Zero-Shot Image Dehazing

Boyun Li, Yuanbiao Gou, Jerry Zitao Liu, Hongyuan Zhu, *Member, IEEE*,
Joey Tianyi Zhou, and Xi Peng, *Member, IEEE*

***Abstract*—In this article, we study two less-touched challenging problems in single image dehazing neural networks, namely, how to remove haze from a given image in an unsupervised and zero-shot manner. To the ends, we propose a novel method based on the idea of layer disentanglement by viewing a hazy image as the entanglement of several "simpler" layers, *i.e.*, a hazy-free image layer, transmission map layer, and atmospheric light layer. The major advantages of the proposed ZID are two-fold. First, it is an unsupervised method that does not use any clean images including hazy-clean pairs as the ground-truth. Second, ZID is a "zero-shot" method, which just uses the observed single hazy image to perform learning and inference. In other words, it does not follow the conventional paradigm of training deep model on a large scale dataset. These two advantages enable our method to avoid the labor-intensive data collection and the domain shift issue of using the synthetic hazy images to address the real-world images. Extensive comparisons show the promising performance of our method compared with 15 approaches in the qualitative and quantitive evaluations. The source code could be found at www.pengxi.me.**

***Index Terms*—Single image dehazing, unsupervised, zero-shot.**

I. Introduction

HAZE is a typical atmospheric phenomenon in which the dust, smoke, and other dry particles obscure the sky. These floating particles greatly absorb and scatter the light, leading to poor contrast and loss of details. Besides the poor visual quality, many vision tasks such as object detection would suffer from performance degradation due to the bad visibility of hazy images. Therefore, as an important visual enhancement technology, image dehazing has been extensively studied and achieved remarkable performance [1]–[8].

In recent, the focus of the community has shifted to detecting and removing haze from a single image [1], [2], [5]. Such a so-called single image dehazing aims to eliminate the scattered light to increase scene visibility and recover haze-free scene, and most of them employ a widely recognized atmospheric scattering model [9]. To estimate the global atmospheric light and the pixel-wise transmission coefficient of the atmospheric scattering model, a variety of methods have been proposed, which could be roughly divided into prior- and learning-based methods.

Prior-based methods are mainly based on some handcrafted priors derived from the image [2], [5], [10]–[13]. For instance, Tan [14] proposes dehazing by maximizing the local contrast of the image, based on the assumption that the clean images are usually more contrast than the hazy images. Berman *et al.* [11] propose a non-local image dehazing method (NLD) based on the assumption that the colors of a haze-free image could be well approximated by a few hundred distinct colors, thus forming tight clusters in RGB space for dehazing. Zhu *et al.* [5] propose color attenuation prior (CAP) that haze will decrease the image saturation and simultaneously increase the brightness. Although remarkable performance has been achieved by these methods, haze removal quality heavily depends on the consistency between the adopted prior and real data distribution

To alleviate the dependence on the predetermined prior, a lot of efforts have been devoted to designing data-driven methods based on deep neural networks [1], [3], [4], [6], [7], [15]. Different from the prior-based method, the data-driven method does not employ handcrafted prior that is assumed to exist in images. Instead, it detects and removes the haze from a single image by directly learning the atmospheric scattering parameters from training data and building the mapping between a hazy image and the corresponding clean one. For example, Cai *et al.* [1] propose a trainable convolution neural network which is trained on a large-scale hazy-clean image pair database. More specifically, it takes a hazy image as the input and outputs the corresponding transmission map which is further used to recover the hazy-free image.

Although data-driven methods have achieved state-of-the-art performance in single image dehazing, they have suffered from the following limitations. To be specific, almost all of them require a large scale hazy-clean image pairs to train their models, and such a requirement is usually satisfied by artificially synthesizing hazy images through the physical model with the handcrafted parameters and the clean image. As pointed out in [16], the synthesized database is less informative and inconsistent with the real hazy images, thus leading to the so-called domain shift issue. Therefore, it is highly expected to develop unsupervised and "zero-shot" models, where "unsupervised" avoids the collection of the

Manuscript received March 23, 2020; revised July 17, 2020 and August 7, 2020; accepted August 8, 2020. Date of publication August 18, 2020; date of current version August 24, 2020. This work was supported in part by NFSC under Grant U19A2081, Grant 61625204, and Grant U19A2078; in part by the Fundamental Research Funds for the Central Universities under Grant YJ201949; in part by the Fund of Sichuan University—Tomorrow Advancing Life; and in part by the Singapore Government's Research, Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering Domain) under Grant A18A1b0045. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abdesselam S. Bouzerdoum. *(Corresponding author: Xi Peng.)*

Boyun Li, Yuanbiao Gou, and Xi Peng are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: liboyun.gm@gmail.com; gouyuanbiao@gmail.com; pengx.gm@gmail.com).

Jerry Zitao Liu is with the TAL AI Laboratory, Beijing 100080, China (e-mail: liuzitao@100tal.com).

Hongyuan Zhu and Joey Tianyi Zhou are with the Research Agency for Science, Technology and Research (A*STAR), Singapore 138634 (e-mail: zhuhongyuan.cn@gmail.com; joey.tianyi.zhou@gmail.com).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2020.3016134

image pairs and "zero-shot" avoids using the information beyond the observed single hazy image itself covered. To the best of our knowledge, such an idea however is less touched so far.

To this end, we propose a Zero-shot Image Dehazing method (ZID). Our idea comes from the elegant assumption of layer disentanglement which views an image as an entanglement of several "simpler" layers/factors. Considering the dehazing task, we specifically view the clean image, transmission map and atmospheric light as three layers that entangle together to form the hazy image. With this idea, the proposed ZID employs three joint subnetworks to disentangle the input hazy image into these three layers, thus recovering the clean image and estimating the haze.

Our method embraces the powerful representative capacity of neural networks, which however is significant different from most of existing deep learning based methods [1], [3], [4], [6], [15], [17]–[19] in the following aspects: 1) the proposed ZID works in an unsupervised rather than supervised manner. In other words, our method does not need the hazy-clean pair images, which avoids the intensive labor for image collection. Though a hint is used to train one subnetwork of our model, it is estimated from the hazy image and no ground-truth is needed; 2) ZID is a "zero-shot" method. In other words, our method does not require training on a dataset like these existing models. Instead, it only exploits the information contained in the observed single hazy image. It is worth noting that the definition of "zero-shot" in our paper is different from the conventional zero-shot learning used in the classification scenario. In brief, the vanilla zero-shot learning often refers to training a model on a dataset and then using the model to predict the unseen categories, whereas our zero-shot setting only refers to using the observed single image and no additional data set are needed. These two differences make our method avoid the labor-intensive data collection and the domain-shift issue of using the synthetic hazy images to address the real-world images.

To summarize, the contributions of this work are given as follows:

- To the best of our knowledge, this work could be one of the first unsupervised and zero-shot deep models for image dehazing. In brief, the proposed method removes haze in an end-to-end manner and does not need extra information beyond the observed hazy image. Note that, the most similar method with our idea may be [20] which is however with significant difference from this work (see Section II for details).
- A jointly learned neural network (*i.e.*, ZID) is proposed, which consists of three joint disentanglement subnetworks. In brief, two convolutional auto-encoders are used to obtain the clean image and the transmission map, and a variational auto-encoder is used to obtain the atmospheric light.

II. Related Work

Most single image dehazing approaches are based on the atmospheric scattering model and their difference mainly lies in the estimation approach of the pixel-wise transmission coefficients and the atmospheric light. Accordingly, most of them could be classified into two categories, *i.e.*, prior- and learning-based methods or called data-driven methods. In this section, we will briefly discuss some typical works of these categories.

A. Prior-Based Methods

To estimate the under-constrained haze generation model, a large number of works adopted various hand-crafted image priors to cast the dehazing as an energy minimization problem. These prior-based methods [2], [5], [11], [14] usually require a non-trivial optimization scheme and the dehazing performance largely depends on the consistency between the adopted prior and the data distribution. In practice, however, these priors would be easily violated, especially, when the background is complex or the illumination is irregular. For example, dark channel prior (DCP) [2] assumes that most local patches in outdoor haze-free images have at least one dark channel whose intensity is close to zero. For the sky region or bright objects which are similar to the atmospheric light, DCP cannot achieve the encouraging result.

Different from existing prior-based approaches, the proposed ZID does not depend on the handcrafted priors, thus avoiding the performance degradation due to the inconsistency between the prior and data distribution. Note that, ZID employs some latent structures to supervise the subnetworks, which however adopts a data-driven rather than handcrafted way. Moreover, almost all these prior-based approaches are based on shallow models, whereas our ZID is a deep neural network.

B. Learning-Based Methods

Motivated by the success of deep learning, some recent works [1], [3], [4], [6], [15], [17], [21]–[29] employ a deep neural network to recover the clean image from a given hazy image in a data-driven way. For example, DehazeNet [1] recovers the haze-free images under the help of the haze-clean image pairs. Multi-scale Convolutional Neural Network (MSCNN) [3] consists of a coarse-scale net which learns a holistic transmission map based on the entire image, and a fine-scale net which locally refines the dehazed results.

Although our method is also a deep learning based method, it is remarkably different from existing approaches in the following aspects. First, most existing learning-based methods work in a supervised manner, whereas the proposed ZID is an unsupervised approach. To be specific, almost all of these methods [1], [3], [4], [6], [15], [17] try to learn a haze removal model by taking the hazy image as the input and the clean image as the label to train the neural network. In contrast, our ZID only takes the hazy image as the input and does not require the ground truth clean image. Second, these learning-based methods usually train a neural network using an image collection, whereas our method only requires a single image. Note that, some existing methods such as [4] have also explicitly utilized the layer disentanglement idea, however, almost all of them are supervised approaches and trained on a large scale dataset, which are significantly different from this work.

C. Advances in Unsupervised and Zero-Shot Methods

Recent years, some unsupervised deep methods [20], [30]–[34] have attracted some attentions in image enhancement and restoration. For instance, Noise2Noise (N2N) [30] leverages the basic statistical reasoning to signal reconstruction by only using the degraded images. Deep image Prior (DIP) [32] finds out that most image statistics are captured by the structure of the convolutional image generator instead of the learning process. Based on this observation, DIP proposes a new approach to recover a clean image using the early-stopping strategy with a CNN-like structure. Though both N2N and DIP have made great progress in image enhancement, there are still some limitations in practice. For example, their performance largely depends on either specially designed datasets or an uncertain early-stopping strategy. Moreover, they are not specifically designed for image dehazing, thus might result in undesirable experimental results.

It is worth noting that ZID remarkably differs from Double-DIP [20] on mainly two aspects. Firstly, the loss function is remarkably different. In brief, our method leverages the property of dark channel by minimizing it into the loss of the J-Net, whereas Double-DIP does not. Secondly, the observation and the working mechanism are totally different. Based on the properties of DIP, Double-DIP adopts three U-Net-like networks, which are similar to DIP. Double-DIP feeds three random noises as inputs into the networks to fit the clean image, transmission map and the global airlight. While ZID is based on the layer disentanglement idea and assumes that haze is a special type of content-independent noise. Based on these assumptions, our method adopts a variational module and directly feeds the hazy image into three subnetworks to disentangle different layers. Moreover, ZID is also different from YOLY [34] in the following two aspects. On the one hand, the loss function is totally different. Specifically, ZID proposes a DCP-like loss to train J-Net, whereas YOLY leverages the property of color attenuation prior (CAP). Moreover, ZID enforces the smooth regularization on the output of both T-Net and A-Net, whereas YOLY only enforces the regularization on A-Net. On the other hand, the network structure is different. ZID adopts a U-Net-like structure [32], whereas YOLY is based on a non-degenerate architecture [35].

III. Proposed Methods

In this section, we introduce the proposed ZID model which consists of a clean image estimation network (J-Net), a transmission map estimation network (T-Net), and an atmospheric light estimation network (A-Net). For clarity, we will first introduce the proposed loss function and then elaborate the implementation details of each subnetwork.

A. The Loss Function

Our idea comes from the observation on the widely-used atmospheric scatter model which describes that the hazing process could be regarded as the entanglement of different factors. Formally,

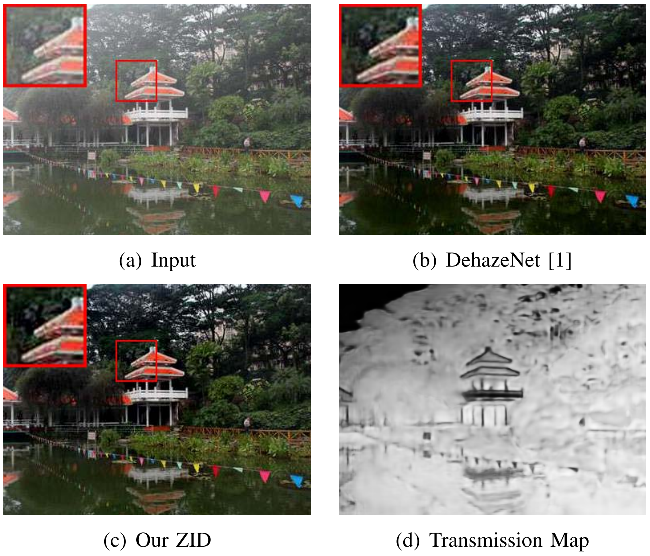

(a) Input
(b) DehazeNet [1]
(c) Our ZID
(d) Transmission Map

Fig. 1. A real world single image haze removal example of the proposed ZID. ZID performs a better visual results compared with DehazeNet [1]. Zooming-in is recommended for the comparisons in more details.

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

where $I(x)$ and $J(x)$ denote the hazy and the clean image, $t(x)$ is the medium transmission map and A is the global atmospheric light on each pixel coordinates. Although some works have pointed out that the core of recovering the clear image from its hazy version is to estimate the parameters in Eq.(1) and some methods [4] have been proposed based on the layer disentanglement idea. To the best of our knowledge, however, there is few efforts have been devoted to developing unsupervised deep methods so far.

In this article, to estimate these parameters from a single image without the help of the ground truth and additive image collection, we proposed ZID which consists of three joint learning subnetworks as shown in Fig. 2. These three subnetworks, *i.e.*, a clean image estimation network $f_J(\cdot)$ (J-Net), a transmission map estimation network $f_T(\cdot)$ (T-Net), and an atmospheric light estimation network $f_A(\cdot)$ (A-Net), are jointly trained via the following loss function

$$\mathcal{L} = \mathcal{L}_{Rec} + \mathcal{L}_A + \mathcal{L}_J + \mathcal{L}_{Reg}. \tag{2}$$

where $\mathcal{L}_{Rec}$ is the reconstruction loss between the input hazy image x and the reconstructed hazy image $I(x)$, $\mathcal{L}_A$ is the loss on the estimated atmospheric light, $\mathcal{L}_J$ is the statistics-based loss on the estimated hazy free image, and $\mathcal{L}_{Reg}$ is the regularization term on the outputs of the subnetworks. We elaborate these three terms and the related subnetworks as follows.

To be specific, $\mathcal{L}_{Rec}$ aims to disentangle each hazy image into different "simpler" layers by minimizing

$$\mathcal{L}_{Rec} = \|I(x) - x\|_p, \tag{3}$$

where $\|\cdot\|_p$ denotes p-norm of a given data matrix. In this article, we simply adopt Frobenius norm. For a given hazy

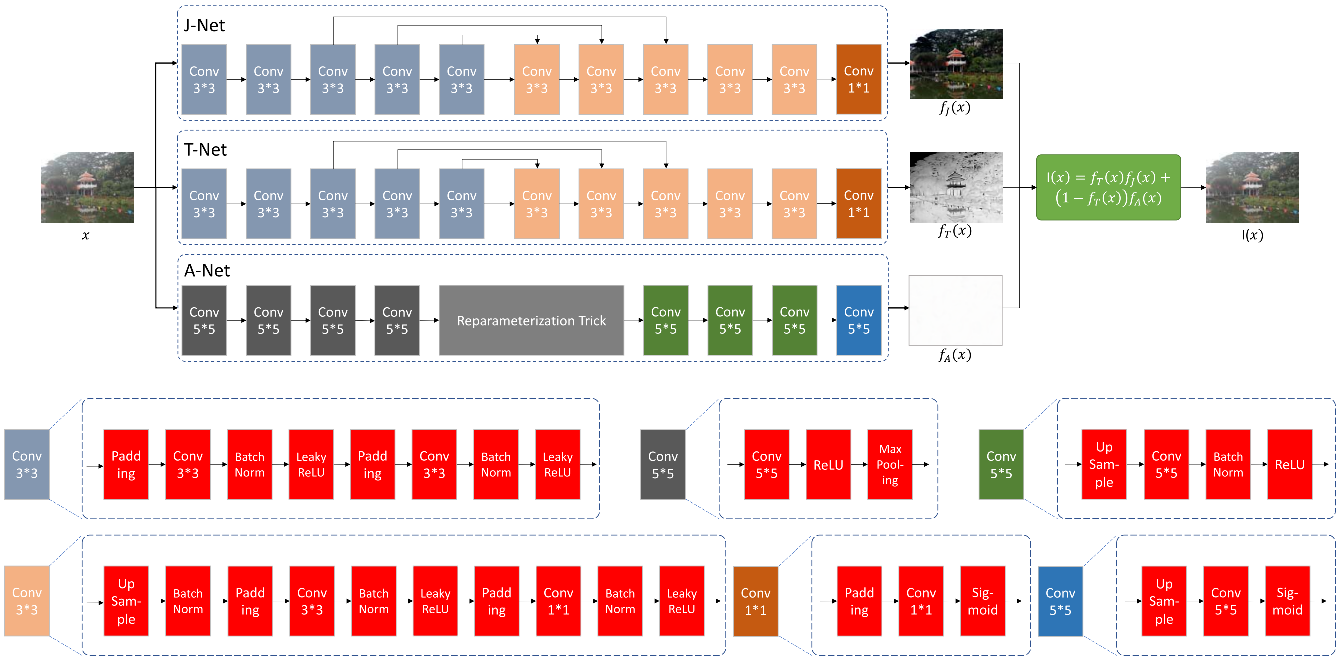


Fig. 2. The framework of our proposed ZID. ZID includes three parts: clean image estimation network (J-Net), the transmission map estimation network (T-Net), and the atmospheric light estimation network (A-Net).

TABLE I

RESULTS ON THE SYNTHETIC INDOOR DATABASE (SOTS). THE BOLD NUMBER INDICATES THE BEST METHOD OF EACH CATEGORY OF METHODS

Metrics	Supervised Methods				Unsupervised Methods						
	DehazeNet	MSCNN	AOD-Net	CAP	DCP	FVR	BCCR	GRM	NLD	N2N	DCPLoss
PSNR	**21.14**	17.57	19.06	19.05	16.62	15.72	16.88	18.86	17.29	14.49	**19.25**
SSIM	0.8472	0.8102	**0.8504**	0.8364	0.8179	0.7483	0.7913	**0.8553**	0.7489	0.7078	0.8320

Metrics	Zero-Shot Methods				
	N2V	DIP	DD	DDIP	Ours
PSNR	10.67	12.28	11.92	16.97	**19.83**
SSIM	0.5397	0.5782	0.6404	0.7147	**0.8353**

TABLE II

RESULTS ON THE SYNTHETIC OUTDOOR DATABASE (HSTS). THE BOLD NUMBER INDICATES THE BEST METHOD OF EACH CATEGORY OF METHODS

Metrics	Supervised Methods				Unsupervised Methods						
	DehazeNet	MSCNN	AOD-Net	CAP	DCP	FVR	BCCR	GRM	NLD	N2N	DCPLoss
PSNR	**24.48**	18.64	20.55	21.53	14.84	14.48	15.08	18.54	18.92	-	**24.44**
SSIM	**0.9153**	0.8168	0.8973	0.8726	0.7609	0.7624	0.7382	0.8184	0.7411	-	**0.9330**

Metrics	Zero-Shot Methods				
	N2V	DIP	DD	DDIP	Ours
PSNR	11.79	14.55	14.66	20.91	**22.65**
SSIM	0.5450	0.5573	0.6409	0.8842	**0.9011**

image x, $I(x)$ is computed using the outputs of the three subnetworks via Eq.(1). Clearly, $\mathcal{L}_{Rec}$ could constrain the entire network including the subnetworks to well reconstruct the hazy image after layer disentanglement. In other words, it guides the layer disentanglement through incorporating the haze creation process. Furthermore, it provides a supervisor to train our T-Net given J-Net and A-Net.

Different from $\mathcal{L}_{Rec}$, $\mathcal{L}_A$ only involves A-Net rather than all the three subnetworks, which aims to disentangle the atmospheric light from x only using variational inference [36]. In mathematical,

$$\mathcal{L}_A = \mathcal{L}_H + \mathcal{L}_{KL} \tag{4}$$

where $\mathcal{L}_H$ is the loss between the disentangled atmospheric light $f_A(x)$ and the initial hint $A(x)$. Note that, $f_A(x)$ and $A(x)$ are different from A in Eq.(1). In brief, A is the "absolutely accurate" global atmospheric light, $f_A(x)$ denotes the estimation given by our A-Net for the input x, and $A(x)$ is the initial hint which is automatically estimated from data. With the above notations, we have

$$\mathcal{L}_H = \|f_A(x) - A(x)\|_F, \tag{5}$$

and $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence which enforces the latent variables $z \in [\mu_z, \sigma_z^2]$ be consistent with a normal Gaussian distribution $\mathcal{N}(0, I)$. To enjoy the end-to-end optimization using the standard stochastic gradient methods, the reparameterization trick is performed on the variational lower bound to yield a lower bound estimator that can be straightforwardly optimized using standard stochastic gradient methods, then we have

$$\begin{aligned}\mathcal{L}_{KL} &= KL(\mathcal{N}(\mu_z, \sigma_z^2)||\mathcal{N}(0, I))\\ &= \frac{1}{2}\sum_i \left((\mu_{z_i})^2 + (\sigma_{z_i})^2 - 1 - \log(\sigma_{z_i})^2\right) \end{aligned} \tag{6}$$

where z_i denotes that the i-th dimension of z and z is learned from the input x.

The loss $\mathcal{L}_A$ is designed based on the following observations or assumptions. To be specific, Eq.(1) illustrates that the haze generation process depends on the transmission map $t(x)$ and the atmospheric light A which are image content dependent and independent, respectively. More specifically, $t(x)$ could be simplified as $t(x) = e^{-\rho d(x)}$, where ρ is the medium extinction coefficient and $d(x)$ is the scene depth. In this article, we assume that A is latently sampled from a Gaussian distribution. As a result, we progressively pass the hazy image x into a neural network to obtain the latent code z of the atmospheric light $f_A(x)$, and then enforce z is recursively sampled from a Gaussian distribution $\mathcal{N}(\mu_z, \sigma_z^2)$ via minimizing $\mathcal{L}_{KL}$ as defined in Eq.(6).

J-Net aims to decompose the hazy image x into the clean image $I(x)$ with the following loss:

$$\mathcal{L}_J = \| \min_{c\in\{r,g,b\}} (J^c(y))\|_p \tag{7}$$

where $J^c(\cdot)$ is the c-color channel of y and y is a local patch of the J-Net output $J(x)$. With such a so-called dark channel loss [2], J-Net incorporates the statistical properties from the recovered "clean images", thus avoiding an explicit ground truth on the recovered image. It should be pointed out that, [16] has recently incorporated the dark channel loss into a neural network, which is remarkably different from our work in the following two aspects. On one hand, DCPLoss [16] uses the corresponding transmission map of the prior as supervisor to compute empirical loss, while our ZID formulates the statistical properties of dark channel prior into our loss to estimate clean images. As a result, our method could avoid performance degradation as shown in our quantitative comparisons. On the other hand, our ZID is a joint learning neural network which enjoys the zero-shot merit as the aforementioned, whereas [16] still requires training on a large scale dataset.

To increase the stability of our model, we enforce the following regularization on the outputs of T-Net and A-Net, *i.e.*, $f_A(x)$ and $f_T(x)$. Mathematically,

$$\mathcal{L}_{Reg} = \lambda_1\mathcal{L}_S(f_A(x)) + \lambda_2\mathcal{L}_S(f_T(x)), \tag{8}$$

where $\{\lambda_i\}_{i=1}^2 \geq 0$ are the balanced factor. $\mathcal{L}_S(f_A(x))$ and $\mathcal{L}_S(f_T(x))$ are with the same form, which are defined as the norm of its Laplacian, i.e.,

$$\mathcal{L}_S(x) = \frac{1}{2m}\sum_{i=1}^m (x_i - \frac{1}{|N(x_i)|}\sum_{y_i\in N(x_i)} y_i)^2, \tag{9}$$

where $\mathcal{N}(x_i)$ is the second order neighborhood of x_i, $|\mathcal{N}(x_i)|$ is the neighborhood size, and m denotes the pixel number of x. Clearly, the regularizations play a role of mean filtering, which enforce $A(x)$ and $t(x)$ to be smooth. Note that, the high-frequency details of the recovered haze-free image will lose if the above regularization is enforced on the output of J-Net.

B. Network Architecture and Implementation

In this section, we elaborate the network structure and the implementation of our ZID. As shown in Fig.(2), given a hazy image x as the input, we simultaneously feed it into J-Net, T-Net, and A-Net to disentangle x into the layer of the clean background, the transmission map, and the atmospheric light. With the outputs of these three networks, we reconstruct the hazy image $I(x)$ at the top of ZID through the atmospheric scattering physical model.

As the aforementioned discussion, the clean background and the transmission map are dependent on the input x. Hence, we adopt a similar network structure for J-Net and T-Net. More specifically, J-Net and T-Net take a U-Net type architecture with the skip-connections by following [32]. The only difference between J-Net and T-Net lies in the top layer. To be specific, J-Net and T-Net are with three and one output channel, respectively. More details could refer to our supplementary materials.

As the global atmospheric light is independent of the image content, which is assumed latently sampling from a Gaussian distribution. Therefore, to implement our A-Net, we adopt a variational auto-encoder [36] structure which consists of a CNN-based encoder, a symmetric decoder, and an intermedia block. To be specific, both the encoder and the decoder consist of four blocks. In the encoder, the blocks are composed of a convolutional layer, a ReLU activation function [37], and a max pooling layer in sequence. In the decoder, the blocks sequentially perform upsampling, convolution, batch normalization [38], and ReLU activation. To learn the latent Gaussian model, the intermedia block will transform the output (*i.e.*, z) of the encoder to the mean (μ_z) and variance (σ_z^2) of a Gaussian distribution through minimizing Eq.(6), namely, $z \to \{\mu_z, \sigma_z^2\}$. With the help of the reparameterization trick, we obtain a reconstruction of the latent code through resampling from the Gaussian distribution, namely, $\mathcal{N}(\mu_z, \sigma_z^2) \to \hat{z}$. After that, $\hat{z}$ is fed into the decoder to obtain the disentangled atmospheric light $f_A(x)$ via minimizing Eq.(5). Note that, we optimize ZID including A-Net, J-Net, and T-Net in an end-to-end manner, and the above introduction (seems like separate steps) is just for better clarity.

IV. Experiments

We carry out experiments on two synthetic datasets and one real-world dataset by comparing with 15 baseline methods in terms of two performance metrics. In the following, we will first demonstrate the experimental setting and then show the qualitative and quantitative results on synthetic and real-world datasets. Then, we will conduct an ablation study and make an experiment to evaluate the influence of parameters.

A. Experimental Settings

In this section, we introduce the details of the used datasets, baselines, evaluation metrics, and implementations.

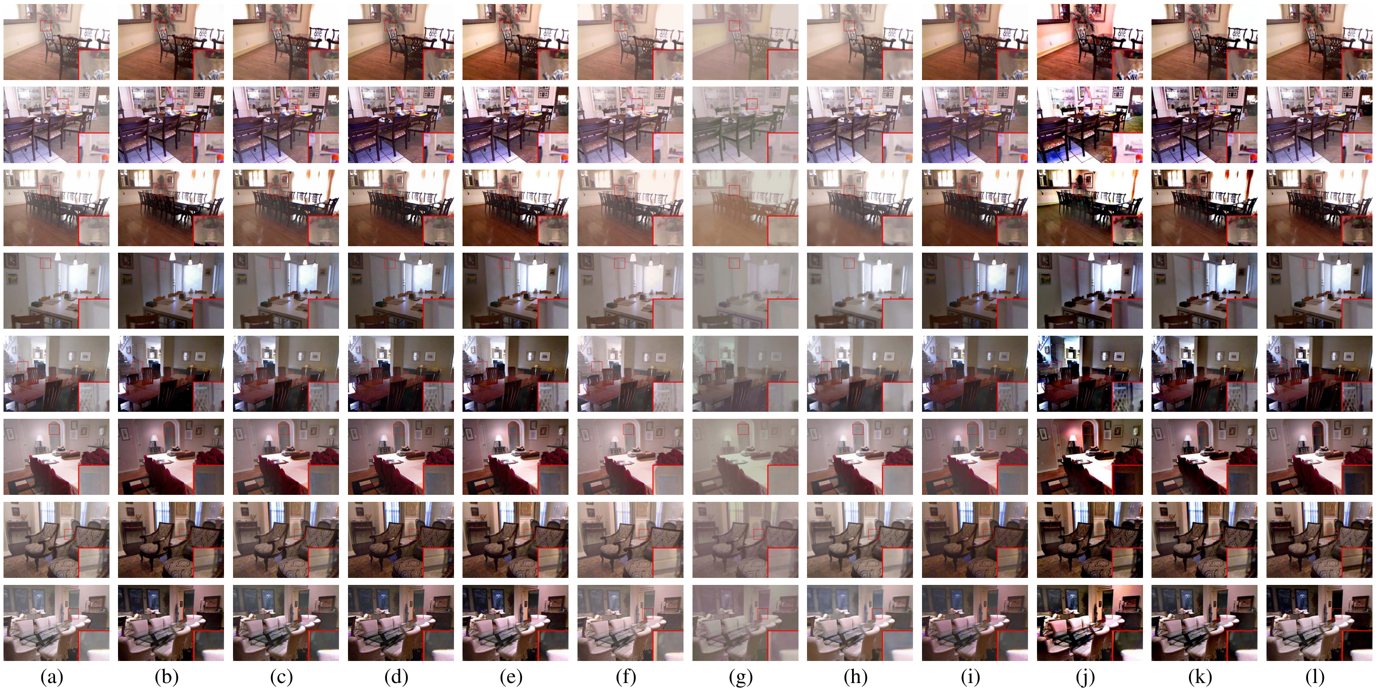


Fig. 3. Comparisons of the SOTA dehazing methods on SOTS. From the left to the right column (*i.e.*, Figs. ((a)–3(l))), the input hazy image, DehazeNet [1], MSCNN [3], AOD-Net [6], DCP [2], N2N [30], N2V [31], DIP [32], DCPLoss [16], DDIP [20], our ZID and the ground truth are presented. Some areas are highlighted by red rectangles and zooming-in is recommended for a better visualization and comparison.

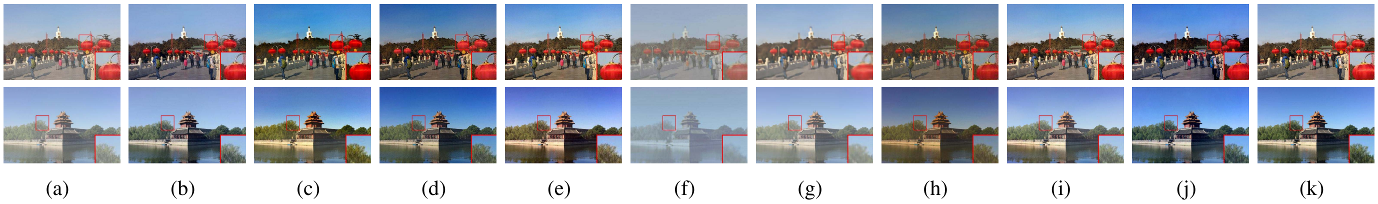


Fig. 4. Comparisons of the SOTA dehazing methods on HSTS. From the left to the right column (*i.e.*, Figs.(4(a)–4(k))), the input hazy image, DehazeNet [1], MSCNN [3], AOD-Net [6], DCP [2], N2V [31], DIP [32], DCPLoss [16], DDIP [20], our ZID and the ground truth are presented. Zooming-in is recommended for a better visualization.

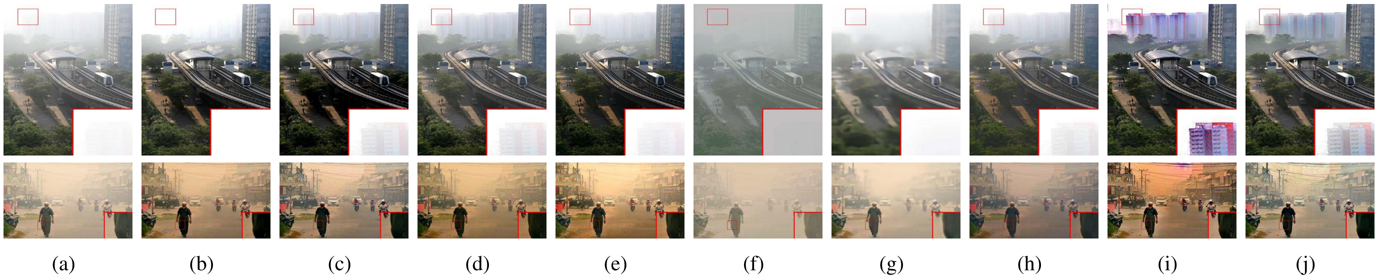


Fig. 5. Comparisons of the SOTA dehazing methods on the Real-World Dataset. From the left to the right column (*i.e.*, Figs.(5(a)–5(j))), the input hazy image, DehazeNet [1], MSCNN [3], AOD-Net [6], DCP [2], N2V [31], DIP [2], DCPLoss [16], DDIP [20] and our method are presented. Some areas are highlighted by red rectangles and zooming-in is recommended for a better visualization and comparison.

1) Datasets: We conduct experiments on a recent large scale dataset, called REalistic Single Image DEhazing (RESIDE) [39] which contains two testing subsets, *i.e.*, SOTS and HSTS. In brief, SOTS consists of 500 indoor hazy images which are synthesized using the physical model with handcrafted parameters. HSTS is an outdoor dataset consisting of 10 synthetic hazy images and 10 real-world hazy images captured from different scenes. What's more, we also manually collect 10 hazy real-world images from the Internet for a more comprehensive investigation.

2) Baselines: For comprehensive comparisons, we compare the proposed ZID with 15 methods which are divided into three groups, namely, four supervised methods and 11 unsupervised methods. To be specific, the supervised methods contain DehazeNet [1], MSCNN [3], AOD-Net [6] and CAP [5]. Note that CAP needs using the depth information of the corresponding clean image, which is thus classified into supervised methods. Regarding the unsupervised family, seven classical unsupervised methods and four recent proposed zero-shot methods are investigated. In details, the classic unsupervised

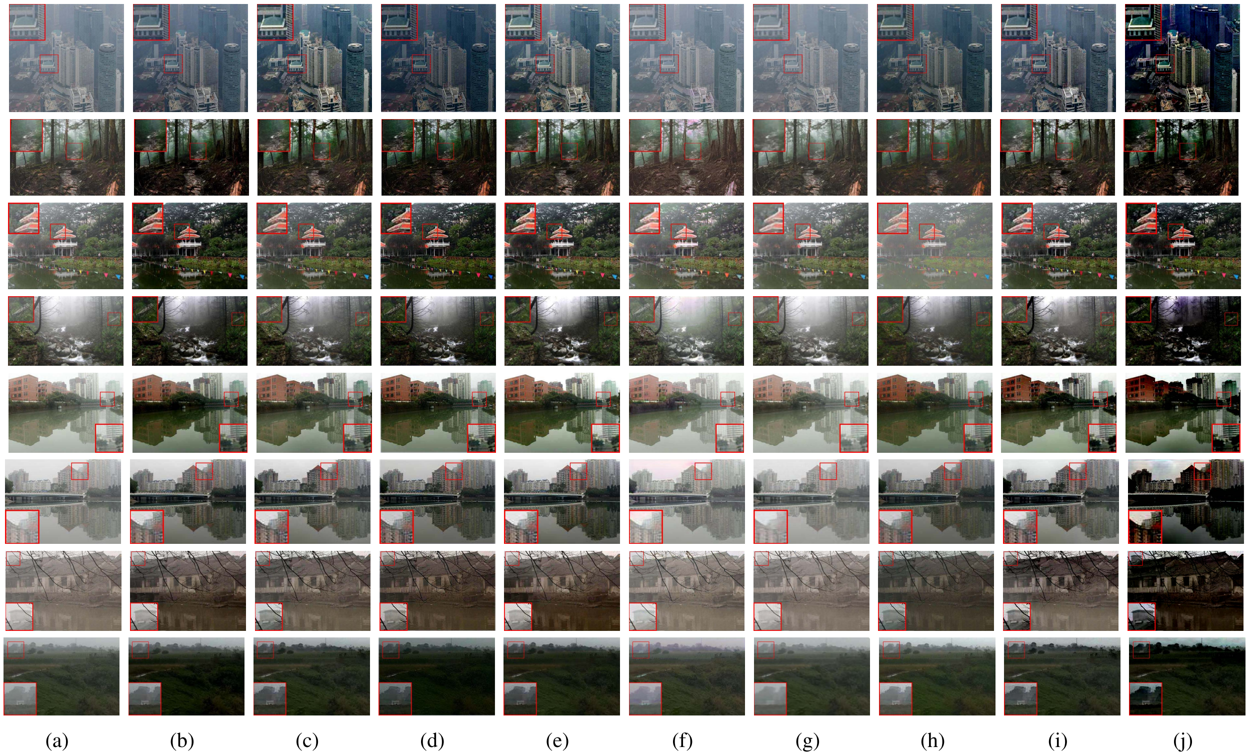


Fig. 6. Comparisons of the SOTA dehazing methods on the Real-World scenes. From the left to the right column (*i.e.*, Figs.(6(a)–6(j))), the input hazy image, DehazeNet [1], MSCNN [3], AOD-Net [6], DCP [2], N2V [31], DIP [2], DCPLoss [16], DDIP [20] and our method are presented. Some areas are highlighted by red rectangles and zooming-in is recommended for a better visualization and comparison.

approaches are DCP [2], FVR [13], BCCR [12], GRM [40], NLD [11], Noise2Noise (N2N) [30] and DCPLoss [16], and the zero-shot methods are Noise2Void (N2V) [31], DIP [32], DeepDecoder (DD) [33], and Double-DIP (DDIP) [20]. Among these unsupervised methods, it should be pointed out that only N2V, DD, DIP, DDIP, and our ZID do not require training data, *i.e.*, they are so-called zero-shot models which only use the given single sample. Note that, the training data partitions of RESIDE are used for the above "trained" methods and the testing partitions are used for inference, whereas the "zero-shot" methods including ours only use the testing partitions.

3) Evaluation Metrics: Like [3], [6], [15], [20], two popular metrics are used in quantitative comparisons, *i.e.*, PSNR and SSIM. Higher value of these metrics, better performance.

4) Experimental Configurations: We conduct experiments on two NVIDIA GeForce RTX 2080Ti GPU in PyTorch. We employ the ADAM optimizer [41] with the default learning rate and the maximal iteration of 500. We set the initial learning rate to 0.001 and do not resize input images. ZID does not use any image augmentation technologies as well. For reproducibility, we do not exhaustively tune parameters for our method. Instead, we simply fix $\lambda_1 = 0.1$ and $\lambda_2 = 0.005$ of Eq.(8) for all the evaluations. To initialize the hint, [2] is used. Regarding some of the baselines, we directly refer to the best result reported in the original works. For the baselines without the corresponding results, we implement experiments by using the source codes provided by the authors and adopting their parameter settings. Our code will be released on Github.

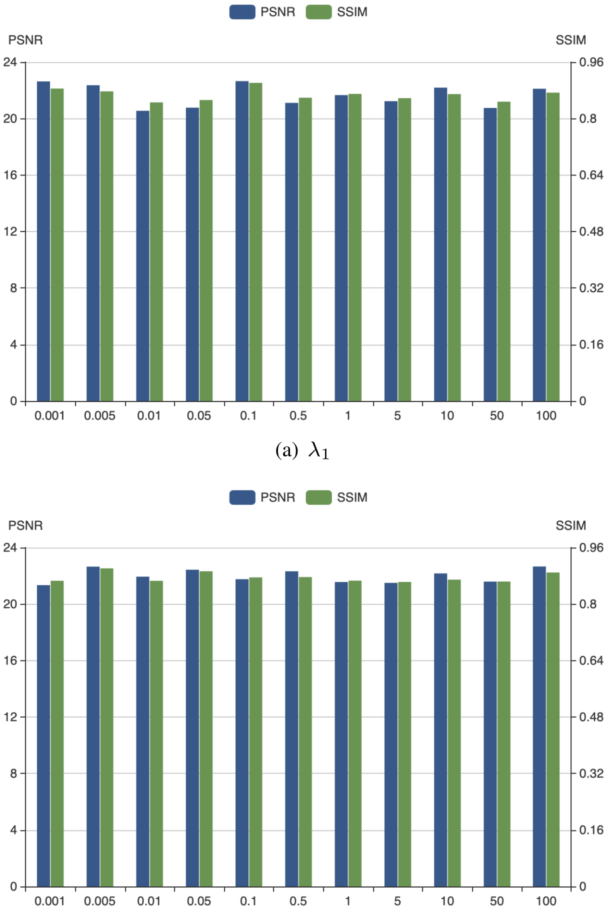
B. Comparisons on Synthetic Datasets

Table I and Fig. 3 report the quantitative and qualitative performance comparisons on the synthetic indoor SOTS dataset. Note that, we do not illustrate the visualization results of FVR, BCCR, GRM, NLD and DD considering the space limitation. From the results, one could have the following observations. First, Table I shows that ZID remarkably outperforms all unsupervised methods. In brief, it is 0.58 and 2.86 higher the best classic unsupervised method (DCPLoss) and the best zero-shot method (DDIP) in PSNR, respectively. Second, Despite the zero-shot and unsupervised characteristics of our ZID, its performance could be superior to all supervised methods excepted DehazeNet in the quantitative comparison. Note that, although ZID achieves a slightly lower PSNR and SSIM compared with DehazeNet, it demonstrates a better visualization result as shown in Fig. 3. For example, relatively, DehazeNet does not well remove the haze in the second image. It indicates the inconsistency between the evaluation metrics and the perceptual quality, as explained in [42]. In brief, the models which excel at minimizing the reconstruction error tend to produce visually unpleasing results, while models that produce results with superior visual quality are rated poorly by evaluation metrics like PSNR and SSIM. Although this problem has been realized by the community of image quality assessment, there are still lacking a better performance metric so far.

TABLE III

ABLATION STUDY ON THE HSTS DATABASE. A-NET (J-NET) DENOTES THAT A-NET ADOPTS THE SAME ARCHITECTURE WITH J-NET

Metrics	w.o. $\mathcal{L}_H$	w.o. $\mathcal{L}_{KL}$	w.o. $\mathcal{L}_A$	$\lambda_1 = 0$	$\lambda_2 = 0$	A-Net (J-Net)	Ours
PSNR	19.93	20.41	20.12	22.24	22.41	18.15	22.65
SSIM	0.8684	0.8724	0.8806	0.8741	0.8920	0.8564	0.9011



(a) λ_1

(b) λ_2

Fig. 7. Influence of parameters on two metrics (PSNR, SSIM). In the experiments, 11 parameter values are investigated to evaluate the influence of λ_1 and λ_2 w.r.t. the metrics, respectively. From the above two figures, one could find that our method is relatively insensitive to the value of parameters.

On the outdoor testing dataset HSTS,

- Table II shows that our method is also superior to most unsupervised baselines in the outdoor scenes. It is worth noting that N2N cannot get corresponding results on HSTS, for N2N requires training on multiple samples from the same scene whereas HSTS only includes a single sample for each scene.
- Although ZID is quantitively worse than the best baseline, it shows better haze-free image recovery performance in the qualitative comparison (see Fig. 4(b) of DehazeNet, Fig. 4(h) of DCPLoss and Fig. 4(j) of ZID). In fact, the haze-free images recovered by our ZID seem more favorite than the ground truth (Fig. 4(k)) because the later might involve haze during data collection.
- The qualitative comparisons show the limitations of the DCPLoss. In brief, DCPLoss tends to recover a darker image. It is mainly because DCPLoss uses the corresponding transmission map of the prior as supervisor to compute empirical loss, while our ZID formulates the statistical properties of dark channel prior into our loss to estimate clean images. As a result, our method could avoid performance degradation.
- On HSTS, ZID takes about 38.14s to handle each image averagely and each iteration only costs 76.28ms. Note that, this is the whole cost of ZID and no training is required in advance.

C. Comparisons on Real World Dataset

To verify the effectiveness of our ZID on real-world hazy images, we conduct qualitative experiments on the HSTS real-world image set and present the results in Fig. 5. It could be seen that our ZID successfully recovers the clean image even though it works in an unsupervised and zero-shot manner. For example, although Double-DIP successfully removes most of the haze in the pictures, it is failed to remove the haze around the people and suffers from the color distortions around the building. In contrast, our method could be immune from these issues.

We also make comparisons on 8 hazy images collected from the Internet by us. As shown in Fig. 6, one could observe that our ZID demonstrates a better visualization result in almost all scenes in the figures. For example, DehazeNet, MSCNN, AOD-Net, DCP, DCPLoss and Double-DIP successfully remove most of the haze in the pictures, but there still contains some haze in the background. In contrast, our method could be immune from these issues and gets the best result.

D. Ablation Study

To demonstrate the effectiveness of our loss function, we conduct an ablation study on the HSTS dataset by removing one of $\mathcal{L}_H$, $\mathcal{L}_{KL}$, and $\mathcal{L}_{Reg}$, where $\lambda_1 = 0$ and $\lambda_2 = 0$ indicate the removal of $\mathcal{L}_{Reg}$ from A-Net and T-Net. To demonstrate the effectiveness of our network structure, we also replace A-Net using an initial hint and adopts a J-Net-like structure as A-Net. From Table III, one could see that: 1) our method benefits from the VAE in haze removal with the formulation of $\mathcal{L}_H$ and $\mathcal{L}_{KL}$; 2) the performance of ZID slightly improved with the regularization on the estimation of atmospheric light and transmission map; 3) the learning process of A-Net improves the quantitive performance than use initial hint directly.

E. Influence of Parameters

Our model requires to specify the value of λ_1 and λ_2 which are trade-off on the regularizations, *i.e.*, $\mathcal{L}_S(f_A(x))$ and $\mathcal{L}_S(f_T(x))$. In this section, we investigate the influence of these parameters to the two metrics (PSNR and SSIM) on the HSTS synthetic dataset. In experiments, we alternatively change the value of λ_1 and λ_2 as indicated in Fig. 7, while accordingly fixing $\lambda_2 = 0.005$ and $\lambda_1 = 0.1$. As shown in the results, one could find that our method is insensitive to the value of λ_1 and λ_2.

V. Conclusion

In this article, we proposed a novel unsupervised and zero-shot single image dehazing method which disentangles a given hazy image into its hazy-free version, transmission map and atmospheric light via three joint subnetworks. In consequence, the model enjoys the interpretability in terms of structure and result. Experimental results on both the synthesis dataset and the real-world dataset demonstrate that the proposed ZID quantitatively outperforms all unsupervised methods and achieves comparable performance with the supervised methods. Besides, it shows a human-favorite result of haze removal in qualitative evaluations. However, ZID has still suffered from some limitations, such as the slow inference speed. In future, we plan to be solve this problem by investigating more efficient network architecture. Besides, it is promising to further improve its performance so that the state of the art could be achieved comparing with supervised deep image dehazing methods.

Acknowledgment

The authors would like to thank an Associate Editor (Prof. Abdesselam Bouzerdoum) and reviewers for the constructive comments and valuable suggestions that greatly improve this work.

References

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

[2] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1956–1963.

[3] G. Tang, L. Zhao, R. Jiang, and X. Zhang, "Single image dehazing via lightweight multi-scale networks," in *Proc. IEEE Int. Conf. Big Data*, Amsterdam, The Netherlands, Dec. 2019, pp. 154–169.

[4] H. Zhu, X. Peng, V. Chandrasekhar, L. Li, and J.-H. Lim, "DehazeGAN: When image dehazing meets differential programming," in *Proc. 37th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 1234–1240.

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

[6] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-net: All-in-One dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4780–4788.

[7] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3194–3203.

[8] W. Ren, J. Pan, H. Zhang, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks with holistic edges," *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 240–259, Sep. 2019.

[9] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Kerkyra, Greece, Sep. 1999, pp. 820–827.

[10] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Deblurring images via dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2315–2328, Oct. 2018.

[11] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1674–1682.

[12] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 617–624.

[13] J.-P. Tarel and N. Hautiere, "Fast visibility restoration from a single color or gray level image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan Sep. 2009, pp. 2201–2208.

[14] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[15] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced Pix2pix dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA USA, Jun. 2019, pp. 8160–8168.

[16] A. Golts, D. Freedman, and M. Elad, "Unsupervised single image dehazing using dark channel prior loss," *IEEE Trans. Image Process.*, vol. 29, pp. 2692–2701, 2020.

[17] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Grid dehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct. 2019, pp. 7314–7323.

[18] R. Liu, P. Mu, J. Chen, X. Fan, and Z. Luo, "Investigating task-driven latent feasibility for nonconvex image modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 7629–7640, 2020.

[19] R. Liu, S. Cheng, Y. He, X. Fan, Z. Lin, and Z. Luo, "On the convergence of learning-based iterative methods for nonconvex inverse problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 3, 2019, doi: 10.1109/TPAMI.2019.2920591.

[20] Y. Gandelsman, A. Shocher, and M. Irani, "Double-DIP: Unsupervised image decomposition via coupled deep-image-priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, p. 11.

[21] J. Park, D. K. Han, and H. Ko, "Fusion of heterogeneous adversarial networks for single image dehazing," *IEEE Trans. Image Process.*, vol. 29, pp. 4721–4732, 2020.

[22] C.-H. Yeh, C.-H. Huang, and L.-W. Kang, "Multi-scale deep residual learning-based single image haze removal via image decomposition," *IEEE Trans. Image Process.*, vol. 29, pp. 3153–3167, 2020.

[23] A. Dudhane and S. Murala, "RYF-net: Deep fusion network for single image haze removal," *IEEE Trans. Image Process.*, vol. 29, pp. 628–640, 2020.

[24] S. E. Kim, T. H. Park, and I. K. Eom, "Fast single image dehazing using saturation based transmission map estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 1985–1998, 2020.

[25] J. Zhang and D. Tao, "FAMED-net: A fast and accurate multi-scale End-to-End dehazing network," *IEEE Trans. Image Process.*, vol. 29, pp. 72–84, 2020.

[26] Q. Liu, X. Gao, L. He, and W. Lu, "Single image dehazing with depth-aware non-local total variation regularization," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5178–5191, Oct. 2018.

[27] A. Wang, W. Wang, J. Liu, and N. Gu, "AIPNet: Image-to-Image single image dehazing with atmospheric illumination prior," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 381–393, Jan. 2019.

[28] Q. Wu, J. Zhang, W. Ren, W. Zuo, and X. Cao, "Accurate transmission estimation for removing haze and noise from a single image," *IEEE Trans. Image Process.*, vol. 29, pp. 2583–2597, 2020.

[29] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2856–2868, Jun. 2018.

[30] J. Lehtinen *et al.*, "Noise2Noise: Learning image restoration without clean data," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 2971–2980.

[31] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2 Void–Learning denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2129–2137.

[32] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 9446–9454.

[33] R. Heckel and P. Hand, "Deep decoder: Concise image representations from untrained non-convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, New Orleans, LA, USA, May 2019, pp. 1–8.

[34] B. Li, Y. Gou, S. Gu, J. Zitao Liu, J. Tianyi Zhou, and X. Peng, "You only look yourself: Unsupervised and untrained single image dehazing neural network," 2020, *arXiv:2006.16829*. [Online]. Available: http://arxiv.org/abs/2006.16829

[35] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8202–8211.

[36] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, Banff, Canada, Apr. 2014, pp. 1–7.

[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines.," in *Proc. Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 807–814.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 448–456.

[39] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[40] C. Chen, M. N. Do, and J. Wang, "Robust image and video dehazing with visual artifact suppression via gradient residual minimization," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 576–591.

[41] J. Kingma and D. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–8.

[42] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," 2018, *arXiv:1809.07517*. [Online]. Available: http://arxiv.org/abs/1809.07517

Boyun Li received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2019, where he is currently pursuing the master's degree in computer science with the College of Computer Science. His research interest includes image restoration.

Yuanbiao Gou received the B.E. degree from the School of Software Engineering, Sichuan University, Chengdu, China, in 2019. He is currently pursuing the M.E. degree with the School of Computer Science, Sichuan University. His current research interest includes image processing.

Jerry Zitao Liu received the Ph.D. degree in computer science from the University of Pittsburgh, under the supervision of Prof. Milos Hauskrecht. He has been leading a team at the TAL Education Group since 2018. His research interests include machine learning, computer vision, natural language processing, and their applications in K-12 education. In these areas, he has published over 40 articles.

Hongyuan Zhu (Member, IEEE) received the B.S. degree in software engineering from the University of Macao, Macau, China, in 2010, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2014. He is currently a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore. His research interests include multimedia content analysis and segmentation.

Joey Tianyi Zhou received the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2015. He is currently a Scientist with the Institute of High Performance Computing, A*STAR, Singapore. His current research interests include differentiable programming, transfer learning, and sparse coding. He was a recipient of the Best Poster Honorable Mention at ACML 2012, the Best Paper Nomination at ECCV 2016, and the NIPS 2017 Best Reviewer Award.

Xi Peng (Member, IEEE) is currently a Full Professor with the College of Computer Science, Sichuan University. His current research interest includes machine intelligence and has authored more than 50 articles in these areas. He has served as an Associate Editor/Guest Editor for six journals, including the IEEE TRANSACTIONS ON SMC: SYSTEMS and the IEEE TRANSACTIONS ON NEURAL NETWORK AND LEARNING SYSTEMS and the Area Chair/Senior Program Committee Member for the conferences such as IJCAI, AAAI, and ICME.