GridDehazeNet+: An Enhanced Multi-Scale Network With Intra-Task Knowledge Transfer for Single Image Dehazing

Xiaohong Liu, Zhihao Shi, *Graduate Student Member, IEEE*, Zijun Wu, Jun Chen, *Senior Member, IEEE*, and Guangtao Zhai, *Member, IEEE*

***Abstract*— Adverse weather conditions such as haze can deteriorate the performance of autonomous driving and intelligent transport systems. As a potential remedy, we propose an enhanced multi-scale network, dubbed GridDehazeNet+, for single image dehazing. The proposed dehazing method does not rely on the Atmosphere Scattering Model (ASM), and an explanation as to why it is not necessarily performing the dimension reduction offered by this model is provided. GridDehazeNet+ consists of three modules: pre-processing, backbone, and post-processing. The trainable pre-processing module can generate learned inputs with better diversity and more pertinent features as compared to those derived inputs produced by hand-selected pre-processing methods. The backbone module implements multi-scale estimation with two major enhancements: 1) a novel grid structure that effectively alleviates the bottleneck issue via dense connections across different scales; 2) a spatial-channel attention block that can facilitate adaptive fusion by consolidating dehazing-relevant features. The post-processing module helps to reduce the artifacts in the final output. Due to domain shift, the model trained on synthetic data may not generalize well on real data. To address this issue, we shape the distribution of synthetic data to match that of real data, and use the resulting translated data to finetune our network. We also propose a novel intra-task knowledge transfer mechanism that can memorize and take advantage of synthetic domain knowledge to assist the learning process on the translated data. Experimental results demonstrate that the proposed method outperforms the state-of-the-art on several synthetic dehazing datasets, and achieves the superior performance on real-world hazy images after finetuning.**

***Index Terms*— Single image dehazing, attention-based feature fusion, intra-task knowledge transfer.**

Manuscript received 29 May 2022; revised 14 August 2022; accepted 22 September 2022. Date of publication 10 October 2022; date of current version 26 January 2023. This work was supported in part by the Shanghai Pujiang Program under Grant 22PJ1406800, in part by the National Natural Science Foundation of China under Grant 62225112 and Grant 61831015, and in part by the Natural Sciences and Engineering Research Council of Canada through a Discovery Grant. The Associate Editor for this article was S. Siri. *(Corresponding authors: Jun Chen; Guangtao Zhai.)*

Xiaohong Liu is with the John Hopcroft Center, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xiaohongliu@sjtu.edu.cn).

Zhihao Shi and Jun Chen are with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada (e-mail: shiz31@mcmaster.ca; chenjun@mcmaster.ca).

Zijun Wu is with the China Telecom Research Institute, Shanghai 200122, China (e-mail: wuzj12@chinatelecom.cn).

Guangtao Zhai is with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhaiguangtao@sjtu.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3210455

I. Introduction

AUTONOMOUS driving systems leverage cameras and sensors as their eyes to see the world. However, in adverse weather conditions, the visibility of autonomous driving systems is dramatically affected, leading to additional difficulties in system control from the analysis of degraded images. Motivated by the fact that haze is one of the leading detrimental factors autonomous driving systems have to deal with [1], [2], [3], and [4], we attempt to address the image dehazing problem in this paper. By restoring clear counterparts from hazy images, dehazing helps mitigate the performance degradation on various down-stream visual tasks such as instance segmentation [5], object detection [6] and object tracking [7], [8], where clear images are generally required as input.

The Atmosphere Scattering Model (ASM) [9] provides a simple approximation of the haze effect. Specifically, it assumes that

$$I_c(x) = J_c(x)t(x) + A(1 - t(x)), \quad c = 1, 2, 3, \tag{1}$$

where $I_c(x)$ ($J_c(x)$) is the intensity of the cth color channel of pixel x in the hazy (clear) image, $t(x)$ is the transmission map, and A is the global atmospheric light intensity. In addition, we have $t(x) = e^{-\beta d(x)}$, where β and $d(x)$ are the atmosphere scattering parameter and the scene depth, respectively. This model indicates that image dehazing is in general an under-determined problem without the knowledge of A and $t(x)$.

As a canonical example of image restoration, the dehazing problem can be tackled using a variety of techniques that are generic in nature. Moreover, many misconceptions and difficulties encountered in image dehazing manifest in other restoration problems as well. Therefore, it is instructive to examine the relevant issues in a broader context, four of which are highlighted below.

1) Role of Physical Model: Many data-driven restoration approaches [13], [14], [15], [16] require synthetic data for training. To produce such data, it is necessary to have a physical model of the relevant image degradation process (*e.g.*, the ASM for the haze effect). A natural question arises whether the design of the image restoration algorithm itself should rely on this physical model. Apparently a model-dependent algorithm may suffer inherent performance loss on real images

due to model mismatch. However, it is often taken for granted that such an algorithm must have advantages on synthetic images produced using the same physical model.

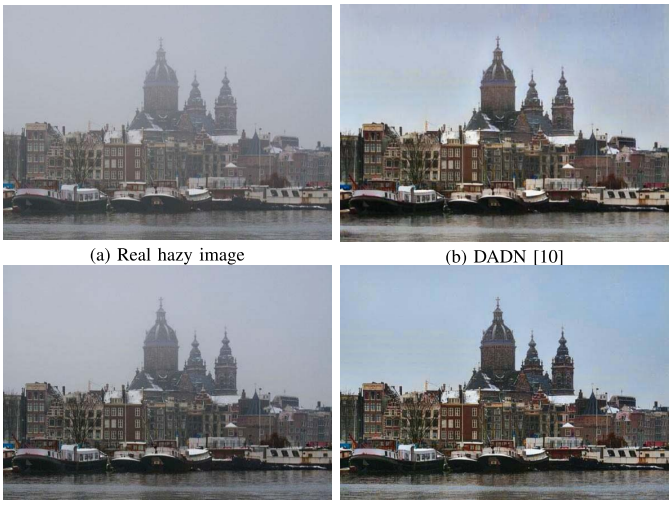
2) Selection of Pre-Processing Method: Pre-processing is widely used in image preparation to facilitate follow-up operations [17]. It can also be used to generate several variants of the given image, providing a certain form of diversity that can be harnessed via proper fusion. However, the pre-processing methods are often selected based on heuristics, thus are not necessarily best suited to the problem under consideration.

3) Bottleneck of Multi-Scale Estimation: Image restoration requires an explicit/implicit knowledge of the statistical relationship between the distorted image and the original clear version. The statistical model needed to capture this relationship often has a huge number of parameters, comparable or even more than the available training data. As such, directly estimating these parameters based on the training data is often unreliable. Multi-scale estimation [18] tackles this problem by i) approximating the high-dimensional statistical model with a low-dimensional one, ii) estimating the parameters of the low-dimensional model based on the training data, iii) parameterizing the neighborhood of the estimated low-dimensional model, performing a refined estimation, and repeating this procedure if needed. It is clear that the estimation accuracy on one scale will affect that on the next scale. Since multi-scale estimation is commonly done in a successive manner, its performance is often limited by a certain bottleneck.

4) Effect of Domain Shift: The effectiveness of supervised learning for image restoration has been widely observed. However, building a large-scale real dataset of distorted images paired with their ground-truth is very expensive and sometimes not even possible [19]. Therefore, in practice one commonly resorts to synthetic data for network training. However, due to domain shift, there is no guarantee that a network trained on synthetic data can generalize well to real data.

The present work can be viewed as a product of our attempt to address the aforementioned generic issues in image restoration. Its main contributions can be summarized as follows:

1) The proposed GridDehazeNet+ (abbreviated as GDN+) does not rely on the ASM for haze removal, yet is capable of outperforming the existing model-dependent dehazing methods even on synthetic images. We also experimentally demonstrate that the dimension reduction offered by the ASM is not necessarily beneficial to network learning, owing to the introduction of undesirable local minima.
2) In contrast to hand-selected pre-processing methods, the pre-processing module in the GDN+ is fully trainable, thus can offer more flexible and pertinent image enhancement.
3) The implementation of attention-based multi-scale estimation on a densely connected grid network allows efficient information exchange across different scales and alleviates the bottleneck issue.
4) To cope with domain shift, certain translated data are generated, by shaping the distribution of synthetic data to match that of real-world hazy images, to finetune our network. Moreover, a novel Intra-Task Knowledge Transfer (ITKT) mechanism is proposed to help the finetuning process on translated data.



(a) Real hazy image (b) DADN [10] (c) ACER-Net [11] (d) GDN+

Fig. 1. Dehazing results for a real hazy image from URHI dataset [12]: (a) a real hazy image, (b) the result based on DADN [10], (c) the result based on ACER-Net [11], and (d) our result. The GDN+ achieves the best visual performance against the others in terms of haze removal and enhanced color contrast.

Benefiting from the overall design, the proposed GDN+ outperforms the State-Of-The-Art (SOTA) methods on several synthetic dehazing datasets and achieves superior performance on real-world hazy images after finetuning. An example is shown in Fig. 1, where our method delivers the most visually appealing dehazing result for a real hazy image from URHI dataset [12].

II. Related Work

Early works on image dehazing either require *multiple* images of the same scene taken under different conditions [20], [21] or side information acquired from other sources [22], [23]. Recent years have seen increasing interest in *single* image dehazing without side information, which is considerably more challenging. To place our work in a proper context, we give a review of existing prior-based and learning-based methods for single image dehazing as well as the recent developments of knowledge distillation and transfer.

A. Prior-Based Single Image Dehazing

A conventional strategy for single image dehazing is to estimate the transmission map $t(x)$ and the global atmospheric light intensity A (or their variants) based on certain assumptions or priors. Then, Eq. (1) is inverted to obtain the dehazed image. Representative works along this line of research include [24], [25], [26], [27], [28]. Specifically, [24] proposed a local contrast maximization method for dehazing, motivated by the observation that clear images tend to have higher contrast as compared to their hazy counterparts. Reference [25] realized haze removal via the analysis of albedo under the assumption that the transmission map and surface shading are locally uncorrelated. Reference [26] proposed the Dark

Channel Prior (DCP), which asserted that pixels in non-haze patches have low intensity in at least one color channel. Reference [27] suggested a machine learning approach that exploits four haze-related features using a random forest regressor. Reference [28] proposed a color attenuation prior that is beneficial to modeling the scene depth of hazy images. Although these methods have enjoyed varying degrees of success, their performances are inherently limited by the accuracy of the adopted assumptions/priors with respect to the target scenes.

B. Learning-Based Single Image Dehazing

With the advance in deep learning techniques and the availability of large synthetic datasets [27], recent years have witnessed the increasing popularity of data-driven methods for image dehazing. These methods largely follow the conventional strategy mentioned above but with reduced reliance on hand-crafted priors. For example, [29] employed a Multi-Scale CNN (MSCNN) that first predicted a holistic transmission map, and refined it locally. Reference [30] proposed a three-layer Convolutional Neural Network (CNN), named DehazeNet, to directly estimate the transmission map from the given hazy image. Reference [31] embedded the ASM into a neural network for joint learning of the transmission map, atmospheric light intensity, and dehazing result. Reference [32] explored the physical model in the feature space (instead of the pixel space) to perform image dehazing.

The AOD-Net [33] represents a departure from the conventional strategy. Specifically, a reformulation of Eq. (1) was introduced in [33] to bypass the estimation of the transmission map and atmospheric light intensity. A close inspection reveals that this reformulation in fact renders the ASM completely superfluous (though this point is not recognized in [33]). In [17], the proposed Gated Fusion Network (GFN) went one step further by explicitly abandoning the ASM in its design, and leveraged several hand-selected pre-processing methods (*i.e.*, white balance, contrast enhancing, and gamma correction) to improve the dehazing results. Recent works mostly followed this model-agnostic design principle and tackled image dehazing with various techniques. By regarding image dehazing as image-to-image translation, [34] constructed an Enhanced Pix2pix Dehazing Network (EPDN) based on the Generative Adversarial Network (GAN), which does not rely on any physical model. Reference [35] capitalized on the attention mechanism and put forward a feature fusion attention network with the flexibility to regulate different types of information. By leveraging the boosting strategy, [36] proposed a boosted decoder that can progressively restore the haze-free image. Reference [11] treated hazy and clear images as negative and positive samples to train the proposed AECR-Net jointly, and the adopted contrastive regularization can be applied to other dehazing methods to further improve their performance.

While there is increasing evidence that model-agnostic image dehazing methods are able to outperform their model-dependent counterparts even if only synthetic data (produced using the physical model) are concerned, the reason behind this puzzling phenomenon is still unclear. In this paper, we attempt to lift the veil by providing a possible explanation together with some supporting experiments.

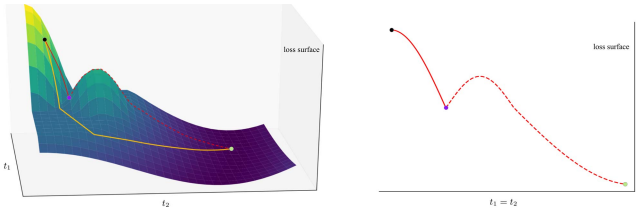
In addition, owing to domain shift, learning-based methods trained on synthetic data tend to generalize poorly over to real data. To mitigate the detrimental effect caused by domain shift, [37] proposed a hybrid approach, where a CNN is trained on synthetic data in a supervised manner, and on real data in an unsupervised manner. To support unsupervised learning, physical priors (*i.e.*, dark channel loss and total variation loss) were employed. Reference [38] followed this line of ideas and proposed a principled synthetic-to-real dehazing framework to finetune a model trained on synthetic data, aiming at improving the generalization performance on real data. However, involving real data in training does not fully address the domain-shift issue. In [10], a Domain Adaptation Dehazing Network (DADN) was proposed by adopting the CycleGAN [19] to deal with the discrepancies between the synthetic domain and real domain.

In view of the fact that unsupervised finetuning guided by physical priors may cause significant artifacts, the GDN+ proposed in the present paper exploits supervised finetuning on translated data to improve the dehazing performance on real data.

C. Knowledge Distillation and Transfer

One popular application of knowledge distillation [39] is for network compression, where the learned logits from a large network (*i.e.,* teacher network) is transferred to a small network (*i.e.,* student network). Compared to the teacher network, the student network is much easier to deploy, possibly at the cost of a potential performance drop. Reference [40] suggested that the intermediate representations from the teacher network can be leveraged to further improve the training process of the student network. In recent years, knowledge distillation has been proved useful not only for network compression, but also for various computer vision tasks, including object detection [41], semantic segmentation [42], image synthesis [43], style transfer [44], *etc.* Knowledge distillation found its first application to single image dehazing in [45], where the teacher and student networks share the same architecture but are responsible for image reconstruction and image dehazing tasks, respectively. In contrast, for the Knowledge Distilling Dehazing Network (KDDN) proposed in [46], the architectures of teacher and student networks are tailored to the designated tasks; besides, multiple features, rather than only one intermediate feature, are distilled to improve the effectiveness of knowledge transfer.

Different from [45] and [46], where knowledge transfer is carried out among heterogeneous tasks, we perform ITKT with teacher and student networks working on the *same* task (*i.e*, dehazing) but taking *different* data as inputs. Intuitively, the synthetic domain knowledge yields useful insights into translated data, where the haze effect does not admit a simple mathematical characterization. Therefore, the characteristics of intermediate features distilled from the teacher network can greatly benefit the learning process of the student network,



(a) Loss surface (b) Constrained loss surface

Fig. 2. On the potential detrimental effect of using the ASM for image dehazing. For illustration purposes, we focus on two color channels of a single pixel and denote the respective transmission maps by t_1 and t_2. Fig. 2 (a) plots the loss surface as a function of t_1 and t_2. It can be seen that the global minimum is attained at a point (see the green dot) satisfying $t_1 = t_2$, which agrees with the ASM. With the black dot as the starting point, one can readily find this global minimum using gradient descent (see the yellow path). However, a restricted search based on the ASM along the $t_1 = t_2$ direction (see the red path) will get stuck at a point indicated by the purple dot (see Fig. 2 (b)). Note that this point is a local minimum in the constrained space but not in the original space, and it becomes an obstruction simply due to the adoption of the ASM.

enabling it to deliver satisfactory dehazing results on real-world hazy images.

III. Method

A. Overview

Here we highlight the following aspects of the proposed GDN+.

1) No Reliance on Atmosphere Scattering Model: Although the model-agnostic approach to single image dehazing has become increasingly popular, no convincing reason has been provided why there is any advantage in ignoring the ASM, as far as the dehazing performance on synthetic images is concerned. The argument put forward in [17] is that estimating $t(x)$ from a hazy image is an ill-posed problem. Nevertheless, this is puzzling since estimating $t(x)$ (which is color-channel-independent) is presumably easier than $J_c(x)$, $c = 1, 2, 3$. In Fig. 2, we offer a possible explanation why it could be problematic if one blindly uses the constraint that $t(x)$ is color-channel-independent to narrow down the search space and why it might be potentially advantageous to relax this constraint in the search of the optimal $t(x)$. However, with this relaxation, the ASM offers no dimension reduction in the estimation procedure. More fundamentally, it is known that the loss surface of a CNN is generally well-behaved in the sense that the local minima are often almost as good as the global minimum [47], [48], [49]. On the other hand, by incorporating the ASM into a CNN, one basically introduces a nonlinear component that is heterogeneous in nature from the rest of the network, which may create an undesirable loss surface. To support this explanation, we provide some experimental results in Section V-E.

2) Trainable Pre-Processing Module: The pre-processing module effectively converts the single image dehazing problem to a multi-image dehazing problem by generating several variants of the given hazy image, each highlighting a different aspect of this image and making the relevant feature information more evidently exposed. In contrast to those hand-selected pre-processing methods adopted in the existing works (*e.g.*, [17]), the proposed pre-processing module is made fully trainable, which is in line with the general preference of data-driven methods over prior-based methods as shown by recent developments in image dehazing. Note that hand-selected processing methods typically aim to enhance certain concrete features that are visually recognizable. However, the exclusion of abstract features is not justifiable. Indeed, there might exist abstract transform domains that better suit the follow-up operations than the image domain. A trainable pre-processing module has the freedom to identify transform domains over which more diversity gain can be harnessed.

3) Enhanced Multi-Scale Estimation: Here the meaning of word *enhanced* is two-fold. First, inspired by [50], we *enhance* the conventional multi-scale network using a novel grid structure. This grid structure has clear advantages over the encoder-decoder structure and the conventional multi-scale structure extensively used in image restoration [17], [51], [52], [53]. In particular, the information flow in the encoder-decoder structure or the conventional multi-scale structure often suffers from the bottleneck effect due to the hierarchical architecture whereas the grid structure circumvents this issue via dense connections across different scales using up-sampling/down-sampling blocks. Second, we further *enhance* the network with Spatial-Channel Attention Blocks (SCABs) that are placed at the junctions where features are exchanged and aggregated. These SCABs enable the network to better exploit the diversity created by the pre-processing module and the information most relevant to the dehazing task.

4) Intra-Task Knowledge Transfer: ITKT refers to leveraging the knowledge acquired from a certain task on *one dataset* to facilitate the learning process of the same task on *another dataset*. In the current context, it is observed that the synthetic domain knowledge is beneficial for handling translated data. Rather than directly finetuning the network on translated data, a teacher-student structure is used to memorize and take advantage of synthetic domain knowledge. To the best of our knowledge, this is the first work that leverages ITKT to improve the dehazing performance on real-world hazy images.

In comparison to the preliminary work GDN [54], the GDN+ is improved in two aspects. First, the GDN only adopts channel-wise attention with the learned weights independent of the target features [54]. In contrast, the GDN+ employs the self-attention mechanism [55], [56], encapsulated in SCABs, to generate feature-adaptive weights. Second, the GDN tends to suffer significant performance drop on real-world hazy images, possibly due to the domain shift between synthetic data in training and real data in testing. To address this issue, we shape the distribution of synthetic data to match that of real data, and use the resulting translated data to finetune the network. In order to memorize and take advantage of synthetic domain knowledge, we propose an ITKT mechanism to assist the learning process on translated data. In addition, more comprehensive performance evaluations are conducted as compared to those in [54]. Specifically, We test more benchmarks with full comparisons to SOTAs, perform the task-driven evaluation, conduct thorough ablation studies, and show the failure cases of our method.

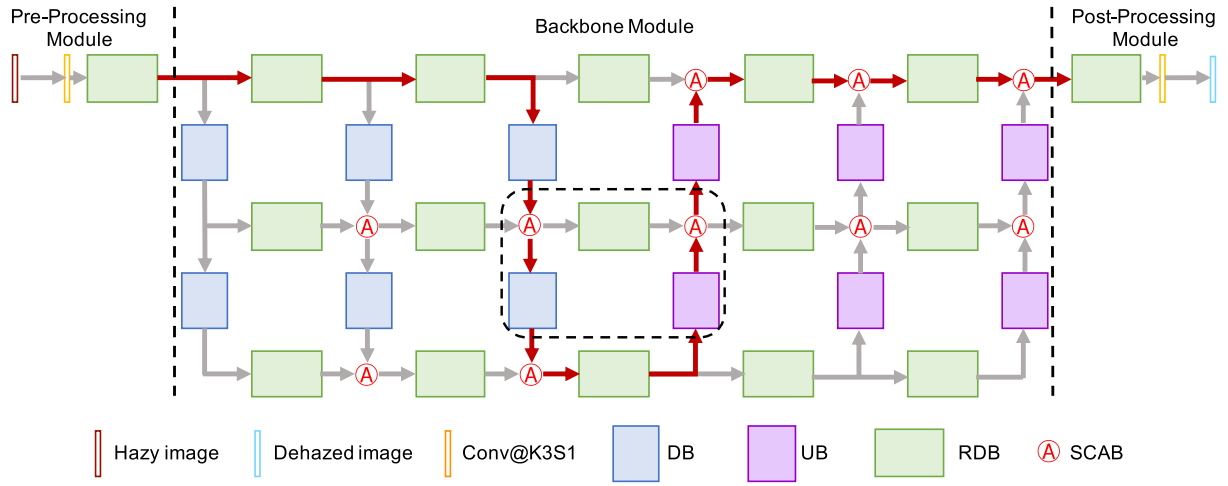


Fig. 3. Architecture of the proposed GridDehazeNet+ (GDN+). Here Conv@KnSm indicates a $n \times n$ convolution with stride m.

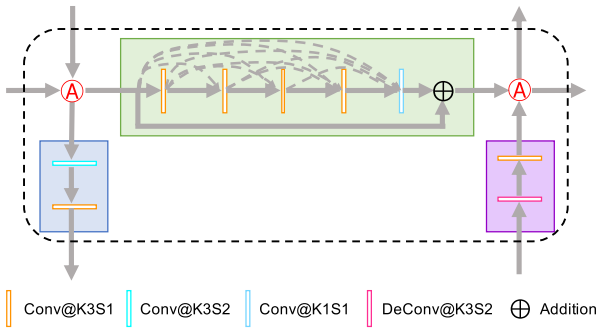


Fig. 4. Illustration of the dashed box in Fig. 3. Here Conv(DeConv)@KnSm indicates a $n \times n$ convolution (deconvolution) with stride m.

B. Network Architecture

The GDN+ consists of three modules, *i.e*, the pre-processing module, the backbone module and the post-processing module. Fig. 3 shows the overall architecture of the proposed network.

The pre-processing module consists of a 3×3 convolution with stride 1 (denoted as Conv@K3S1) and a Residual Dense Block (RDB) [52]. It generates 16 feature maps, which will be referred to as the learned inputs, from the given hazy image.

The backbone module is an improved version of GridNet [50] originally proposed for semantic segmentation. It performs enhanced multi-scale estimation based on the learned inputs. We choose a grid structure with three rows and six columns. Each row corresponds to a different scale and consists of five RDBs with the number of feature maps unaltered. Each column can be regarded as a bridge that connects different scales via Upsampling Blocks (UBs) or Downsampling Blocks (DBs). In each UB (DB), the size of feature maps is increased (decreased) by a factor of 2 while the number of feature maps is decreased (increased) by the same factor. Here upsampling/downsampling is implemented using convolution instead of traditional methods such as bilinear or bicubic interpolation. Fig. 4 provides a detailed illustration of the RDB, UB, and DB in the dash box in Fig. 3. Each RDB consists of five convolutions: the first four are used to increase the number of feature maps while the last one fuses these feature maps. The output is then combined with the input of this RDB via channel-wise addition. Following [52], the growth rate in RDB is set to 16. The UB and DB are structurally the same except that they respectively use Convolution (Conv) and DeConvolution (DeConv)) to adjust the size of feature maps. In GDN+, except for the first convolution in the pre-processing module and the 1×1 convolution in each RDB, all other convolutions are activated by ReLU. To strike a balance between the output size and the computational complexity, we set the number of feature maps at three different scales to 16, 32, and 64, respectively.

Since dehazed images constructed directly from the output of the backbone module tend to contain artifacts, we introduce a post-processing module to further improve the quality. The structure of the post-processing module is symmetrical to that of the pre-processing module.

It is worth noting that the GDN+ subsumes some existing networks as special cases. For example, the red path in Fig. 3 shows an encoder-decoder network that can be obtained by pruning the GDN+. As another example, removing the exchange branches (*i.e*., the middle four columns in the backbone module) from GDN+ leads to a conventional multi-scale network.

C. Feature Fusion With Spatial-Channel Attention Blocks

Since the appearance of haze in real world is usually nonhomogeneous and different channels of learned features may not be of the same importance for the dehazing process, we embed certain judiciously constructed SCABs into the network to enable adaptive feature fusion. The SCAB employs spatial and channel-wise attentions [56], realized respectively by the Spatial Attention Block (SAB) to deal with distinct haze effects at different positions of an image, and the Channel Attention Block (CAB) to perform importance-aware exploitation of feature maps. The SAB applies the average and max poolings along the channel axis to aggregate the local information on different feature maps, and the two pooled results are concatenated and fed into a convolution to generate

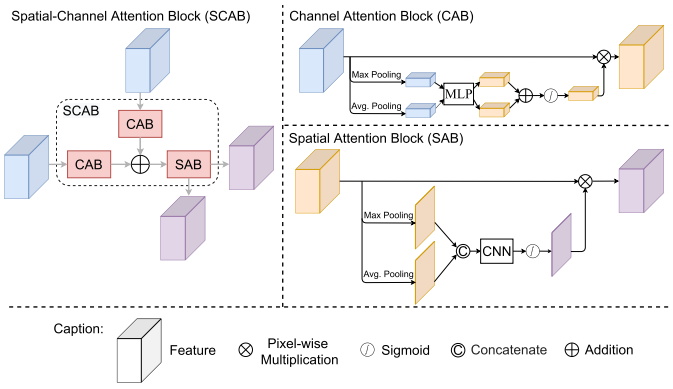


Fig. 5. Illustration of the spatial-channel attention block (SCAB).

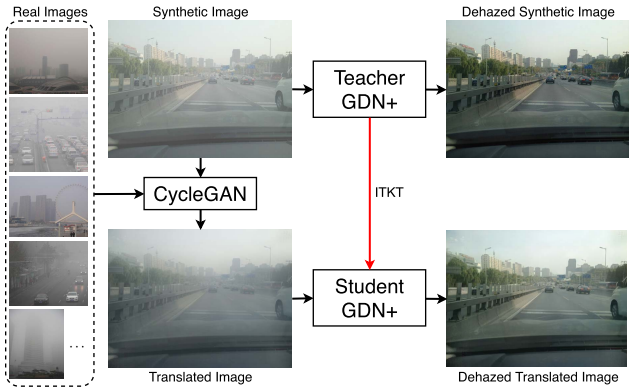


Fig. 6. Flowchart of the proposed ITKT mechanism.

Fig. 7. Visualization of the haze effect before (shown in the 1st row) and after (shown in the 2rd row) the translation.

the spatial attention map. The CAB applies the average and max poolings along the spatial axis instead; the pooled features are adjusted by a shared multi-layer perceptron which explores the inter-channel relationship to consolidate the important information; the adjusted versions are then added together and passed through a Sigmoid function to produce the channel attention coefficients. Finally, the spatial attention map and channel attention coefficients act back on the corresponding input features to enable self-adaptation.

As illustrated in Fig. 5, each SCAB consists of two CABs and one SAB. The features from horizontal and vertical streams are first accommodated by two distinct CABs to strengthen the relevant characteristics via channel-wise attention. The outputs of the two CABs are added together and then fed into a SAB for spatial adaptation. Let $F^h_{i,j}$ and $F^v_{i,j}$ denote respectively the features from the horizontal stream and vertical stream at the fusion position (i, j) in the backbone module, where $i = 0, 1, 2$ and $j = 0, 1, \ldots, 5$. Let $f^h_{i,j}(F \mid \Theta^h_{i,j})$ and $f^v_{i,j}(F \mid \Theta^v_{i,j})$ denote respectively the CAB operations for the horizontal stream and vertical stream at the fusion position (i, j), where F represents an arbitrary input feature, and $\Theta^h_{i,j}$, $\Theta^v_{i,j}$ are the trainable weights. Similarly, let $g_{i,j}(F \mid \Phi_{i,j})$ denote the SAB operation at the fusion position (i, j), where $\Phi_{i,j}$ is the trainable weight. The proposed SCAB can be expressed as

$$\tilde{F}_{i,j} = g_{i,j}(f^c_{i,j}(F^c_{i,j} \mid \Theta^c_{i,j}) + f^r_{i,j}(F^r_{i,j} \mid \Theta^r_{i,j}) \mid \Phi_{i,j}), \quad (2)$$

where $\tilde{F}_{i,j}$ is the output feature of the SCAB. Note that SCABs endow the GDN+ with the ability to fuse features from different scales adaptively. Quite remarkably, our experimental results indicate that it suffices to use SCABs with a small number of trainable weights to substantially boost the overall performance.

D. Intra-Task Knowledge Transfer

We use the CycleGAN [19] to convert ASM-based synthetic data to more realistic-looking translated data, which can be regarded as samples from the distribution of real-world hazy images. As the real haze effect captured by translated data does not admit a simple mathematical characterization, the learning process on translated data is more difficult than that on synthetic data. Therefore, to memorize and take advantage of synthetic domain knowledge, we propose an ITKT mechanism to reduce the finetuning difficulty on translated data. The overall flowchart of the proposed ITKT mechanism is demonstrated in Fig. 6. The teacher GDN+ is pre-trained on synthetic data, and its learned weights are utilized to initialize the student GDN+. During the finetuning process, the teacher GDN+ is responsible for memorizing and providing the synthetic domain knowledge to the student GDN+, thus its weights are fixed. The student GDN+, equipped with this knowledge, is finetuned on translated data in a supervised manner to improve the dehazing performance on real-world hazy images. Note that the teacher and student networks have the freedom to adopt their own architectures as long as the synthetic domain knowledge is properly transferred.

As shown in Fig. 6 and Fig. 7, the haze effect of synthetic images is noticeably different from that of translated ones, which is a clear indicator of domain shift. Benefiting from ITKT, the performance drop on real-world hazy images is significantly alleviated. In Sec. V-H, we also evaluate the effectiveness of ITKT by directly finetuning the GDN+ on translated data. Our experimental results show that the dehazing performance deteriorates as a consequence of this change.

E. Loss Function

In total, three different loss functions are employed to train the proposed network: 1) the fidelity loss L_F, 2) the perceptual loss L_P, and 3) the intra-task knowledge transfer loss L_{KT}. Their definitions and the underlying rationale are detailed below.

1) Fidelity Loss: The commonly used fidelity losses include L_1 and MSE. The MSE loss is very sensitive to outliers,

thus might suffer from gradient explosion [57]. Although the L_1 loss does not have this issue, it is not differentiable at zero. The smooth L_1 loss can be regarded as an integration of these two losses, thus inherits their merits and avoids their drawbacks. Therefore, we use it as our fidelity loss to quantitatively measure the difference between the dehazed image and the ground-truth.

Let $\hat{J}_c(x)$ denote the intensity of the cth color channel of the pixel x in the dehazed image, and N denote the total number of pixels in one channel. Our fidelity loss can be expressed as

$$L_F = \frac{1}{3N}\sum_{c=1}^{3}\sum_{x=1}^{N} h(\hat{J}_c(x) - J_c(x)), \tag{3}$$

where

$$h(e) = \begin{cases} 0.5e^2, & \text{if } |e| < 1, \\ |e| - 0.5, & \text{otherwise.} \end{cases} \tag{4}$$

2) Perceptual Loss: As a complement to the pixel-level fidelity loss, the perceptual loss [58] leverages multi-scale features extracted from a pre-trained deep neural network to quantify the overall *perceptual* difference between the dehazed image and the ground-truth. In this work, we use the VGG16 [59] pre-trained on ImageNet [60] as our loss network and extract the features from the last layer of each of the first three stages (*i.e.*, Conv1-2, Conv2-2 and Conv3-3). The perceptual loss can be defined as

$$L_P = \frac{1}{3}\sum_{l=1}^{3} \frac{1}{C_l H_l W_l} ||\phi_l(\hat{J}) - \phi_l(J)||_2^2, \tag{5}$$

where $\phi_l(\hat{J})$ ($\phi_l(J)$), $l = 1, 2, 3$, denote the aforementioned three VGG16 feature maps associated with the dehazed image $\hat{J}$ (the ground truth J), and C_l, H_l, and W_l specify the dimension of $\phi_l(\hat{J})$ ($\phi_l(J)$).

3) Intra-Task Knowledge Transfer Loss: To effectively transfer the synthetic domain knowledge, we design an ITKT loss that guides the features from the student network to mimic the ones from the teacher network by reducing their L_1 distance. Three intermediate features from the first scale of the backbone module after the SCAB-based fusion are selected. According to our experiments, this selection induces the best dehazing performance among the candidates that have been considered. Following the notation in Sec. III-C, we denote these features by $\tilde{F}_{0,3}$, $\tilde{F}_{0,4}$, $\tilde{F}_{0,5}$, and use the superscripts t and s to indicate whether they come from the teacher or student network. Our ITKT loss can be expressed as

$$L_{KT} = \frac{1}{3}\sum_{j=3}^{5} ||\tilde{F}_{0,j}^{s} - \tilde{F}_{0,j}^{t}||_1. \tag{6}$$

4) The Overall Loss: The overall loss L_S of our GDN+ is a linear combination of fidelity loss L_F, perceptual loss L_P, and ITKT loss L_{KL}, which can be formulated as

$$L_S = L_F + \lambda_P L_P + \lambda_{KL} L_{KL}, \tag{7}$$

where λ_P and λ_{KL} are used to balance the loss components. According to our experiments, they are set to 0.04 and 0.01 respectively.

TABLE I

STATISTICAL DISTANCES OF SYNTHETIC AND TRANSLATED IMAGES TO REAL-WORLD HAZY IMAGES IN TERMS OF KULLBACK–LEIBLER DIVERGENCE

Distance	KLD
Syn↔Real	0.1397
Tran↔Real	**0.0387**

F. Training Dataset

The RESIDE [12] is a large-scale dataset that contains an Indoor Training Set (ITS), an Outdoor Training Set (OTS), a Synthetic Object Testing Set (SOTS), a set of Unannotated real Hazy Images (URHI), and a real Task-driven Testing Set (RTTS). The ITS and OTS are generated from clear images based on the ASM via proper choices of the scattering coefficient β and the atmospheric light intensity A. Following DADN [10], we use the exactly same dataset that consists of 6,000 images with 3,000 from ITS and 3,000 from OTS to train our GDN+. Since different dehazing methods may originally adopt different training datasets (*e.g.*, AOD-Net [33] was trained using 27,256 synthetic hazy images while ACER-Net [11] was trained on ITS that only has 13,990 images), for fair comparisons, we laboriously retrain all the methods under consideration on the aforementioned dataset by following their respective training strategies.

To finetune the GDN+, we select 1,000 real-world hazy images from RTTS, and utilize the CycleGAN to convert 6,000 synthetic images to translated ones with the distribution matched to that of real-world hazy images. Note that these 6,000 translated images should not be considered as additionally introduced data since they are generated from the training data *per se*. Fig. 7 visualizes the haze effect before and after this translation. In addition, we adopt the Kullback–Leibler Divergence (KLD) to measure the statistical distance between translated and real-world hazy images (denoted as Tran↔Real), as well as that between synthetic and real-world ones (denoted as Syn↔Real), where the real-world hazy images are from URHI. The corresponding results are shown in Tab. I. Since the lower KLD value stands for the higher similarity between two distributions, it is clear that the distribution of synthetic images has been shaped to better approximate that of real-world hazy images after this translation, resulting in a more realistic appearance.

IV. DATA PREPARATION

A. Testing Dataset

For testing, in total 6 dehazing datasets are used. Four of them are synthetic datasets and the rest two are real datasets. These testing datasets differ in size and haze distribution. We elaborate them as follows:

- **SOTS** [12] is an ASM-based dataset that comprises 500 indoor hazy images and 500 outdoor hazy images roughly of size 620×460.
- **Middlebury** [61] is an ASM-based dataset that consists of 23 indoor hazy images roughly of size $2,880 \times 1,988$.

TABLE II

QUANTITATIVE EVALUATIONS ON FOUR DEHAZING BENCHMARKS. FOR EACH METHOD, AVERAGE PSNR/SSIM VALUES ARE REPORTED. **RED** AND BLUE INDICATE THE BEST AND THE SECOND BEST PERFORMANCE. THE NUMBER OF PARAMETERS AND RUNTIME ARE ALSO PROVIDED

Method	SOTS		Middlebury		HazeRD		O-HAZE		# Param.(M)	Runtime (s)
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
DCP [26]	15.49	0.64	15.91	0.81	14.01	0.39	15.80	0.33	-	166.18
MSCNN [29]	17.57	0.81	13.35	0.78	15.57	0.42	19.60	0.63	0.008	2.71
DehazeNet [30]	21.14	0.85	13.28	0.79	15.54	0.41	19.37	0.64	0.008	3.70
AOD-Net [33]	19.06	0.85	13.86	0.79	15.63	0.45	18.85	0.70	**0.002**	0.46
GFN [17]	22.30	0.88	14.38	0.81	13.98	0.37	22.83	0.77	0.212	0.84
EPDN [34]	23.82	0.89	15.11	0.83	17.37	0.56	19.82	0.75	17.380	0.32
KDDN [46]	30.09	0.97	17.27	0.87	16.44	0.83	25.45	0.78	2.405	0.37
DADN [10]	27.76	0.93	15.93	0.70	18.07	0.63	21.69	0.74	54.591	1.13
ACER-Net [11]	31.61	**0.98**	17.48	0.86	15.88	0.81	25.87	0.84	2.614	0.35
GDN [54]	27.75	0.96	14.46	0.83	14.51	0.79	23.51	0.83	0.958	**0.28**
GDN+	**32.15**	**0.98**	**18.04**	**0.88**	**19.54**	**0.87**	**26.10**	**0.85**	0.961	0.30

- **HazeRD** [62] is an ASM-based dataset that contains 75 outdoor hazy images roughly of size $3,873 \times 2,516$.
- **O-HAZE** [63] has 45 outdoor hazy images roughly of size $5,456 \times 3,632$. Instead of relying on ASM for synthesizing the haze effect, hazy images are produced by a professional haze machine and consequently more realistic. Since the haze distribution of this dataset is different from that of ASM-based datasets, following the testing protocol of previous dehazing works, we adopt the training/testing splits in [64] to train and test the GDN+ and other methods chosen for comparison.
- **37Real** [65] collects 37 real-world hazy images roughly of size 768×512. This is a commonly used benchmark for testing the performance of dehazing methods in the real world.
- **URHI** [12] contains $4,809$ real-world hazy images of various sizes (ranging from 400×350 to $2,000 \times 1,000$).

Unless otherwise specified, the *pre-trained* GDN+ is tested on synthetic datasets to demonstrate the superiority of our network design while the *finetuned* GDN+ is tested on real datasets to verify the mitigation of domain gap attributed to ITKT. For simplicity, we do not explicitly differentiate them since it is easy to tell the difference based on the testing datasets.

V. EXPERIMENTAL RESULTS

We conduct extensive experiments to demonstrate that the proposed GDN+ outperforms the SOTA methods on synthetic datasets and delivers visually more satisfactory results on real datasets after finetuning. The experiments also provide useful insights into the constituent modules of the GDN+ and solid justifications for the effectiveness of the proposed ITKT mechanism.

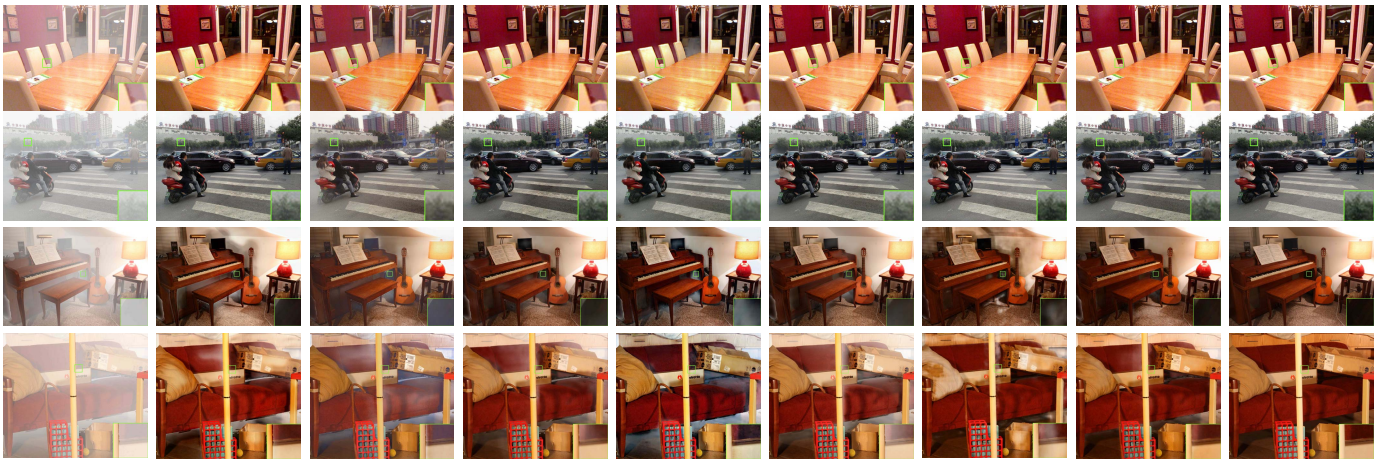
A. Experimental Setup

The GDN+ is first trained on synthetic data for 100 epochs and then finetuned on translated data for another 100 epochs. We randomly crop a patch of size 240×240 from each image. For training, the Adam optimizer is adopted, where β_1 and β_2 take the default values of 0.9 and 0.999, respectively. The batch size is set to 16 with the initial learning rate $1e$-3 that will be reduced by half every 20 epochs. The training is carried out on a PC with two NVIDIA GeForce GTX 2080Ti, but only one GPU is used for testing.

We compare the proposed GDN+ with 10 methods including DCP [26], MSCNN [29], DehazeNet [30], AOD-Net [33], GFN [17], EPDN [34], KDDN [46], DADN [10], ACER-Net [11], and GDN [54], where the DCP is the only non-learning-based method. Although ACER-Net is the current SOTA, DADN achieves better visual quality on real-world hazy images. Therefore, we consider both of them as the representatives of existing dehazing methods. For quantitative comparisons, we leverage the Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) to evaluate the dehazing results of different methods on synthetic datasets. Since the ground-truth of real-world hazy images are not available in real datasets, the Fog Aware Density Evaluator (FADE) [66], a no-reference image quality assessment tool specifically designed for image dehazing task, is used as an alternative to support quantitative evaluations.

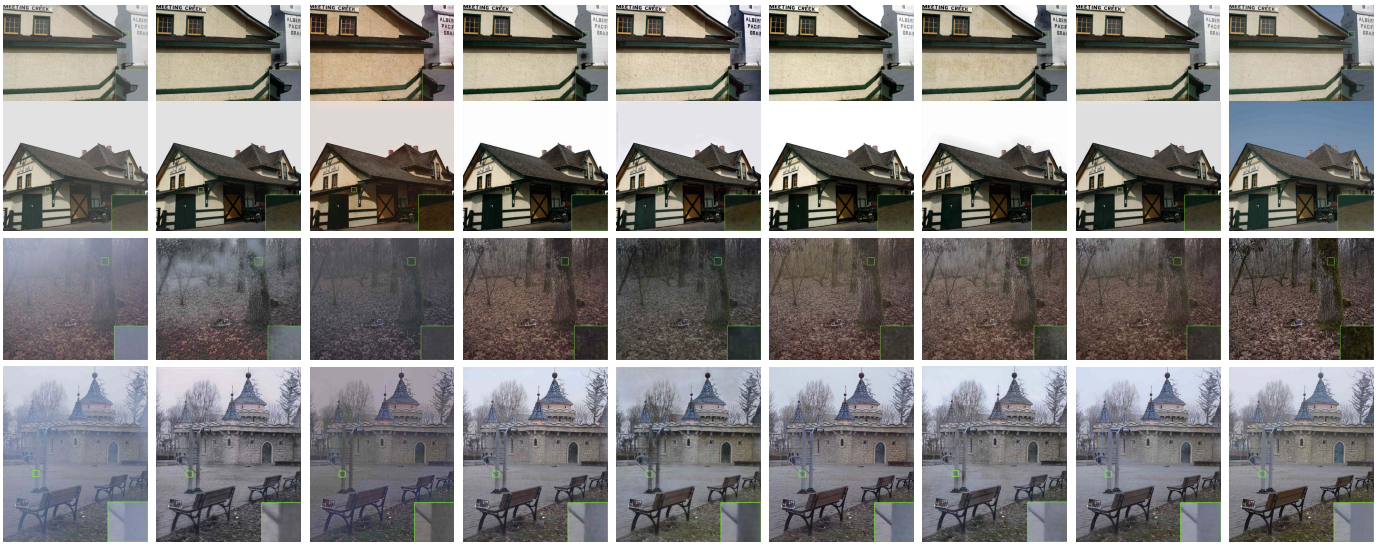
B. Evaluation on Synthetic Data

We conduct evaluations on the 4 synthetic datasets, *i.e.*, SOTS, Middlebury, HazeRD, and O-HAZE. Comparisons in terms of average PSNR/SSIM values can be found in Tab II. It is evident that the proposed GDN+ outperforms all other methods chosen for comparison, and has a significant improvement over its preliminary version GDN (*e.g.*, 4.4 dB on SOTS). Besides, for each method, we also demonstrate the number of trainable parameters in million (M), and the runtime in second (s) on a $1,080$P fake dataset, where all pixel values are set to 1. Except for DCP that only works on CPU, the runtime of all other methods are tested on GPU. The proposed GDN+ has much fewer parameters than ACER-Net and DADN, and our un-optimized code takes about $0.3s$ to process one $1,080$P image, faster than DADN and ACER-Net. The comparison between GDN+ and GDN reveals that



(a) Hazy inputs (b) GFN (c) EPDN (d) KDDN (e) DADN (f) ACER-Net (g) GDN (h) GDN+ (i) Ground truth

Fig. 8. Visual comparisons on SOTS (the first two rows) and Middlebury (the last two rows). Zoom in for details.



(a) Hazy inputs (b) GFN (c) EPDN (d) KDDN (e) DADN (f) ACER-Net (g) GDN (h) GDN+ (i) Ground truth

Fig. 9. Visual comparisons on HazeRD (the first two rows) and O-HAZE (the last two rows). Zoom in for details.

the adoption of self-attentions greatly improves the dehazing performance with a negligible impact on the model size (*i.e.*, +0.003M) and runtime (*i.e.*, +0.02s).

We demonstrate the visual comparisons on SOTS and Middlebury in Fig. 8, and HazeRD and O-Haze in Fig. 9. GFN and EPDN exhibit limited performance on dense haze removal (*e.g.*, the 3rd row in Figs. 9 (b-c)). KDDN, ACER-Net, and GDN still retain a non-negligible amount of haze in some cases (*e.g.*, the 4th row in Figs. 8 (d, f, g)). DADN tends to cause color distortions (*e.g.*, the 1st row in Fig. 9 (e)). In comparison, the dehazing results of GDN+ are visually most similar to the ground-truth as they are free of color distortion and contain very little residual haze.

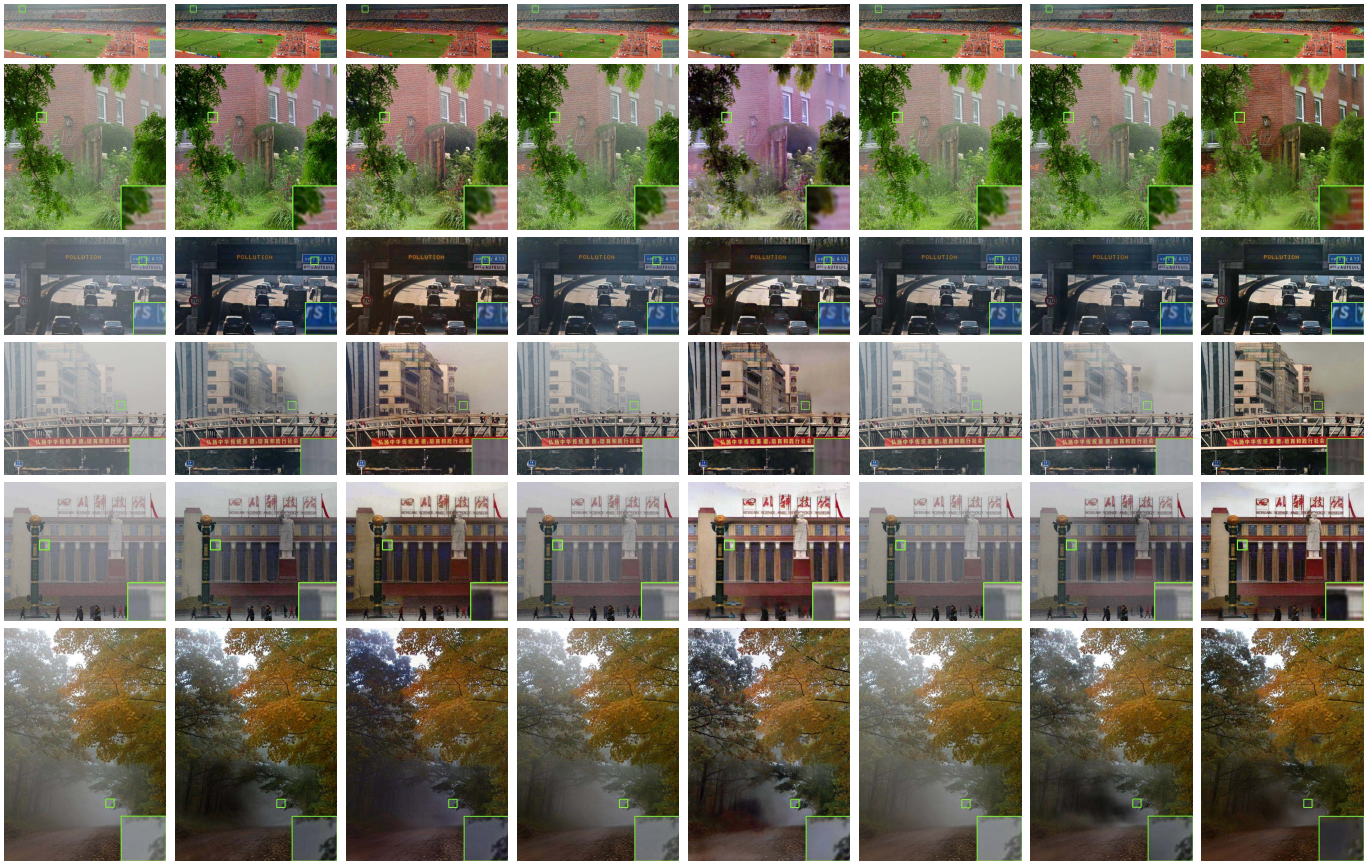
TABLE III

QUANTITATIVE EVALUATIONS ON TWO REAL DATASETS USING THE FADE METRIC. THE LOWER VALUE INDICATES THE BETTER DEHAZING PERFORMANCE

Method	37Real	URHI
Hazy	1.0037	2.3230
DADN	0.5788	1.0561
ACER-Net	0.6961	1.5927
GDN	0.6290	1.4852
GDN+	**0.5184**	**0.9094**

C. Evaluation on Real Data

In Fig. 10, we perform visual comparisons between GDN+ and other methods on 2 real datasets, *i.e.*, 37Real and URHI. Except for DADN, the dehazing performance of other methods under comparison deteriorates significantly due to domain shift. The performance drop of DADN is the least among them. However, similar to the situation on synthetic data, severe color distortion may occur, which seriously compromises the visual quality of its dehazed images. In comparison, the proposed GDN+ removes haze more thoroughly and is free of color distortion. Moreover, the objective assessment is conducted by leveraging FADE metric, where we quantitatively compare the GDN+ with the SOTAs and our precedent work GDN.



(a) Hazy inputs (b) GFN (c) EPDN (d) KDDN (e) DADN (f) ACER-Net (g) GDN (h) GDN+

Fig. 10. Visual comparisons of different methods on real-world hazy images: the ones in the first 2 rows are from 37Real, and the rest are from URHI. Zoom in for details.

The results are shown in Tab. III, where the lower FADE value indicates the better dehazing performance. Unless otherwise specified, the best results for all tables in this paper are highlighted in **bold**. It is evident that GDN+ surpasses the SOTAs on FADE, and the dehazing performance on real data is significantly improved as compared to GDN. This improvement can be attributed to the proposed ITKT mechanism that successfully alleviates domain shift between synthetic data and real data.

D. Task-Driven Evaluation

Since image dehazing methods can be utilized as a pre-processing step to alleviate the performance degradation of high-level tasks such as object detection or recognition in the presence of adverse weather conditions, the improvement of detection accuracy resulted from dehazing has been regraded as an indicator to evaluate the dehazing methods [33], [67]. We leverage YOLOv3 [68] to detect image objects. For the test dataset, although RTTS provides hazy images with annotated object classes and relevant bounding boxes, annotation defects occur in this dataset where plenty of objects that should have been annotated are left out. This flaw results in a severe consequence that true-positive detections are considered as false-positive ones when the ground-truth label is missing. To address this issue, we select 100 real-world hazy images that have full annotations from RTTS for task-driven evaluation, and these images do not overlap with the ones used to generate translated data. It should be mentioned that we do not finetune image dehazing and YOLOv3 jointly. Instead, image dehazing methods only serve as a pre-processing step to remove haze in input images.

TABLE IV
TASK-DRIVEN EVALUATION. THE mAP SCORES ARE REPORTED IN %

Method	mAP (%)	Gain
Hazy	67.1	-
DADN	66.9	−0.2
ACER-Net	67.4	0.3
GDN	67.9	0.8
GDN+	**71.0**	**2.9**

Tab. IV shows mean Average Precision (mAP) scores on the dehazed images for different methods. Although DADN visually performs better than ACER-Net and GDN on real-world hazy images, it surprisingly reports the lowest mAP, even worse than direct detection on hazy images. One possible reason is that DADN introduces noise during dehazing, which is the culprit for detection degradation. For ACER-Net, the detection gain on its dehazed images is marginal. In comparison, the dehazed images from our GDN+ are most beneficial to YOLOv3 with a gain of 2.9%. Besides, owing to the novel SCAB and domain shift mitigation, GDN+ surpasses our

(a) Hazy image (b) DADN (c) ACER-Net (d) GDN (e) GDN+

Fig. 11. Task-Driven Evaluation. YOLOv3 is leveraged to detect objects on real-world hazy images and the dehazing results from different methods.

TABLE V
QUANTITATIVE COMPARISONS OF DIFFERENT ESTIMATION STRATEGIES

Variant	SOTS		HazeRD	
	PSNR	SSIM	PSNR	SSIM
Indirect	31.82	0.9732	17.81	0.8455
Direct	**32.15**	**0.9777**	**19.54**	**0.8697**

TABLE VI
QUANTITATIVE COMPARISONS OF DIFFERENT INPUT TYPES

Variant	SOTS		HazeRD	
	PSNR	SSIM	PSNR	SSIM
Original	30.04	0.9697	17.69	0.8486
Derived	29.08	0.9635	18.04	0.8580
Learned	**32.15**	**0.9777**	**19.54**	**0.8697**

precedent work GDN by 2.1%. We also demonstrate two sets of object detection results in Fig. 11. It can be seen that the dehazed image from DADN does suffer from the noise that is not originally present, which matches our analysis above. Due to the superior dehazing performance of GDN+, YOLOv3 can now detect the objects that are not detectable from hazy images (see the 2nd row in Fig. 11).

E. Necessity of Atmosphere Scattering Model

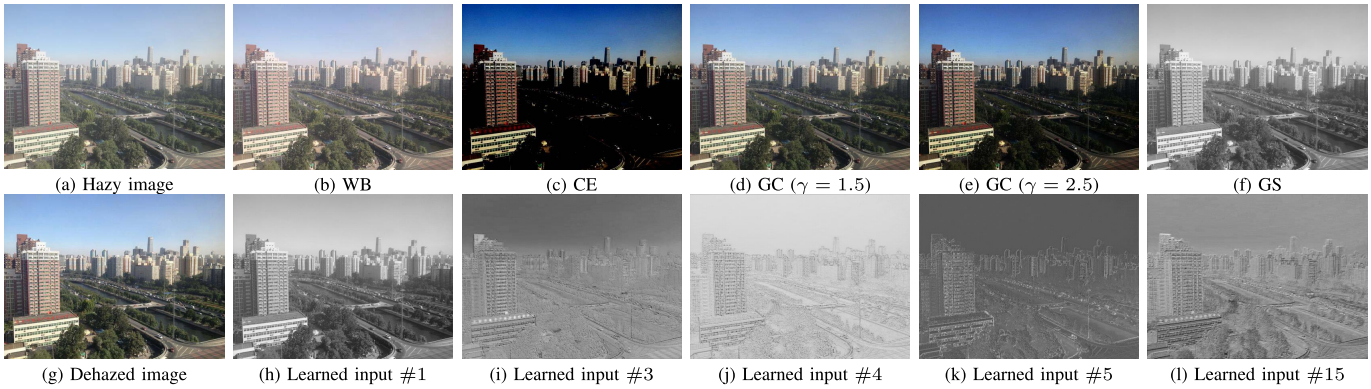
To gain a better understanding of the difference between the adopted direct estimation strategy where the ASM is completely bypassed (denoted as *Direct*), and the indirect estimation strategy where the transmission map and the atmospheric light intensity are first estimated in order to calculate the dehazing result (denoted as *Indirect*), we adjust the GDN+ to make it follow the indirect estimation strategy instead. Specifically, we modify the convolution at the output end (*i.e.*, the rightmost Conv@K3S1 in Fig. 3) so that it outputs two feature maps rather than three. The first feature map is used as the estimated transmission map while the mean of the second one serves as the estimated atmospheric light intensity. These two estimates are then substituted into Eq. (1) to calculate the dehazing result. This variant of GDN+ is trained in the same way as detailed in Sec. V-A, and evaluated on SOTS and HazeRD. It is worth noting that both SOTS and HazeRD are synthetic datasets based on ASM. Therefore, as far as this kind of testing datasets are concerned, the indirect estimation strategy essentially takes advantage of the ASM as a perfect prior. However, as shown in Tab. V, although adopting the ASM leads to a significant reduction in the number of parameters to be estimated, it in fact incurs performance degradation. This indicates that incorporating the ASM into GDN+ does have a detrimental effect on the loss surface.

F. Utility of Learned Inputs

The pre-processing module of GDN+ produces 16 learned inputs in total. Here we build two variants of GDN+ to demonstrate the diversity gain offered by these learned inputs. For the first variant (denoted as *Original*), we remove the pre-processing module and replace the first 3 learned inputs by the RGB channels of the given RGB hazy image and the rest by all-zero feature maps. For the second variant (denoted as *Derived*), the learned inputs are substituted with the same number of derived inputs generated by hand-selected pre-processing methods. More specifically, we generate 16 derived inputs, 3 from the given hazy image, 3 from the White Balanced (WB) image, 3 from the Contrast Enhanced (CE) image, 6 from two Gamma Corrected (GC) images with γ set to 1.5 and 2.5 respectively, and 1 from the Gray-Scaled (GS) image. Fig. 12 shows the derived and learned inputs of a hazy image.

Although the hand-selected pre-processing methods can create diversified inputs, our pre-processing module is considerably more flexible and adaptive in finetuning the given image to better suit the follow-up process (*e.g.,* the learned inputs #3 and #5 enhance different aspects of the given hazy image and are complement to each other). More interestingly, the learned input #1 resembles a GS image, even though this is not prescribed. This shows that our pre-processing module is capable of mimicking hand-selected pre-processing methods when it is beneficial to do so.

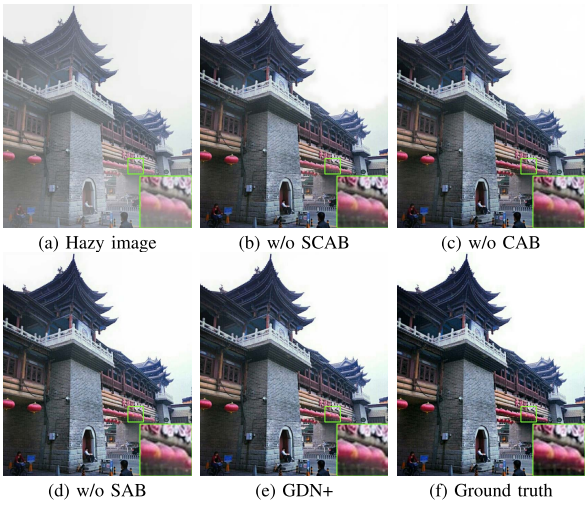
To further validate the effectiveness of learned inputs, we follow the same experimental setup to train both variants,



(a) Hazy image (b) WB (c) CE (d) GC ($\gamma = 1.5$) (e) GC ($\gamma = 2.5$) (f) GS

(g) Dehazed image (h) Learned input #1 (i) Learned input #3 (j) Learned input #4 (k) Learned input #5 (l) Learned input #15

Fig. 12. Visualization of the derived and learned inputs for a hazy image from SOTS.



(a) Hazy image (b) w/o SCAB (c) w/o CAB

(d) w/o SAB (e) GDN+ (f) Ground truth

Fig. 13. Dehazing results of attention variants on a hazy image from SOTS.

(a) Hazy image (b) Pre-trained

(c) w/o ITKT (d) w/ ITKT

Fig. 14. Qualitative comparisons of ITKT-related variants on real hazy images from URHI.

and quantitatively evaluate their dehazing performance on the SOTS and HazeRD. Tab. VI shows that the GDN+ with learned inputs (denoted as *Learned*) outperforms the *Original* and *Derived* versions in terms of PSNR and SSIM metrics. Note that the *Derived* version performs worse than the *Original* version on SOTS. A possible explanation is that the generated derived inputs may fail to compensate the given hazy image, but rather hinder the haze removal in some cases.

G. Validation of Overall Design

The proposed GDN+ is a multi-scale network that is enhanced in two aspects: 1) a grid structure with dense connections across different scales to facilitate the information exchange, and 2) a novel SCAB that is capable of fusing features based on their relative importance. To demonstrate the effectiveness of the adopted grid structure, we consider the following two variants: 1) the Encoder-Decoder Network (*EDNet*) obtained by pruning the GDN+ (see the red path in Fig. 3), and 2) the conventional multi-scale network (*MSNet*) that removes all exchange branches except for the first and the last ones in order to maintain the minimum connection. To validate the proposed SCAB, we consider the following three variants: 1) the GDN+ without SCABs (*w/o SCAB*), 2) the GDN+ with CAB-absent SCABs (*w/o CAB*), and 3) the GDN+ with SAB-absent SCABs (*w/o SAB*). Moreover,

TABLE VII

QUANTITATIVE COMPARISONS OF DIFFERENT VARIANTS OF GDN+

Variant	SOTS		HazeRD	
	PSNR	SSIM	PSNR	SSIM
EDNet	25.98	0.9446	15.86	0.8182
MSNet	27.65	0.9467	16.51	0.8208
w/o SCAB	27.85	0.9657	15.91	0.8264
w/o CAB	29.31	0.9734	16.12	0.8295
w/o SAB	31.80	0.9759	19.05	0.8648
w/o post-processing	31.79	0.9755	19.28	0.8669
w/o perceptual loss	31.82	0.9760	19.32	0.8676
GDN+	**32.15**	**0.9777**	**19.54**	**0.8697**

to validate the efficacy of our post-processing module and the adopted perceptual loss, we build two other variants, one without the post-processing module (*w/o post-processing*) and the other without using the perceptual loss (*w/o perceptual loss*). All these variants are trained in the same way as before and are tested on the SOTS and HazeRD.

The quantitative comparisons are shown in Table VII. Compared to *EDNet* and *MSNet*, the proposed GDN+ achieves favorable dehazing results owing to the superiority of the grid structure. Besides, it can be seen that the variants *w/o SAB* and *w/o CAB* both outperform the baseline *w/o SCAB* though the performance gain from CAB appears to be more significant

(a) Hazy image (b) ACER-Net (c) DADN (d) GDN (e) GDN+

Fig. 15. Limitations. Our GDN+ may fail while dealing with extremely dense haze in distant scenes. Besides, the GDN+ might amplify the shot noise in the hazy images.

TABLE VIII
QUANTITATIVE COMPARISONS OF ITKT-RELATED VARIANTS ON SOTS-T

Variant	SOTS-T	
	PSNR	SSIM
pre-trained	21.08	0.8004
w/o ITKT	23.70	0.8606
w/ ITKT	**24.66**	**0.8751**

than that from SAB. Benefiting from the contributions of both CAB and SAB, the GDN+ with SCABs delivers further elevated performance. We also visualize the respective advantages of spatial and channel-wise attentions in Fig. 13. As compared to GDN+, the variant *w/o post-processing* is inferior owing to the potential residual artifacts from the backbone module, and the variant *w/o perceptual loss* has a degraded performance that validates the benefit of supervising perceptual difference. The above results provide a fairly comprehensive justification for the overall design of GDN+.

H. Effectiveness of Intra-Task Knowledge Transfer

To convincingly demonstrate the effectiveness of the proposed ITKT mechanism (denoted as *w/ ITKT*), we consider a variant that trains the GDN+ directly on translated data (denoted as *w/o ITKT*). We also convert the original SOTS to a translated version, named SOTS-T, for quantitative comparisons. Besides, the GDN+ pre-trained on synthetic data is also tested on SOTS-T (denoted as *pre-trained*).

As shown in Tab. VIII, higher PSNR and SSIM values are achieved while the ITKT mechanism is adopted. This validates the synthetic domain knowledge can benefit the learning process of translated data. Moreover, from Figs. 14(c-d), *w/ ITKT* removes haze more thoroughly than *w/o ITKT*, and produces more appealing dehazing results. As for *pre-trained*, although it works well on synthetic data, the dehazing performance on real data is rather limited as shown in Fig. 14 (b). This dramatic performance drop is owing to the domain shift between training and testing data. Therefore, it is necessary to conduct training on real data or those with (approximately) the same distribution. This is exactly the rationale of creating and utilizing the translated data to finetune the GDN+.

It is worth emphasizing that the proposed ITKT is generic in nature and can be easily employed in other learning-based dehazing methods to improve their performance on real-world hazy images.

I. Limitations

Fig. 15 demonstrates the limitations of the proposed GDN+. In the first row, we show a real-world hazy image that suffers from extreme dense haze. Our GDN+ performs favorably against SOTAs in the nearby areas where the haze is relatively light, but it cannot fully remove the haze in distant scenes (*e.g.,* see the buildings in this image). Besides, the dehazing task could become more intractable when the image is not only severely degraded by extremely dense haze but also contaminated by shot noise. For instance, as shown in the second row of Fig. 15, after employing GDN+, the shot noise is greatly amplified and more perceptible. This problem is not only encountered in the proposed method, but also exists in the dehazing results of DADN. The problem of joint dehazing and denoising is beyond the scope of the present paper and will be treated in our future work.

VI. CONCLUSION

We have proposed an enhanced multi-scale network and demonstrated its competitive performance for single image dehazing. The design of this network involves several ideas. We adopt a densely connected grid structure to facilitate the information exchange across different scales. A Novel SCAB, constructed based on the idea of self-attentions, is placed at the junctions of the grid structure to enable adaptive feature fusion. The issue of domain shift is addressed by converting synthetic data to translated data with the distribution matched to that of real-world hazy images. We further propose a novel ITKT mechanism that leverages the synthetic domain knowledge to assist the learning process on translated data.

Due to the generic nature of its building components, the proposed network is expected to be applicable to a wide range of image restoration problems. Investigating such applications is an endeavor well worth undertaking.

Our work also sheds some light on the puzzling phenomenon regarding the use of the ASM in image dehazing, and suggests the need to rethink the role of physical models in the design of image restoration algorithms.

REFERENCES

[1] Z. Cao *et al.*, "Haze removal of railway monitoring images using multi-scale residual network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7460–7473, Dec. 2021.

[2] A. Mehra, M. Mandal, P. Narang, and V. Chamola, "ReViewNet: A fast and resource optimized network for enabling safe autonomous driving in hazy weather conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4256–4266, Jul. 2021.

[3] G. Kim and J. Kwon, "Deep illumination-aware dehazing with low-light and detail enhancement," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2494–2508, Mar. 2022.

[4] P. Liu, C. Zhang, H. Qi, G. Wang, and H. Zheng, "Multi-attention DenseNet: A scattering medium imaging optimization framework for visual data pre-processing of autonomous driving systems," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 1, 2022, doi: 10.1109/TITS.2022.3145815.

[5] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9157–9166.

[6] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[7] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Adaptive correlation filters with long-term and short-term memory for object tracking," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 771–796, Aug. 2018.

[8] Q. Wang, Y. Zheng, P. Pan, and Y. Xu, "Multiple object tracking with correlation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3876–3886.

[9] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, 2002.

[10] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain adaptation for image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2808–2817.

[11] H. Wu *et al.*, "Contrastive learning for compact single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10551–10560.

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

[13] X. Liu, L. Kong, Y. Zhou, J. Zhao, and J. Chen, "End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2416–2425.

[14] X. Liu, K. Shi, Z. Wang, and J. Chen, "Exploit camera raw data for video super-resolution via hidden Markov model inference," *IEEE Trans. Image Process.*, vol. 30, pp. 2127–2140, 2021.

[15] Z. Shi *et al.*, "Learning for unconstrained space-time video super-resolution," *IEEE Trans. Broadcast.*, vol. 68, no. 2, pp. 345–358, Jun. 2022.

[16] Z. Shi, X. Xu, X. Liu, J. Chen, and M.-H. Yang, "Video frame interpolation transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17482–17491.

[17] W. Ren *et al.*, "Gated fusion network for single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.

[18] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8260–8269.

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[20] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 713–724, Jun. 2003.

[21] S. Shwartz, E. Namer, and Y. Y. Schechner, "Blind haze separation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2006, pp. 1984–1991.

[22] S. G. Narasimhan and S. K. Nayar, "Interactive (de)weathering of an image using physical models," in *Proc. IEEE Workshop Color Photometric Methods Comput. Vis.*, Paris, France, vol. 6, Oct. 2003, pp. 1–8.

[23] J. Kopf *et al.*, *Deep Photo: Model-Based Photograph Enhancement and Viewing*, vol. 27. New York, NY, USA: ACM, 2008.

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

[25] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, p. 72, Aug. 2008.

[26] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[27] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2995–3000.

[28] 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.

[29] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 154–169.

[30] 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.

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

[32] J. Dong and J. Pan, "Physics-based feature dehazing networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 188–204.

[33] 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)*, Oct. 2017, pp. 4770–4778.

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

[35] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11908–11915.

[36] H. Dong *et al.*, "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2157–2167.

[37] L. Li *et al.*, "Semi-supervised image dehazing," *IEEE Trans. Image Process.*, vol. 29, pp. 2766–2779, 2019.

[38] Z. Chen, Y. Wang, Y. Yang, and D. Liu, "PSD: Principled synthetic-to-real dehazing guided by physical priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7180–7189.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[40] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.

[41] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4933–4942.

[42] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 346–362.

[43] H. Yin *et al.*, "Dreaming to distill: Data-free knowledge transfer via DeepInversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8715–8724.

[44] X. Chen, Y. Zhang, Y. Wang, H. Shu, C. Xu, and C. Xu, "Optical flow distillation: Towards efficient and stable video style transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 614–630.

[45] H. Wu, J. Liu, Y. Xie, Y. Qu, and L. Ma, "Knowledge transfer dehazing network for NonHomogeneous dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 478–479.

[46] M. Hong, Y. Xie, C. Li, and Y. Qu, "Distilling image dehazing with heterogeneous task imitation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3462–3471.

[47] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proc. Artif. Intell. Statist.*, 2015, pp. 192–204.

[48] F. Draxler, K. Veschgini, M. Salmhofer, and F. A. Hamprecht, "Essentially no barriers in neural network energy landscape," 2018, *arXiv:1803.00885*.

[49] Q. Nguyen and M. Hein, "The loss surface and expressivity of deep convolutional neural networks," 2018, *arXiv:1710.10928*.
[50] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," 2017, *arXiv:1707.07958*.
[51] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2502–2510.
[52] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
[53] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8174–8182.
[54] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.
[55] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.
[56] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[57] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[58] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 694–711.
[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[60] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[61] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer, "D-HAZY: A dataset to evaluate quantitatively dehazing algorithms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2226–2230.
[62] Y. Zhang, L. Ding, and G. Sharma, "HazeRD: An outdoor scene dataset and benchmark for single image dehazing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3205–3209.
[63] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 754–762.
[64] C. Ancuti, C. O. Ancuti, and R. Timofte, "NTIRE 2018 challenge on image dehazing: Methods and results," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW)*, Jun. 2018, pp. 891–901.
[65] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, vol. 34, no. 1, p. 13, Dec. 2014.
[66] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3888–3901, Nov. 2015.
[67] W. Yang *et al.*, "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020.
[68] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

Xiaohong Liu received the B.E. degree in communication engineering from Southwest Jiaotong University, China, in 2014, the M.A.Sc. degree in electrical and computer engineering from the University of Ottawa, Canada, in 2016, and the Ph.D. degree in electrical and computer engineering from McMaster University, Canada, in 2021. He is a Tenure-Track Assistant Professor with the John Hopcroft Center, Shanghai Jiao Tong University. His research interests include video super-resolution/interpolation, image dehazing/deraining, and image forgery detection.

He received the Ontario Graduate Scholarship in 2019, the NSERC Alexander Graham Bell Canada Graduate Scholarship-Doctoral in 2020, the Borealis AI Global Fellowship Award in 2020, and the Chinese Government Award for Outstanding Self-Financed Students Abroad in 2021. He serves as a Reviewer for several IEEE journals, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Zhihao Shi (Graduate Student Member, IEEE) received the B.E. degree in communication engineering from Zhengzhou University, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, McMaster University, Canada. His research interests include image dehazing/deraining, video super-resolution, and other low-level computer vision problems.

Zijun Wu received the B.S. degree from North China Electric Power University, China, in 2018, and the M.A.Sc. degree from McMaster University, Canada, in 2021. She is currently a Researcher with China Telecom. Her research interests include image and video processing.

Jun Chen (Senior Member, IEEE) received the B.E. degree in communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2004 and 2006, respectively.

He was a Post-Doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL, USA, from September 2005 to July 2006; and a Post-Doctoral Fellow at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, from July 2006 to August 2007. Since September 2007, he has been with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he is currently a Professor. His research interests include information theory, machine learning, wireless communications, and signal processing.

Dr. Chen was a recipient of the Josef Raviv Memorial Postdoctoral Fellowship in 2006, the Early Researcher Award from the Province of Ontario in 2010, the IBM Faculty Award in 2010, the ICC Best Paper Award in 2020, and the JSPS Invitational Fellowship in 2021. He held the title of the Barber–Gennum Chair in information technology from 2008 to 2013 and the Joseph Ip Distinguished Engineering Fellow from 2016 to 2018. He served as the Editor for the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING from 2020 to 2021. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY.

Guangtao Zhai (Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow, from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen–Nuremberg, Germany. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. His research interests include multimedia signal processing and perceptual signal processing. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012.