been proposed to estimate $\mathbf{t}$ and $\mathbf{A}$ to reconstruct $\mathbf{J}$. For example, Yang *et al.* [26] unrolled an iterative algorithm into a deep learning framework to estimate the dark channel and transmission priors. Yuan *et al.* [27] combined Network-in-Network with multi-scale CNN to estimate $\mathbf{t}$. Ren *et al.* [28] also proposed a multi-scale deep neural network to estimate $\mathbf{t}$. These methods are all limited by their structures since only transmission map $\mathbf{t}$ is estimated through CNN frameworks. To incorporate further essential information, Li *et al.* [29] proposed a dehazing network where both $\mathbf{t}$ and $\mathbf{A}$ are encoded into one unit. Recently, in Guo *et al.*'s work [30], they introduced a shared-encoder multi-decoders architecture to be trained jointly to estimate $\mathbf{t}, \mathbf{A}$ which is proven to be very effective and achieved top place in NTIRE-2019 Dehazing Contest.

With the promise of Generative Adversarial Networks (GANs) [31] in many computer vision tasks, they have also exhibited their advantages in image dehazing tasks. In [32], the authors developed a discriminator to judge whether the corresponding dehazed image and the estimated transmission map are real or fake. In [33], the authors proposed a framework where discriminator guides the generator to create realistic images on a coarse scale while the enhancer following the generator produces realistic images on a fine scale.

Semi-supervised learning and unsupervised learning for training dehazing models have also been explored. In [34], the authors presented a semi-supervised learning algorithm in which the deep CNN has a supervised learning branch and an unsupervised learning branch. In [35], a cycle GAN is trained through unsupervised learning to remove the reliance on degraded and ground-truth image pairs.

Although all these methods have offered significant practical benefits for image dehazing, most of these methods were developed based on the assumption of the uniform haze, which may not be valid in many practical scenarios. Hence, in our proposed work, we tackle this issue by designing shared ensemble models wherein sub-models are trained partially jointly and are combined effectively to gain the benefits of different dehazing networks. This strategy enables our models to offer better performance on the non-homogeneous hazy images as well as uniform hazy images.

3. Ensemble Dehazing Network

One of the advantages of ensemble techniques is the reduced variance of overall estimated models, which is particularly beneficial in the case of non-homogeneous haze. In non-homogeneous dehazing problem, mappings from scenes of multiple haze level to haze-free scene are learned. Those non-linear mappings have distinct properties and therefore ideally are supposed to be modeled separately for avoiding confusion to the neural network. Inspired by this observation, we focus on ensembling multiple dehazing networks and facilitating them to be expert at different aspects of non-homogeneous haze. However, since sub-models in ensemble model are always less capable due to smaller structure, the number of distinct sub-models is usually required to be large to ensure considerable improvement. Since it would make the ensemble model cumbersome, we aim to address this issue by boosting the individual sub-model's capacity and at the same time maintaining their differences. In this way, we can benefit from the ensemble technique and avoid suffering from huge architecture. We observed that information sharing between two distinct sub-models is beneficial to improvement of individual model. Based on this observation, first we propose EDN-3J and EDN-AT which focus on ensembling sub-models sharing a common encoder. Second, we propose EDN-EDU which focus on ensembling two dehazing networks sequentially to let them recover information hierarchically and be trained jointly. In this section, we would illustrate details of the proposed architectures of three ensemble dehazing networks and strategies we adopted for optimizing each model.

3.1. Network Building Blocks

In our proposed models, the main building blocks are encoder and decoders, which are based on Densely Connected Network(DCN) [36] because of its compelling advantages such as the alleviation of vanishing-gradient problem, strengthening of features propagation and features reuse. The details of architectures of EDN-3J, EDN-AT, and EDN-EDU are shown below:

1) Encoder: We use pre-trained blocks that have been used for image classification tasks in [36] for the encoder, since they have already been trained on a vast amount of natural images and possess feature extracting capacity. The encoder consists of a base block, 4 Dense blocks(DB), 4 transition blocks and a residual block. Details are shown in Table. 1. The description of blocks in the table includes 4 items:

- Block: name of the current block.
- Input: name of the blocks which outputs the input of the current block.
- Structure: operating layers which might be convolution layers with kernel size and number of layers specified, max- or average-pooling of specific size, fully connected layers, etc.
- Output: $h \times w \times c$, where h, w, c are dimensions of height, width and channels of current block's output.

Table 1: Encoder Structure

Block	**Base.0**	**Dense.1**	**Trans.1**	**Dense.2**
Input	input patch/image	Base.1	Dense.1	Trans.1
Structure	$\begin{bmatrix} 7 \times 7 \text{ conv.} \\ 3 \times 3 \text{ max-pool} \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 3 \times 3 \text{ conv.} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 2 \times 2 \text{ avg-pool} \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 3 \times 3 \text{ conv.} \end{bmatrix} \times 12$
Output	$64 \times 64 \times 64$	$64 \times 64 \times 256$	$32 \times 32 \times 128$	$32 \times 32 \times 512$
Block	**Trans.2**	**Dense.3**	**Trans.3**	**Dense.4**
Input	Dense.2	Trans.2	Dense.3	Trans.3
Structure	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 2 \times 2 \text{ avg-pool} \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 3 \times 3 \text{ conv.} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 2 \times 2 \text{ avg-pool} \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 3 \times 3 \text{ conv.} \end{bmatrix} \times 12$
Output	$16 \times 16 \times 256$	$16 \times 16 \times 1024$	$8 \times 8 \times 512$	$8 \times 8 \times 768$
Block	**Trans.4**	**Res.4**		
Input	Dense.4	Trans.4		
Structure	$\begin{bmatrix} 1 \times 1 \text{ conv.} \\ 2 \times 2 \text{ avg-pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ conv.} \\ 3 \times 3 \text{ conv.} \end{bmatrix} \times 2$		
Output	$16 \times 16 \times 128$	$16 \times 16 \times 128$		

2) Decoder: the decoders are trained from scratch without specific initialization. The structure of decoder is similar to that of encoder shown in Table. 2, which consists of 4 Dense blocks, corresponding transition blocks, a residual block and several convolution layers as refinement blocks at the end of the decoder. In decoders, 3 channel attention blocks are embedded to reinforce the informative channels and to suppress less useful channels of feature maps. The structures of channel attention module and residual block are illustrated in Fig. 2 and Fig. 3. It is observed that the models have relatively stable behaviour after training phase and avoid severe over-fitting with channel attention modules incorporated. The number of channels in the output layer, denoted as X, depends on the functionality of decoder, which is shown in Table. 3

Table 2: Decoder Structure

Block	**CA.4**	**Dense.5**	**Trans.5**	**Res.5**
Input	[Res.4, Trans.2]	CA.4	Dense.5	Trans.5
Structure	$\begin{bmatrix}\text{fully connected}\end{bmatrix}$	$\begin{bmatrix}\text{batch norm}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 7$	$\begin{bmatrix}1\times 1\text{ conv.}\\ \text{upsample }2\end{bmatrix}$	$\begin{bmatrix}3\times 3\text{ conv.}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 2$
Output	$16\times 16\times 128$	$16\times 16\times 640$	$32\times 32\times 128$	$32\times 32\times 128$
Block	**CA.5**	**Dense.6**	**Trans.6**	**Res.6**
Input	[Trans.1, Res.5]	CA.5	Dense.6	Trans.6
Structure	$\begin{bmatrix}\text{fully connected}\end{bmatrix}$	$\begin{bmatrix}\text{batch norm}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 7$	$\begin{bmatrix}1\times 1\text{ conv.}\\ \text{upsample }2\end{bmatrix}$	$\begin{bmatrix}3\times 3\text{ conv.}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 2$
Output	$32\times 32\times 128$	$32\times 32\times 384$	$128\times 128\times 64$	$64\times 64\times 64$
Block	**Dense.7**	**Trans.7**	**Res.7**	**Dense.8**
Input	Res.6	Dense.7	Trans.7	Res.7
Structure	$\begin{bmatrix}\text{batch norm}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 7$	$\begin{bmatrix}1\times 1\text{ conv.}\\ \text{upsample }2\end{bmatrix}$	$\begin{bmatrix}3\times 3\text{ conv.}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 2$	$\begin{bmatrix}\text{batch norm}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 7$
Output	$64\times 64\times 64$	$64\times 64\times 128$	$128\times 128\times 32$	$128\times 128\times 32$
Block	**Trans.8**	**Res.8**	**CA.8**	**Refine.9**
Input	Dense.8	Trans.8	Res.8	CA.8
Structure	$\begin{bmatrix}1\times 1\text{ conv.}\\ \text{upsample }2\end{bmatrix}$	$\begin{bmatrix}3\times 3\text{ conv.}\\ 3\times 3\text{ conv.}\end{bmatrix}\times 2$	$\begin{bmatrix}\text{fully connected}\end{bmatrix}$	$\begin{bmatrix}3\times 3\ conv\\ 32\times 32\text{ avg-pool}\\ 1\times 1\text{ conv.}\\ \text{upsample}\end{bmatrix}$
Output	$256\times 256\times 16$	$256\times 256\times 16$	$256\times 256\times 20$	$256\times 256\times 1$
Block	**Refine.10**	**Refine.11**	**Refine.12**	**Refine.13**
Input	Refine.9	Refine.9	Refine.12	[Refine.9, .10, .11, .12]
Structure	$\begin{bmatrix}16\times 16\text{ avg-pool}\\ 1\times 1\text{ conv.}\\ \text{upsample}\end{bmatrix}$	$\begin{bmatrix}8\times 8\text{ avg-pool}\\ 1\times 1\text{ conv.}\\ \text{upsample}\end{bmatrix}$	$\begin{bmatrix}4\times 4\text{ avg-pool}\\ 1\times 1\text{ conv.}\\ \text{upsample}\end{bmatrix}$	3×3 conv.
Output	$256\times 256\times 1$	$256\times 256\times 20$	$256\times 256\times 1$	$256\times 256\times X$

Table 3: Number of Channels in Output Layers of Decoders

EDN-3J	decoder.J1	decoder.J2	decoder.J3	decoder.W
X	3	3	3	1
EDN-AT	decoder.A1	decoder.A2	decoder.W	decoder.T
X	3	3	2	1

3) U-net: the U-net [37] structure is shown in Table. 4. In total, it consists of 15 convolution layers.

Table 4: U-net Structure

Block	**unet.down**	**unet.conv**	**unet.up**	**unet.out**
Input	Refine.14	unet.down	unet.conv	unet.out
Structure	$\begin{bmatrix}3\times 3\text{ conv.}\\ 3\times 3\text{ conv.}\\ \text{max-pool}\end{bmatrix}\times 3$	$\begin{bmatrix}3\times 3\text{ conv.}\\ 3\times 3\text{ conv.}\end{bmatrix}$	$\begin{bmatrix}3\times 3\text{ conv.}\\ 3\times 3\text{ conv.}\\ \text{max-pool}\end{bmatrix}\times 3$	$\begin{bmatrix}3\times 3\text{ conv.}\end{bmatrix}$
Output	$32\times 32\times 256$	$32\times 32\times 512$	$128\times 128\times 64$	$256\times 256\times 3$

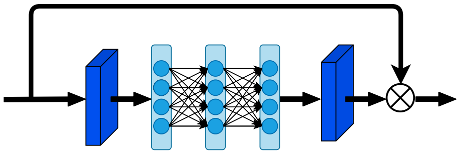

Figure 2: Structure of channel attention module(CA.4, CA.5, CA.8 in Table. 2). It pools channel information from multi-channel feature map of last layer and passes it into a 2-layer fully-connected network of which output is recovered to multiply with skipped feature map.

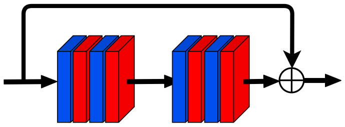

Figure 3: Structure of residual block(Res.4, Res.5, Res.6, Res.7, Res.8 in Table. 1 and 2). It adds skip connection with outputs of four convolution layers.

3.2. EDN-3J Model

The EDN-3J model utilizes the encoder in Table. 1 which is initialized as mentioned above. There are 4 decoders denoted as decoder.J1, decoder.J2, decoder.J3 and decoder.W. Decoder.J's, which are constructed based on Table. 2, outputs dehazed images $\mathbf{J}_1, \mathbf{J}_2$, $\mathbf{J}_3$ which are learned using different reconstruction loss. Aiming to find the best version of combination, decoder.W generates 3 weight maps by the 3-channel output layer. Elements of each weight map are constrained between 0 and 1 and are restricted to have sum to be 1. The formulation of final out-

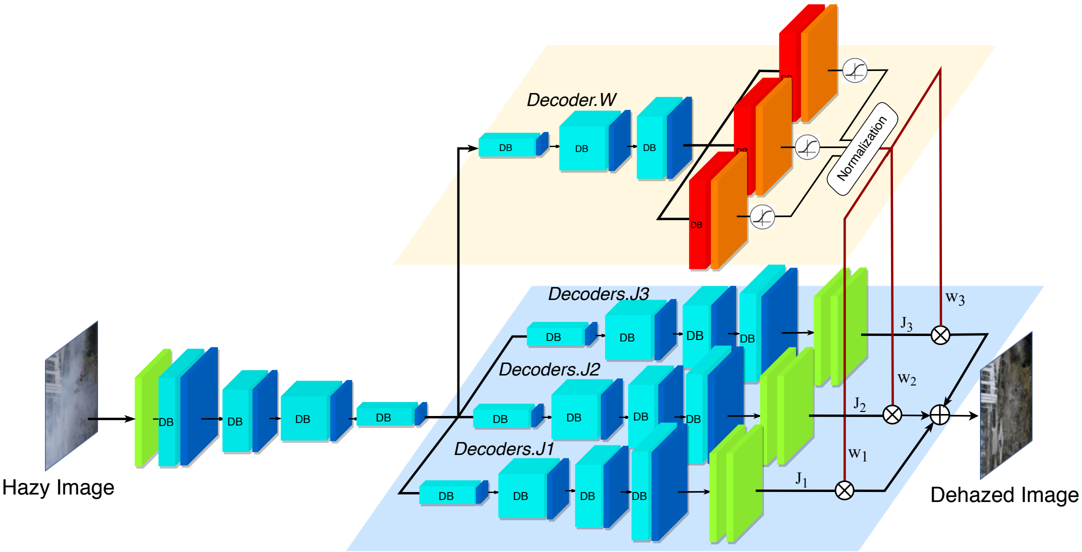

Figure 1: Architecture of the proposed 'EDN-3J' model. In EDN-3J, 3 decoders estimate $\mathbf{J}_1, \mathbf{J}_2$ and $\mathbf{J}_3$ directly and are combined via $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ generated by decoder.W. Each decoder consists of convolution layers, attention modules, residual blocks and Dense blocks(DB).

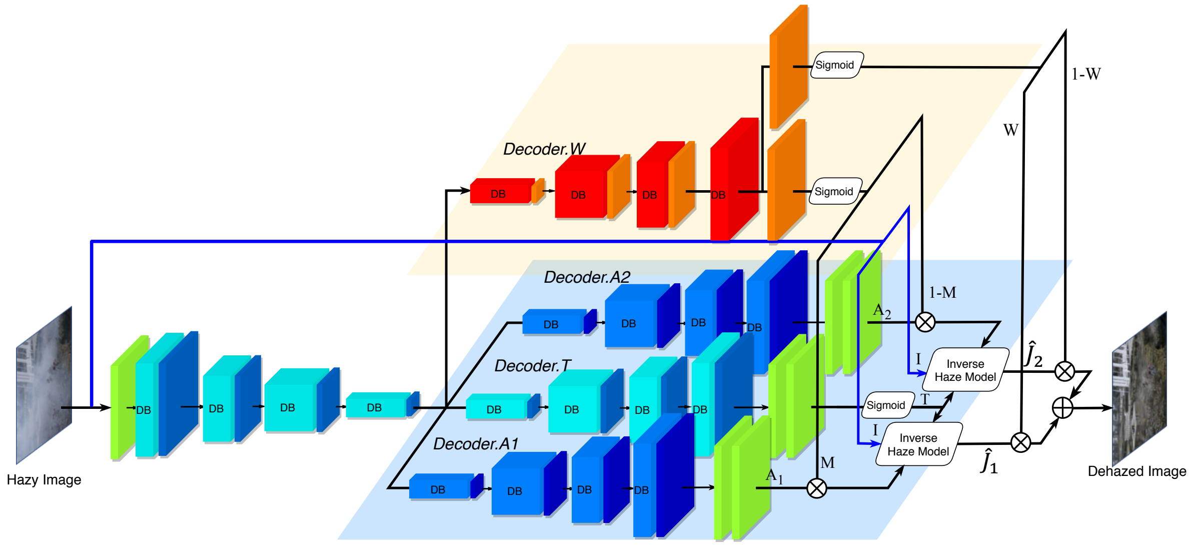

Figure 4: Architecture of the proposed 'EDN-AT' model. Ambient light maps $\mathbf{A}_1, \mathbf{A}_2$, transmission map $\mathbf{t}$ are estimated by decoder.A1, decoder.A2 and decoder.T respectively. $\mathbf{A}_1, \mathbf{A}_2$, which are then used to estimate clean images via inverse haze model in Eq. 5, are multiplied with attention maps $\mathbf{m}$ and $1-\mathbf{m}$. Final output is a combination of two sub-models' weighted outputs.

put is shown in Eq. 2:

$$\hat{\mathbf{J}} = \mathbf{w}_1 \cdot \hat{\mathbf{J}}_1 + \mathbf{w}_2 \cdot \hat{\mathbf{J}}_2 + \mathbf{w}_3 \cdot \hat{\mathbf{J}}_3 \tag{2}$$

where $\cdot$ represents element-wise multiplication, $\hat{\mathbf{J}}_i := f_{J_i}(\mathbf{I})$ and $\hat{\mathbf{w}}_i := \frac{f_{w_i}(\mathbf{I})}{f_{w_1}(\mathbf{I})+f_{w_2}(\mathbf{I})+f_{w_3}(\mathbf{I})}$ for $i = 1, 2, 3$ and $f_{w_i} \in [0, 1]$. The overall architecture is illustrated in Fig. 1.

The joint training of encoder and decoders is accomplished by minimizing the following objective function:

$$\begin{aligned} \mathcal{L} = \|\hat{\mathbf{J}}_1 - \mathbf{J}\|_2^2 + \lambda_1\|\hat{\mathbf{J}}_2 - \mathbf{J}\|_1 + \lambda_2\mathcal{L}_{SSIM}(\hat{\mathbf{J}}_3, \mathbf{J}) \\ + \lambda_3\|\hat{\mathbf{J}} - \mathbf{J}\|_2^2 \ + \ \lambda_4\mathcal{L}_{vgg}(\hat{\mathbf{J}}, \mathbf{J}) \end{aligned} \tag{3}$$

where $\lambda_i, i \in \{1, 2, 3, 4\}$ are parameters for weighing different loss terms, $\mathbf{J}$ is ground truth, $\mathcal{L}_{SSIM}(x, y) := 1 - SSIM(x, y)$ is dissimilarity measurement based on Structural Similarity Index Measure(SSIM) [38] and $\mathcal{L}_{vgg}$ is a perceptual loss that measures high-level differences, like contents and style discrepancies between images alike. $\mathcal{L}_{vgg}$ is calculated by inputting $\mathbf{J}$ and $\hat{\mathbf{J}}$ into a pre-trained *VGG16* network [39] and measuring their output features' difference in the sense of L2-norm given by:

$$\mathcal{L}_{vgg} = \sum_{i=1}^{3}\|g_i(\hat{\mathbf{J}}) - g_i(\mathbf{J})\|_2^2 \tag{4}$$

where the $g_i(\cdot)$ represents the operator of feature extraction conducted by *VGG16* model.

In the EDN-3J model, we are combining shared sub-models that are learned under different reconstruction loss functions. This offers a great benefit in the challenging scenario of non-homogeneous haze, since each loss function has certain advantages thereby motivates each sub-model to have distinct characteristics. Weight maps are generated and used to ensure that outputs of sub-models are combined effectively for the best version of final output.

3.3. EDN-AT Model

Although the EDN-3J model shows promise in dehazing non-homogeneous corrupted images, lack of knowledge about physical haze model would make it inferior. If we can estimate the physical parameters, the clean images can be recovered by:

$$\hat{\mathbf{J}} = \frac{\mathbf{I} - (1 - \hat{\mathbf{t}})\hat{\mathbf{A}}}{\hat{\mathbf{t}}} \tag{5}$$

To explore the merits of the physical haze model and also the ensemble scheme, we propose EDN-AT which consists of one shared encoder constructed and initialized in a similar fashion as EDN-3J.

Four decoders are connected to the encoder in a parallel way, denoted as decoder.T, decoder.A1, decoder.A2 and decoder.W. The dehazed image is estimated by combining the outputs of two sub-models:

$$\hat{\mathbf{J}}(x) = \hat{\mathbf{w}} \cdot \hat{\mathbf{J}}_1(x) + (1 - \hat{\mathbf{w}}) \cdot \hat{\mathbf{J}}_2(x) \tag{6}$$

where $\hat{\mathbf{J}}_1, \hat{\mathbf{J}}_2$ are sub-models' outputs, which are estimated through inversion of physical haze model. In addition to the parameters $\mathbf{A}_1, \mathbf{A}_2, \mathbf{t}$ in normal haze model, we embedded a novel parameter: attention map $\mathbf{M}$ to further enable two sub-models to focus on distinct aspects. The formation of

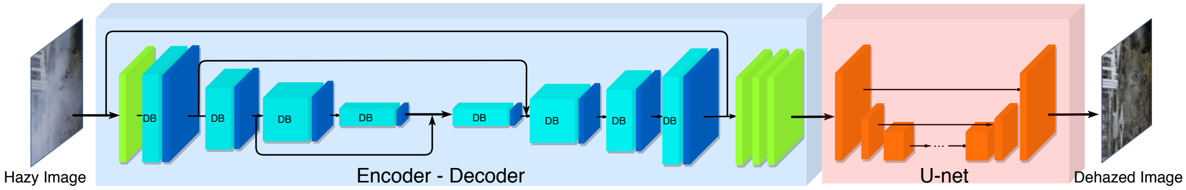

Figure 5: The architecture of the proposed 'EDN-EDU' model, which consists of a Dense encoder, Dense decoder, and a U-net.

$\hat{\mathbf{J}}_1, \hat{\mathbf{J}}_2$ is shown as follows:

$$\hat{\mathbf{J}}_1(x) = \frac{\mathbf{I}(x) - \hat{\mathbf{A}}_1(x) \cdot \hat{\mathbf{M}}(x) \cdot (1 - \hat{\mathbf{t}}(x))}{\hat{\mathbf{t}}(x)} \tag{7}$$

$$\hat{\mathbf{J}}_2(x) = \frac{\mathbf{I}(x) - \hat{\mathbf{A}}_2(x) \cdot (1 - \hat{\mathbf{M}}(x)) \cdot (1 - \hat{\mathbf{t}}(x))}{\hat{\mathbf{t}}(x)} \tag{8}$$

where $\mathbf{I}$ is hazy image, $\hat{\mathbf{t}}$ is common transmission map estimated by decoder.T, $\hat{\mathbf{A}}_1, \hat{\mathbf{A}}_2$ are ambient light intensities in two haze models estimated by decoder.A1 and decoder.A2, $\hat{\mathbf{w}}$ and $\hat{\mathbf{M}}$ are weight map and attention mask generated by 2-channel output layer from decoder.W.

The motivation here is that since the haze is primarily governed by the atmospheric light, we obtain two different ambient light maps $\mathbf{A}'s$: one for reconstructing dense hazy regions and one for light hazy regions. With the help of multiplication of attention map $\mathbf{m}$ and its complementary part $1 - \mathbf{m}$, the models would gain more expressing power to regress the non-linear mapping from dense hazy scene to haze-free scene and corresponding mapping of non-dense haze scene. The transmission map $\mathbf{t}$ remains the same for both the models as we assume $\mathbf{t}$ in both model depends on the pixel locations which do not vary. The overall architecture of EDN-AT is illustrated in Fig. 4.

The training of EDN-AT is performed by minimizing the following loss function:

$$\begin{aligned} \mathcal{L} = \|\hat{\mathbf{J}}_1 - \mathbf{J}\|_2^2 + \lambda_1 \|\hat{\mathbf{J}}_2 - \mathbf{J}\|_1 + \lambda_2 \|\hat{\mathbf{J}} - \mathbf{J}\|_2^2 \\ +\ \lambda_3\ \mathcal{L}_{SSIM}(\hat{\mathbf{J}}, \mathbf{J})\ +\ \lambda_4 \mathcal{L}_{vgg}(\hat{\mathbf{J}}, \mathbf{J}) \end{aligned} \tag{9}$$

where $\lambda_i, i \in \{1, 2, 3, 4\}$ are parameters that weigh each term, $\mathbf{J}$ is ground truth and $\mathcal{L}_{vgg}$ is perceptual loss perceived by *VGG16* network. The loss function contains 4 terms constraining outputs of sub-models and final output of the model. The shared encoder structure enhance both sub-models' individual performance significantly. Meanwhile, different reconstruction losses plus attention maps force the sub-models in EDN-AT to generate diverse outputs. With the help of learned weight maps, the final output is able to recover images in less hazy region and estimate scenes in dense hazy region.

3.4. EDN-EDU Model

In contrast to the ensemble dehazing networks introduced above, here we demonstrate a simple but effective cascading ensemble model of two dehazing blocks. In this model, a DenseNet based encoder-decoder block and a U-net based block are cascaded to form a direct mapping from hazy images to clean images, hence called EDN-EDU model. The sequential ensemble of an encoder-decoder and a U-net enables the overall model to recover scene information in hazy images hierarchically, thereby enhancing its capacity of feature extraction and recovery. The detailed structure of encoder, decoder and U-net block are shown in Table. 1, 2 and 4. The architecture of the cascaded model is shown in Fig. 5. The training of EDN-EDU is performed by minimizing the following loss function:

$$\begin{aligned} \mathcal{L} = \|\hat{\mathbf{J}}_1 - \mathbf{J}\|_2^2\ +\ \lambda_1 \|\hat{\mathbf{J}} - \mathbf{J}\|_1\ +\ \lambda_2 \|\hat{\mathbf{J}} - \mathbf{J}\|_2^2 \\ +\lambda_3\ \mathcal{L}_{SSIM}(\hat{\mathbf{J}}, \mathbf{J})\ +\ \lambda_4 \mathcal{L}_{vgg}(\hat{\mathbf{J}}, \mathbf{J}) \end{aligned} \tag{10}$$

where $\hat{\mathbf{J}}_1$ is output of dense network and $\hat{\mathbf{J}}$ is the output of U-net which also happens to to be the output of EDN-EDU model, $\mathcal{L}_{vgg}$ and $\mathcal{L}_{SSIM}$ are defined in the same way as those in EDN-3J and EDN-AT are defined.

4. Experiments

In this section we present the procedures for pre-processing training datasets and setup for the experiments.

4.1. Datasets

The EDN-3J, EDN-AT and EDN-EDN models are trained using the NTIRE-2020 non-homogeneous Dehazing dataset [22]. The dataset contains haze-free images and hazy images of the same scene. The training dataset consists of 45 pairs of hazy images and their corresponding haze-free ground truth. To enable the networks to have a better generalizing ability, we also include the outdoor images from the NTIRE-2018 Dehazing dataset [3] and NTIRE-2019 Dehazing dataset [23] for training. Both datasets contain images with homogeneous haze while haze in the NTIRE-2019 dataset has a higher density. To obtain a sizeable amount of training data, we extract patches of size 256 × 256 from these images. To further extend our training dataset, the following augmentation techniques are used: **1**) horizontal flip, rotation by $90°, 180°$ and $270°$; **2**) the original images are resized to 256 × 256 and then the same augmentation techniques are applied to them and are included in the training dataset.

4.2. Training Setup

We use a batch size of four for training with Adam [40] as the optimizer. The initial learning rate is 1×10^{-4} which is reduced to its 40% after every 10 epochs. All the 3 models are trained for 60 epochs.

5. Experimental Results

In this section we present the numerical and visual performance of our proposed models, which include ablation study and comparison with state-of-the-art methods. The evaluation metrics used to quantify the performance are Peak Signal-to-Noise Ratio(PSNR), Structural Similarity Index Measure(SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) metric. LPIPS is a novel metric that measures perceptual similarity using deep features of two images extracted by some well-known deep learning frameworks. The lower LPIPS score indicates a higher similarity between two images. The evaluation datasets include the validation dataset of NTIRE-2019 dehazing challenge and validation dataset of NTIRE-2020 dehazing challenge. The ground-truth images are only available for NTIRE-2019 dataset, while for NTIRE-2020 dataset, we submitted the results on the challenge server to obtain the results of our methods along with the competing methods.

5.1. Ablation Study

We performed an ablation study on EDN-3J and EDN-AT models to investigate the effects of different blocks. The quantitative results of the study are reported in Table. 5. In EDN-3J model, we investigated the impacts of using different losses terms for each decoder. We noticed that assigning distinct loss functions to different decoders have positive effects on the models. SSIM is particularly crucial to increasing the capacity of this ensemble model. Besides, we removed the third decoder that is constrained by $\mathcal{L}_{SSIM}$ and observed the result. In EDN-AT model, we investigated the case of attention mask $\mathbf{m}$ being removed, the case of single physical haze model with attention mask and the case of single physical haze model. From the ablation study of EDN-AT, it is observed that utilizing two haze models and combining them effectively can boost the performance. With attention map incorporated in the model, dehazed images can be further enhanced. This study is performed on NTIRE-2020 validation dataset.

Table 5: Ablation Study

	Decoder		Loss Terms				
	$\mathbf{J_1}$& $\mathbf{J_2}$	$\mathbf{J_3}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_{SSIM}$	PSNR	SSIM
EDN-3J	√	√	√	√	√	18.01	0.61
	√	√	√	√ √		17.85	0.61
	√	√		√ √ √		17.82	0.60
	√		√	√		17.70	0.58
	$\mathbf{m}$	$\mathbf{t}$	$\mathbf{A_1}$	$\mathbf{A_2}$	$\mathbf{W}$	PSNR	SSIM
EDN-AT	√	√	√	√	√	18.52	0.63
		√	√	√	√	18.46	0.62
	√	√	√			18.27	0.60
		√	√			18.20	0.59

5.2. Comparison with State-of-the-art Methods

The state-of-the-art methods used for comparisons include: TIP'15 [41], TIP'16 [28], CVPR'16 [42], ICCV'17 [29], CVPR'18 [43], CVPRW'18 [32], and CVPRW'19 [30]. Table. 6 and Table. 7 report the quantitative results of EDN-3J, EDN-AT and EDN-EDU models on NTIRE'19 validation dataset and NTIRE'20 validation dataset. From Figure. 6 and 7, it can be observed that images generated by EDN-AT and EDN-EDU models are visually better compared to the state-of-the art methods. Further, quantitatively EDN-AT and EDN-EDU produced the best PSNR values compared to the state of the alternatives on both the datasets. Among the variants of our own methods, EDN-AT and EDN-EDU produced the best results compared to EDN-3J. This is due to the facts that EDN-AT has knowledge of physical model imbibed into it and EDN-EDU has a sequential hierarchical structure with stronger feature extracting-recovering capacity thereby leading to better performance. The NTIRE-2019 validation dataset has uniform haze which is a special case of non-homogenous haze. From the results of comparisons, the issue of uniform haze is also effectively addressed by our EDN models.

Table 6: The average PSNR/SSIM of different methods over NTIRE-2019 **validation** dataset.

Team	Contest Method	PSNR	SSIM
Other Methods	TIP15[41]	13.29	0.38
	TIP16[28]	14.56	0.42
	CVPR16[42]	15.98	0.45
	ICCV17[29]	15.67	0.51
	CVPR18[43]	16.30	0.48
	CVPRW18 [32]	15.69	0.47
	CVPRW19[30]	17.15	0.52
Ours	EDN-3J	16.87	0.49
	EDN-AT	**17.44**	**0.55**
	EDN-EDU	17.21	0.51

Table 7: The average PSNR/SSIM of different methods over NTIRE-2020 **validation** dataset.

Team	Contest Method	PSNR	SSIM
Other Methods	TIP15 [41]	14.59	0.55
	TIP16 [28]	15.94	0.57
	CVPR16 [42]	16.13	0.60
	ICCV17 [29]	17.97	0.62
	CVPR18 [43]	17.90	0.63
	CVPRW18 [32]	18.23	0.62
	CVPRW19 [30]	18.45	**0.64**
Ours	EDN-3J	18.01	0.61
	EDN-AT	18.52	0.63
	EDN-EDU	**18.92**	0.63

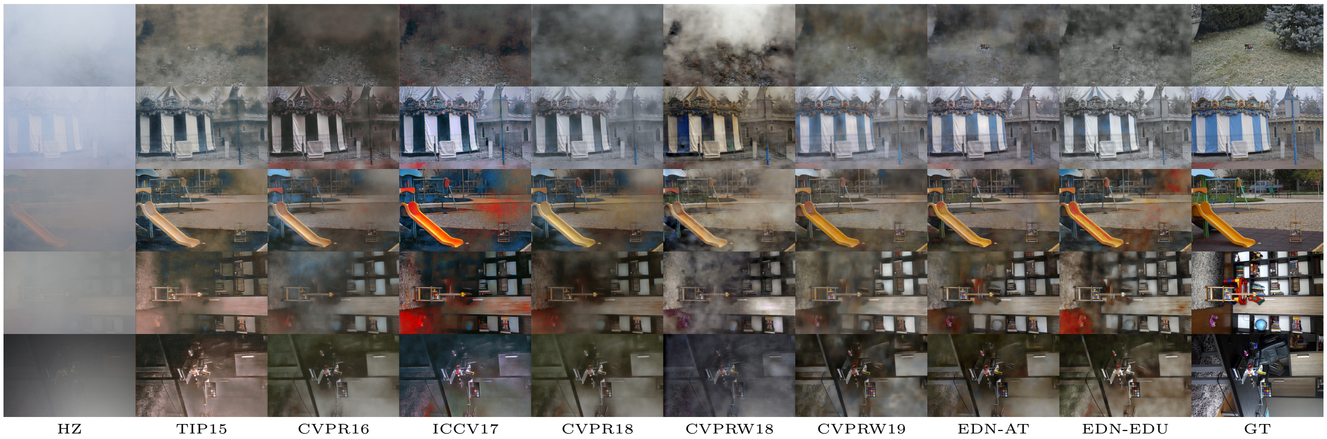

Figure 6: Visual results of different state-of-the-art methods on validation dataset of NTIRE-2019 Dehazing Competition.

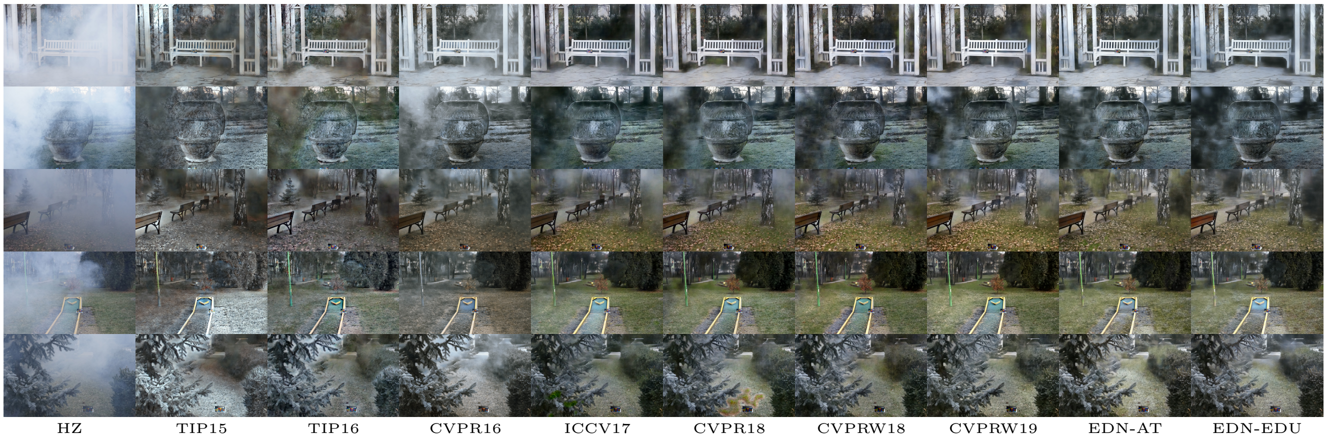

Figure 7: Visual results of different state-of-the-art methods on validation dataset of NTIRE-2020 Dehazing Competition. The GT images for this dataset are not provided.

NTIRE-2020 Dehazing Challenge The haze presented in images from NTIRE-2020 Dehazing dataset is non-homogeneous, which is a new challenge that has not been addressed in the previous literature. In evaluation phase of NTIRE-2020 Dehazing challenge, both fidelity and perceptual quality are taken into consideration.

Table 8: The average PSNR/SSIM/LPIPS of top methods over NTIRE-2020 **testing** dataset.

Team	Contest Method	PSNR	SSIM	LPIPS
Top methods	method1	**21.91**$^{(1)}$	0.69$^{(2)}$	0.361
	method2	21.60$^{(2)}$	0.67	0.363
	method3	21.41$^{(3)}$	**0.71**$^{(1)}$	0.267$^{(2)}$
	method4	20.85$^{(4)}$	0.69$^{(2)}$	0.285$^{(3)}$
	method5	20.11$^{(5)}$	0.66	0.351
	method6	20.10$^{(6)}$	0.69$^{(2)}$	0.330
Ours	EDN-EDU	19.76$^{(7)}$	0.67$^{(7)}$	0.289$^{(4)}$
	EDN-AT	19.22	0.66	**0.266**$^{(1)}$
	EDN-3J	18.58	0.63	0.303

Table. 8 includes the top methods from the contest in terms of the performance of PSNR. The superscription of the number represents the ranking of the model in terms of the corresponding metric. As we can see from the table, EDN-AT model ranks 1^{st} in LPIPS while EDN-EDU model ranks 7^{th} in PSNR, 4^{th} in LPIPS and 7^{th} in SSIM metric. These results validate the effectiveness of our proposed models.

6. Conclusion

We developed ensemble dehazing networks to address the challenge of non-homogeneous haze. We proposed three new models that are particularly effective in recovering clean images from non-homogeneous haze. Our EDN-AT and EDN-EDU models achieved excellent results in the NTIRE-2020 Dehazing Challenge. EDN-AT model benefits from the incorporation of the physical model into the deep-learning framework, while EDN-EDU benefits from the rich modeling capacity of a sequential hierarchical framework. As a future direction, we can extend our ensemble framework by combining all the 3 proposed models efficiently thereby increasing the performance further.

References

[1] 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. 1

[2] C. Ancuti et al., “I-haze: a dehazing benchmark with real hazy and haze-free indoor images,” in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2018, pp. 620–631. 1

[3] C. O. Ancuti et al., “O-haze: a dehazing benchmark with real hazy and haze-free outdoor images,” in *Proc. IEEE Conf. Workshop on Comp. Vis. Patt. Recog.*, 2018, pp. 754–762. 1, 6

[4] C. Ancuti et al., “Night-time dehazing by fusion,” in *Proc. IEEE Conf. on Image Proc.*, 2016, pp. 2256–2260. 1

[5] C. O. Ancuti et al., “Locally adaptive color correction for underwater image dehazing and matching,” in *Proc. IEEE Conf. on Image Proc.*, 2017, pp. 1–9. 1

[6] C. O. Ancuti and C. Ancuti, “Single image dehazing by multi-scale fusion authors,” *IEEE Trans. on Image Proc.*, vol. 22, no. 8, pp. 3271–3282, 2013. 1

[7] W. Ren et al., “Gated fusion network for single image dehazing,” in *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, 2018, pp. 3253–3261. 1

[8] C. Chen, M. N. Do, and J. Wang, “Robust image and video dehazing with visual artifact suppression via gradient residual minimization,” in *Proc. IEEE European Conf. on Comp. Vision*. Springer, 2016, pp. 576–591. 1

[9] R. Li et al., “Single image dehazing via conditional generative adversarial network,” in *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, 2018, pp. 8202–8211. 1

[10] S. G. Narasimhan and S. K. Nayar, “Chromatic framework for vision in bad weather,” in *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, 2000, vol. 1, pp. 598–605. 1

[11] Z. Li et al., “Simultaneous video defogging and stereo reconstruction,” in *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, 2015, pp. 4988–4997. 1

[12] E. J. McCartney, “Optics of the atmosphere: scattering by molecules and particles,” *New York, John Wiley and Sons, Inc., 1976. 421 p.*, 1976. 1

[13] T. Guo et al., “Deep wavelet prediction for image super-resolution,” in *Proc. IEEE Conf. Workshop on Comp. Vis. Patt. Recog.*, 2017. 1

[14] T. Guo et al., “Orthogonally regularized deep networks for image super-resolution,” in *Proc. IEEE Int. on Conf. Acoustics, Speech, and Signal Proc.*, 2018, pp. 1463–1467. 1

[15] T. Guo, H. S. Mousavi, and V. Monga, “Adaptive transform domain image super-resolution via orthogonally regularized deep networks,” *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4685–4700, 2019. 1

[16] S. Nah et al., “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, 2017, pp. 3883–3891. 1

[17] J. Xie, L. Xu, and E. Chen, “Image denoising and inpainting with deep neural networks,” in *Proc. Advances in Neural Information Proc. Systems*, 2012, pp. 341–349. 1

[18] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT’2010*, pp. 177–186. Springer, 2010. 1

[19] C. O. Ancuti, C. Ancuti, F.-A. Vasluianu, and R. T. et al., “Ntire 2020 challenge on nonhomogeneous dehazing,” *IEEE CVPR, NTIRE Workshop*, 2020. 2

[20] T. G. Dietterich et al., “Ensemble learning,” *The handbook of brain theory and neural networks*, vol. 2, pp. 110–125, 2002. 2

[21] C. M. Bishop et al., *Neural networks for pattern recognition*, Oxford university press, 1995. 2

[22] C. O. Ancuti, C. Ancuti, and R. Timofte, “NH-HAZE: An image dehazing benchmark with nonhomogeneous hazy and haze-free images,” *IEEE CVPR, NTIRE Workshop*, 2020. 2, 6

[23] C. O. Ancuti et al., “Dense haze: A benchmark for image dehazing with dense-haze and haze-free images,” *arXiv preprint arXiv:1904.02904*, 2019. 2, 6

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595. 2

[25] T. Guo, V. Cherukuri, and V. Monga, “Dense ‘123’ color enhancement dehazing network,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[26] D. Yang and J. Sun, “Proximal dehaze-net: A prior learning-based deep network for single image dehazing,” in *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[27] K. Yuan, J. Wei, W. Lu, and N. Xiong, “Single image dehazing via nin-dehazenet,” *IEEE Access*, vol. 7, pp. 181348–181356, 2019. 2

[28] W. Ren et al., “Single image dehazing via multi-scale convolutional neural networks,” in *Proc. IEEE European Conf. on Comp. Vision*. Springer, 2016, pp. 154–169. 2, 7

[29] B. Li et al., “Aod-net: All-in-one dehazing network,” in *Proc. IEEE Conf. on Comp. Vis.*, 2017. 2, 7

[30] T. Guo, X. Li, V. Cherukuri, and V. Monga, “Dense scene information estimation network for dehazing,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 7

[31] I. Goodfellow et al., “Generative adversarial nets,” in *Proc. Advances in Neural Information Proc. Systems*, 2014, pp. 2672–2680. 2

[32] H. Zhang, V. Sindagi, and V. M. Patel, “Multi-scale single image dehazing using perceptual pyramid deep network,” in *Proc. IEEE Conf. Workshop on Comp. Vis. Patt. Recog.*, 2018, pp. 902–911. 2, 7

[33] Y. Qu, Y. Chen, J. Huang, and Y. Xie, “Enhanced pix2pix dehazing network,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[34] L. Li, Y. Dong, W. Ren, J. Pan, C. Gao, N. Sang, and M.-H. Yang, “Semi-supervised image dehazing,” *IEEE Transactions on Image Processing*, vol. 29, pp. 2766–2779, 2019. 3

[35] L.-Y. Huang, J.-L. Yin, B.-H. Chen, and S.-Z. Ye, "Towards unsupervised single image dehazing with deep learning," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2741–2745. 3

[36] G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, 2017, pp. 4700–4708. 3

[37] 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. 4

[38] Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004. 5

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 5

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 7

[41] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. on Image Proc.*, vol. 24, no. 11, pp. 3522–3533, 2015. 7

[42] B. Cai et al., "Dehazenet: An end-to-end system for single image haze removal," *IEEE Trans. on Image Proc.*, vol. 25, no. 11, pp. 5187–5198, 2016. 7

[43] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, 2018, pp. 3194–3203. 7