# TransER: Hybrid Model and Ensemble-based Sequential Learning for Non-homogenous Dehazing

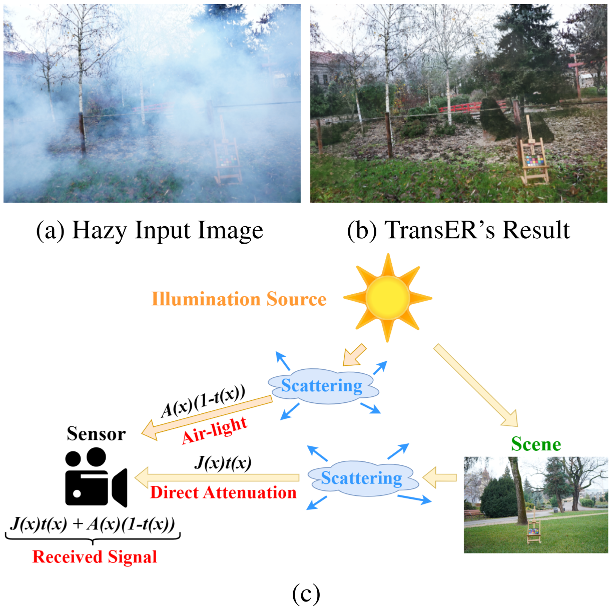
Trung Hoang, Haichuan Zhang, Amirsaeed Yazdani, Vishal Monga
School of Electrical Engineering and Computer Science
Pennsylvania State University, Unversity Park, USA
{tvh5567, haichuan, auy200, vum4}@psu.edu

## Abstract

*Image dehazing is one of the most challenging imaging inverse problems that estimates the haze-free images from hazy ones. While recent transformer/convolutional neural network-based methods have shown excellent performance in handling both homogeneous and non-homogeneous dehazing problems, these networks are often trained end-to-end to estimate the haze-free image directly and require a large number of parameters. In this work, we propose a novel, lightweight two-stage deep network for non-homogeneous dehazing. In particular, our proposed method, denoted as* ***TransER****, consists of two separate deep neural networks which are* ***Trans****Conv Fusion Dehaze (TFD) model in Stage I and Lightweight* ***E****nsemble* ***R****econstruction (LER) network in Stage II. The first model (TFD) using transformer-based encoder and decoders generates two estimates of the haze-free image: a parameter-based dehazed output based on the physical modeling of the problem and a pseudo haze-free output generated directly by the model in an end-to-end fashion. LER in stage II reconstructs the final dehazed output fusing the two estimates from stage I. We incorporate knowledge distillation to develop a teacher network with the same architecture as LER, allowing it to supervise the intermediate features. Extensive experiments performed on challenging real and synthetic scene image datasets (NTIRE 2019-2023, and RESIDE-indoor) demonstrate that TransER can outperform many state-of-the-art competing methods while using a significantly lower number of parameters. The source code is available at https://github.com/trungpsu1210/TransER.*

## 1. Introduction

The restoration of hazy images, also known as image dehazing, is a crucial topic in low-level vision, as haze is a common natural phenomenon caused by floating particles in the atmosphere that scatter or absorb light. The presence of haze in digital images can obscure or partially block objects, leading to color and texture distortion, which can negatively impact high-level vision tasks such as image classification, segmentation, or object detection. Therefore, restoring hazy images has become a challenging ill-posed problem that has received increasing attention in the computer vision community.


(a) Hazy Input Image (b) TransER's Result

(c)

Figure 1. (a) The hazy input image of NH-Haze 2023 test set [9]. Our proposed method (TransER) is able to produce haze-free image with high perceptual quality (b). (c) is the formation of hazy image in atmospheric model [35].

Image dehazing aims to restore the clear/haze-free image from the observed hazy image. This problem has been actively researched over the past two decades [2,4–6,20,21,36,44,45,53], in which the majority of work can be divided into two categories: single image dehazing and multi-image dehazing. However, multi-image dehazing requires multiple images of the same scene under different weather or environmental conditions, which may not always be available [28]. Thus single image dehazing has gained popularity [13]. For the single image dehazing problem, there is a physical haze model introduced by [35], which is illustrated in Figure 1(c) and can be described by the equation:

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

where $I$ is the captured hazy image, $J$ is the true scene radiance, $A$ is the ambient light intensity, and $t$ is the transmission map. We use $A(x)$ to describe haze with varying density for the non-homogeneous dehazing task. The transmission map can be mathematically expressed as $t(x) = e^{-\beta d(x)}$, where $\beta$ represents the attenuation coefficient of the atmosphere and $d$ is the scene depth. The main challenge in single image dehazing is the ill-posed nature of the problem, as described in Eq. (1), which requires significant modeling capacity. It can be observed from Eq. (1) that there are several potential solutions for any given hazy image input. Besides, recovering $J$ becomes much simpler if information regarding $d$ is available, but in practice, depth information is rarely accessible.

Earlier single image dehazing methods attempt to recover the haze-free image $J$ based on the captured hazy image $I$ via estimating $t$ and $A$ by applying dark channel priors and color regularizers, etc [1, 10, 17, 22, 46, 57]. Recent research has shifted towards end-to-end Vision Transformer/CNN-based image dehazing methods that directly learn hazy-to-clear image translation [13, 18, 30, 39, 44, 45, 50]. However, there exist several issues: 1) these methods usually require a large number of image pairs $\{I, J\}$ during the training process, such as 13990 in Reside-ITS dataset [27]. 2) Several existing state-of-the-art methods excel in either homogeneous or non-homogeneous dehazing, but not both [18, 30]. 3) Many previous approaches focus on performance improvement by increasing model complexity, resulting in resource limitations for mobile or embedded devices. For instance, TNN [54] and TDN [30], which are two winner methods for non-homogeneous dehazing tasks on NTIRE 2021 [8] and NTIRE 2020 [7] Challenges respectively have over 45 million parameters.

Images with non-homogeneous haze have varying levels of haze across different regions, presenting a challenging task for image dehazing algorithms. Recently, NTIRE organized several non-homogeneous dehazing challenges [7–9] and introduced several small-scale real-world datasets. For instance, the dataset used in the NTIRE 2023 Challenge [9] features high-resolution non-homogeneous haze with sharp changes in haze level from certain regions to others, adding to the difficulties posed by limited training data and complex hazy patterns. The high resolution of the digital images in this challenge makes both the training and inference process more challenging.

To overcome the challenges mentioned above and effectively handle the complex distribution of non-homogeneous hazy images, we propose a novel two-stage lightweight learning-based deep method called **TransER**, which is inspired by vision transformer [16], dehazing fusion network [20, 21, 36, 53], ensemble learning [14], and knowledge distillation [23]. **TransER** comprises two deep networks: the **Trans**Conv Fusion Dehaze (TFD) model in Stage I and the Lightweight **E**nsemble **R**econstruction (LER) network in Stage II and learns from hazy images to reconstruct high-quality, haze-free images. Despite using a small number of training samples, our two-stage method achieves high fidelity and perceptual performance in both homogeneous and non-homogeneous datasets, as shown by an example in Figure 1(a) and (b) on the NTIRE 2023 test set. In particular, TFD estimates the scene information via $A$ and $t$ respectively and contains one more Vision Transformer-based decoder to jointly recreate the clean image along with $A$ and $t$. Notably, we encouraged our second network to imitate the simple image reconstruction task from Teacher Reconstruction Network (TRN) model by supervising the intermediate features in the second stage of TransER. We designed the LER as an ensemble technique to ensure robust performance in reconstructing haze-free images. In summary, our contributions are as follows:

- A two-stage deep network (TransER) to reconstruct haze-free images from hazy inputs. TransER includes two novel models (TFD and LER) and incorporates teacher-student learning from a TRN network learned image reconstruction task.
- A new TransConv Dehazing (TCD) block combines the feature attention module (FAM) with Vision Transformer, enabling it to extract both local and global information simultaneously. This integration adds flexibility in effectively handling haze in images.
- Our proposed method is extensively evaluated on various datasets and analyzed comprehensively through ablation analysis, demonstrating its effectiveness and outperforming many state-of-the-art methods on both dense haze and non-homogeneous haze scenes.

## 2. Related Work

**Single Image Dehazing.** Image dehazing methods can be divided into two categories: prior-based and deep learning-based. Prior-based methods rely on handcrafted priors such as dark channel prior (DCP) [22], color-lines [17], haze-lines [10], and color attenuation prior (CAP) [57], rank-one prior [29]. While these methods can produce images with good visibility, they may lead to unrealistic results when scenes do not fit these priors. For example, DCP performs poorly in large white regions due to its empirical statistics. Deep learning-based methods differ from prior-based methods in that they generalize from large-scale datasets to estimate haze-free images. Various methods have been proposed to estimate the transmission map and global atmospheric light [12,21,26,41] in the physical model [35] or directly learn hazy-to-clean image translation through end-to-end training [15, 31, 39, 40, 42, 50]. Notably, AtJ-DH *et al.* [21] used a shared-encoder multi-decoders architecture to be trained jointly to estimate $A$ and $t$ which is proven to be very effective and achieved top place in NTIRE 2019 Dehazing Contest [3], while two end-to-end DW-GAN [18] and TDN [30] networks, which are the two winners of non-homogeneous dehazing NTIRE Challenges recently, have been proposed to directly reconstruct the haze-free images without using atmospheric scattering model. In addition to this, many researchers have developed GAN-based architectures [19] to improve their results [40, 54]. In [18, 54], an additional discriminator was incorporated as a regularization loss term to evaluate the authenticity of the reconstructed haze-free images. Qu *et al* [40] proposed an enhanced Pix2Pix network based on GAN for dehazing, that can strengthen the dehazing effect in both color and details. Very recently, Vision Transformer (ViT) [16] has shown promising results in low-level and high-level computer vision tasks [32, 47, 48, 55], and has also demonstrated advantages in image dehazing. DeHamer [13] proposed a novel transmission-aware 3D position embedding to involve haze density-related prior information into Transformer, while in [44], a U-Net-like vision transformer model was designed with various improvements such as modified normalization layer, activation function, and spatial information aggregation scheme. Although these methods have shown promising achievements in homogeneous dehazing task due to the availability of large training datasets, they struggle with the high-resolution input hazy images and complex haze distributions found in the real-world NH-Haze 2023 dataset [9].

**Knowledge Distillation.** Knowledge distillation [23] is a technique used to transfer knowledge from one deep learning model (the teacher) to a student model. This method has been successfully applied to a wide range of tasks, including image classification, object detection, and image segmentation [16]. In the field of image dehazing, [49] proposed a dual-network approach for knowledge transfer, where the teacher network learned the distribution of clear images through an image reconstruction task, and provided prior knowledge to assist the dehazing network in restoring clear images from hazy ones. Our work is inspired by this approach, but we apply it in different ways. Particularly, in Stage II of TransER, our student model reconstructs the final clear image from parameter-based dehazed and pseudo haze-free images, while the teacher model performs a clear image reconstruction task.

**Ensemble learning.** Ensemble learning has been shown to effectively reduce variance in neural networks [14]. It is well-established that an ensemble model can outperform a single network when used in isolation [11]. In the context of non-homogeneous dehazing, researchers have explored various methods to improve performance. For instance, [53] proposed a sequential hierarchical ensemble of two different dehazing networks with varying modeling capacities to generate clear images. Similarly, [54] developed a learnable fusion tail that effectively fuses the outputs from two different neural network branches. In our proposed method, TFD generates two pseudo haze-free images, each of which focuses on different haze regions. Our ensemble model, LER, which can be classified as a Mixtures of Experts, combines the two TFD's outputs to generate the final haze-free image. Furthermore, unlike the existing state-of-the-art works that are trained end-to-end, our approach consists of two separate stages and does not include any additional pre/post-processing operations.

## 3. Proposed Method

To overcome the challenge of single image dehazing, particularly in non-homogeneous scenarios, we developed a two-stage deep network called **TransER**. Our proposed network leverages the strengths of two different learning-based models in each stage. **Trans**Conv Fusion Dehaze (TFD) model is able to utilize information from regions with varying levels of haze, while the Lightweight **E**nsemble **R**econstruction (LER) network treats distinct levels of haze differently. In this section, we present the structure of our proposed network as well as loss functions.

### 3.1. Network Structure

The proposed method, illustrated in Figure 2, consists of a two-stage deep network that transforms a hazy input image to a haze-free output. Note for brevity, some connections from encoder to decoder are omitted. The objectives of the two stages are as follows:

- **Stage I**: TFD learns to generate two different clean images from a hazy input: a pseudo haze-free through an independent decoder and a parameterized clear image via physical inspiration [35].
- **Stage II**: LER is designed to generate the final clean image by utilizing the TFD's estimated outputs, and to enhance its performance, we incorporate knowledge distillation from the teacher model.

#### 3.1.1 Stage I design

**TransConv Fusion Dehaze Network**. The proposed TFD network is mainly composed by the following build-

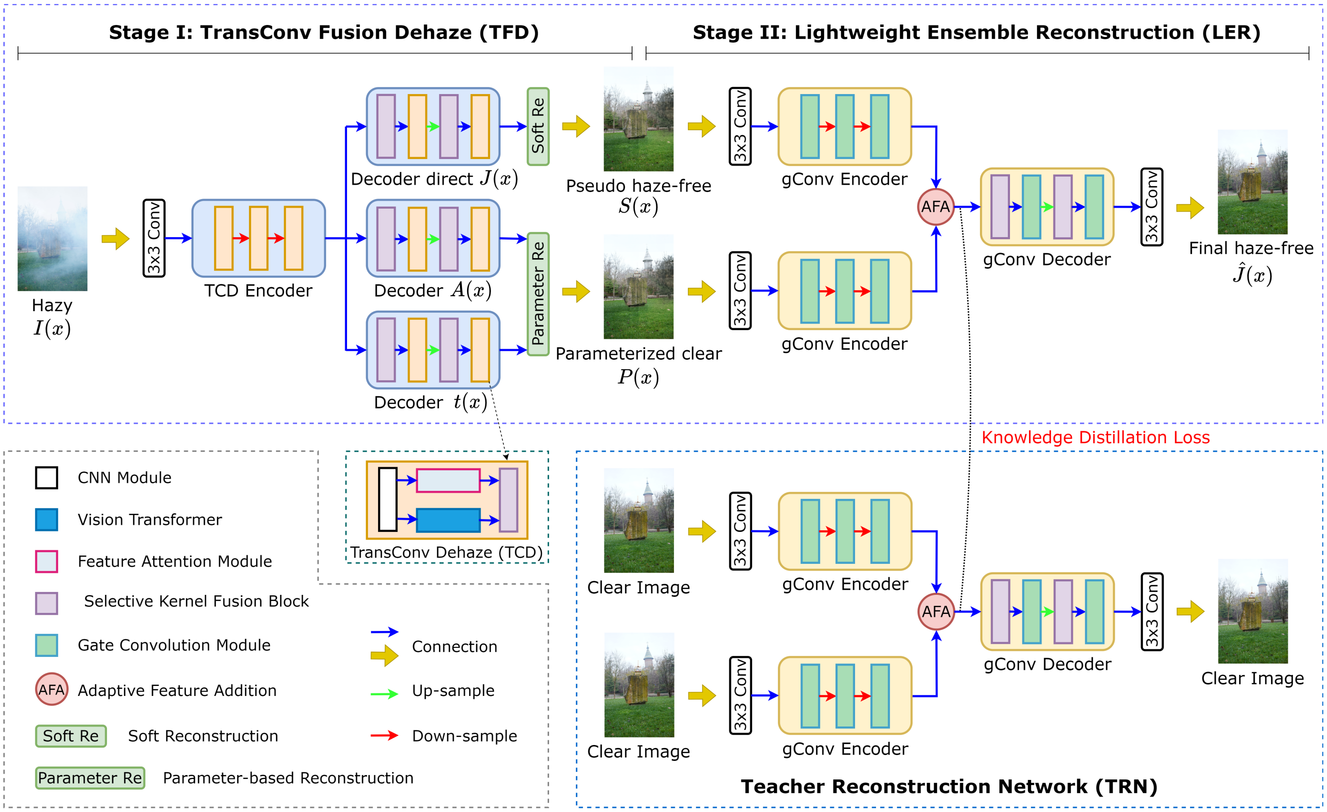


Figure 2. Illustration of our model architecture - TransER.

ing blocks: 1) one shared encoder, which is constructed based on the novel TransConv Dehaze module, and 2) three separate decoders which have similar structures as the encoder. Skip connections are used between the encoder and the decoders as in U-net. The complete network structure is illustrated in Figure 2. To have better information flow as suggested in [25], several feature maps from encoder and decoders are connected by Selective Kernel Fusion (SKF) [45]. Figure 3(c) illustrates the architecture of SKF, it is a channel attention-based fusion technique that has been shown to be more efficient than simple concatenation. The network architecture includes three decoders to predict the different estimated values $\hat{A}(x)$, $\hat{t}(x)$, and $\hat{J}_{direct}(x)$, where $x$ is the pixel location. Then, the estimated value of $\hat{J}_{AT}$ is obtained using the physical model, and the estimated values of $\hat{A}(x)$ and $\hat{t}(x)$ by the equation:

$$\hat{J}_{AT}(x) = \frac{I(x) - \hat{A}(x)(1 - \hat{t}(x))}{\hat{t}(x)} \tag{2}$$

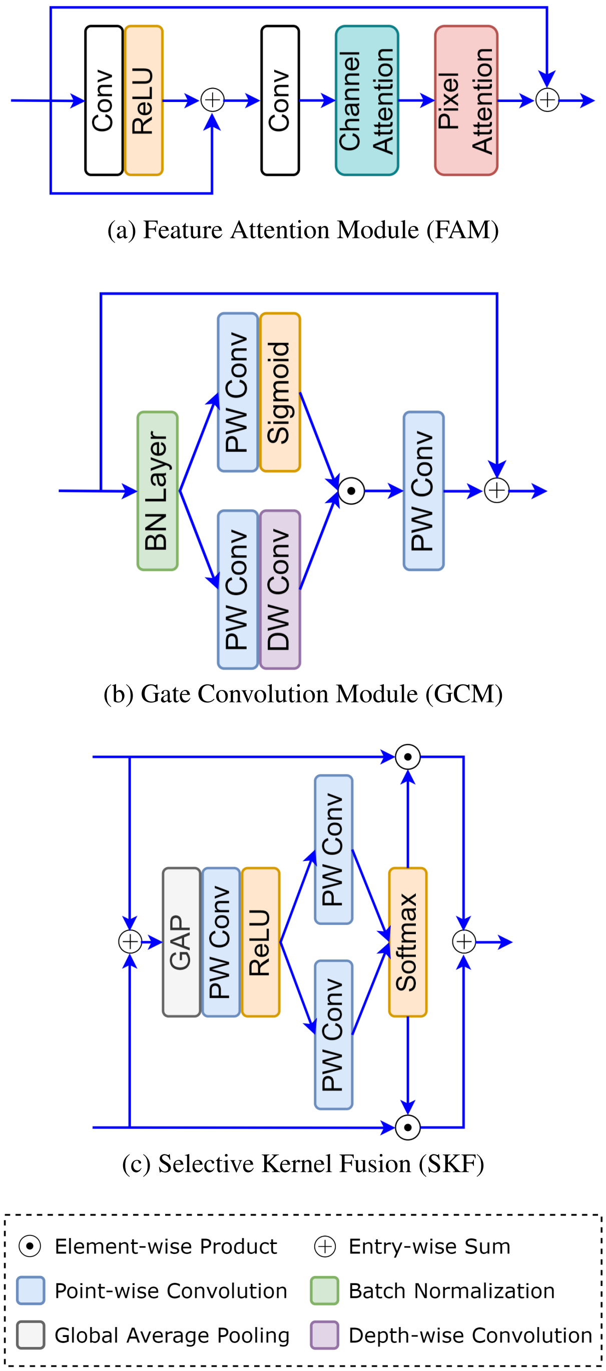
where $I$ is the input hazy image. In the Figure 2, $\hat{J}_{direct}$ is denoted as pseudo haze-free $S(x)$, while $\hat{J}_{AT}$ is the parameterized clear $P(x)$. As described in Section 2, $S(x)$ has better performance than $P(x)$ in regions with dense haze, while for regions with shallow haze, $P(x)$ performs better.

**TransConv Dehaze Module.** In scenes with non-homogeneous haze, the distribution of haze is not uniform across all image pixels. To address this issue, we draw inspiration from the feature attention module (FAM) [39] and modified Vision Transformer (ViT) [44] and propose a new TransConv Dehaze (TCD) block. The TCD block extracts feature maps in parallel, and the output from the FAM and ViT is fused using the SKF module. The FAM includes both channel and pixel attention blocks, as shown in Figure 3(a). By incorporating FAM into TFD, the network is able to focus more on relevant information such as textures, colors, and dense haze regions. This additional flexibility in dealing with non-homogeneous haze enables the proposed method to achieve superior performance.

### 3.1.2 Stage II design

We propose a dual network comprising of a teacher and student model, which are trained on different tasks while sharing the same architecture.

**Lightweigh Ensemble Reconstruction.** We propose a lightweight and straightforward ensemble reconstruction network called LER. LER follows a multi-level information extraction approach similar to U-net and is composed of two encoders and one decoder that employs the gate con-



(a) Feature Attention Module (FAM)

(b) Gate Convolution Module (GCM)

(c) Selective Kernel Fusion (SKF)

Figure 3. The architecture of feature attention module (FAM), gate convolution Module (GCM), and selective kernel fusion (SKF).

volution module (GCM) [45]. The GCM is a residual block that utilizes gating mechanisms and its architecture is depicted in Figure 3(b). The feature maps are extracted from two different inputs $S(x)$ and $P(x)$ by the encoders, and then combined using an adaptive feature addition (AFA) to effectively preserve information. Besides, the SKF module is used to dynamically fuse the feature maps from the encoders and decoder.

**Knowledge Transfer.** To enhance the naturalness of the generated clean images, we introduce a teacher network, TRN, which is trained on the task of clear image reconstruction. The TRN provides prior knowledge to the LER model through distillation loss. By leveraging the intermediate feature maps, LER can learn the distribution of clear images from TRN, thereby improving performance.

### 3.2. Network Learning Loss

The stages of the TransER method have been optimized with five loss functions which are the $L_1$ reconstruction loss $\mathcal{L}_{L_1}$, perceptual loss $\mathcal{L}_p$, MS-SSIM loss $\mathcal{L}_{ssim}$, standard deviation loss $\mathcal{L}_{std}$, and knowledge distillation loss $\mathcal{L}_{KD}$.

**Stage I optimization**. The total loss of stage I can be summarized as follows:

$$\mathcal{L}_{S_I} = \mathcal{L}_{L_1} + \lambda_1 \mathcal{L}_p + \lambda_2 \mathcal{L}_{ssim} + \lambda_3 \mathcal{L}_{std} \tag{3}$$
$$\mathcal{L}_{L_1} = |J - S|_1 + |J - P|_1$$
$$\mathcal{L}_p = ||G(J) - G(S)||_2^2 + ||G(J) - G(P)||_2^2$$
$$\mathcal{L}_{ssim} = -\texttt{SSIM}(J, S) - \texttt{SSIM}(J, P)$$
$$\mathcal{L}_{std} = \sigma_{\hat{A}}^2$$

**Stage II optimization**. The total loss of stage II can be derived as:

$$\mathcal{L}_{S_{II}} = \mathcal{L}_{L_1} + \gamma_1 \mathcal{L}_p + \gamma_2 \mathcal{L}_{ssim} + \gamma_3 \mathcal{L}_{KD} \tag{4}$$
$$\mathcal{L}_{L_1} = |J - \hat{J}|_1 \quad ; \quad \mathcal{L}_p = ||G(J) - G(\hat{J})||_2^2$$
$$\mathcal{L}_{ssim} = -\texttt{SSIM}(J, \hat{J})$$
$$\mathcal{L}_{KD} = |\texttt{Mid}_{TRN}(J, J) - \texttt{Mid}_{LER}(S, P)|_1$$

where $\lambda_1, \lambda_2, \lambda_3, \gamma_1, \gamma_2$, and $\gamma_3$ are hyperparameters used to balance the contribution of each loss term. We use $\mathcal{L}_{L_1}$ as the main loss function to train networks, which is reported by [7, 56] that training with $L_1$ loss achieved a better performance than $L_2$ loss in terms of PSNR and SSIM metrics in many image restoration tasks. $S$, $P$, and $\hat{J}$ are pseudo haze-free $S(x)$, parameterized clear $P(x)$, and final haze-free images $\hat{J}(x)$ respectively (as shown in Figure 2). $\mathcal{L}_p$ is obtained from the outputs of feature extraction layers of a pre-trained VGG16 [43] where $G(.)$ is the function representing the features extracted from $3^{rd}$, $8^{th}$, and $15^{th}$ layers of the VGG model. $\mathcal{L}_{ssim}$ is used to maximize the value of $\texttt{SSIM}$, which is refereed as Multi-Scale Structure Similarity (MS-SSIM) [51] while $\hat{A}$ is regularized by minimizing its variance $\sigma_{\hat{A}}^2$ through $\mathcal{L}_{std}$. We employ $L_1$ loss to calculate the $\mathcal{L}_{KD}$, in which $\texttt{Mid}_{TRN}(.)$ and $\texttt{Mid}_{LER}(.)$ denote the intermediate feature maps (after AFA's output) of TRN and LER models. Additionally, we only use $L_1$ loss to optimize the TRN network.

# 4. Dataset, Training, Testing and Metric

## 4.1. Dataset

In this study, we evaluate the performance of TransER on both synthetic and real-world datasets. We use the Indoor Training Set (ITS) and the Synthetic Objective Testing Set (SOTS) from RESIDE [27] for the synthetic dataset. For the real-world datasets, we experiment with Dense-Haze used in NTIRE 2019 Dehazing Challenge [3], NH-Haze from NTIRE 2020 Dehazing Challenge [7], NH-Haze 2 in NTIRE 2021 Dehazing Challenge [8], and high resolution NH-Haze 3 from NTIRE 2023 Dehazing Challenge [9].

RESIDE provides a benchmark for single image dehazing, with a collection of large-scale training and testing images across indoor and outdoor scenarios. We train our proposed method using 13,990 synthetic images from ITS and evaluated it on 500 indoor images from SOTS. For real-world evaluation, we use the Dense-Haze, NH-Haze, NH-Haze 2, and NH-Haze 3 datasets provided by the NTIRE Challenges. Dense-Haze consists of 45 training, 5 validation, and 5 testing dense hazy images, while NH-Haze contains 45 training data, 5 validation data, and 5 testing data. NH-Haze 2 and NH-Haze 3 contain 35 and 50 non-homogeneous images respectively. We utilized the official training, validation, and testing split for Dense-Haze and NH-Haze. However, because the ground truth images for validation and testing sets have not yet been released, we choose the first 20 pairs as the training set and the remaining 5 pairs to evaluate for NH-Haze 2 experiments, and we only present qualitative results on the high resolution NH-Haze 3 dataset.

## 4.2. Training

During the training process, we randomly apply augmentations, including horizontal flipping and rotation by $90°$, $180°$, and $270°$. Input patches of size $256 \times 256$ are extracted from the training images. We use the AdamW optimizer [34] with default values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rates for TFD, TRN, and LER models are set to $\{4, 1, 1\} \times 10^{-4}$, respectively, and we apply the cosine annealing schedule [33] to gradually reduce the learning rate from the initial values to $\{4, 1, 1\} \times 10^{-7}$. We adapt a multi-step strategy to train our proposed method, TransER, as described below:

- **Step 1 - TFD Training**. In this step, we only train the TFD model for 8000 epochs with a batch size of 16. The values of the hyperparameters $\lambda_1$, $\lambda_2$, and $\lambda_3$ are set to 0.1, 0.5, and 0.0001 respectively, through cross-validation [37].

- **Step 2 - TRN Training**. The TRN network is trained using a batch size of 32 and optimized for 80 epochs in the clean reconstruction task.

- **Step 3 - LER Training**. In the last step, we freeze the parameters of the TFD and TRN models and only update the parameters of the LER network. We train LER for 100 epochs using a batch size of 32. The hyperparameters of loss functions, $\gamma_1, \gamma_2$, and $\gamma_3$ are $0.1$, $0.2$, and $1.0$ respectively.

Our models are implemented using Pytorch framework [38] and all experiments are performed on two NVIDIA Titan X (12GB) GPUs [1]

## 4.3. Testing and Metric

The proposed method involves a sequential processing of the input hazy image through TFD and LER modules to generate the final haze-free output. For quantitative evaluation, we adopt the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) which are often used as criteria for evaluating image quality in dehazing task. Notably, we do not employ commonly used model-ensemble or self-ensemble techniques to produce the clean image.

# 5. Experimental Results

In this section, we present the experimental results of our TransER method, including an ablation study of different components and loss terms, as well as comparisons with state-of-the-art methods.

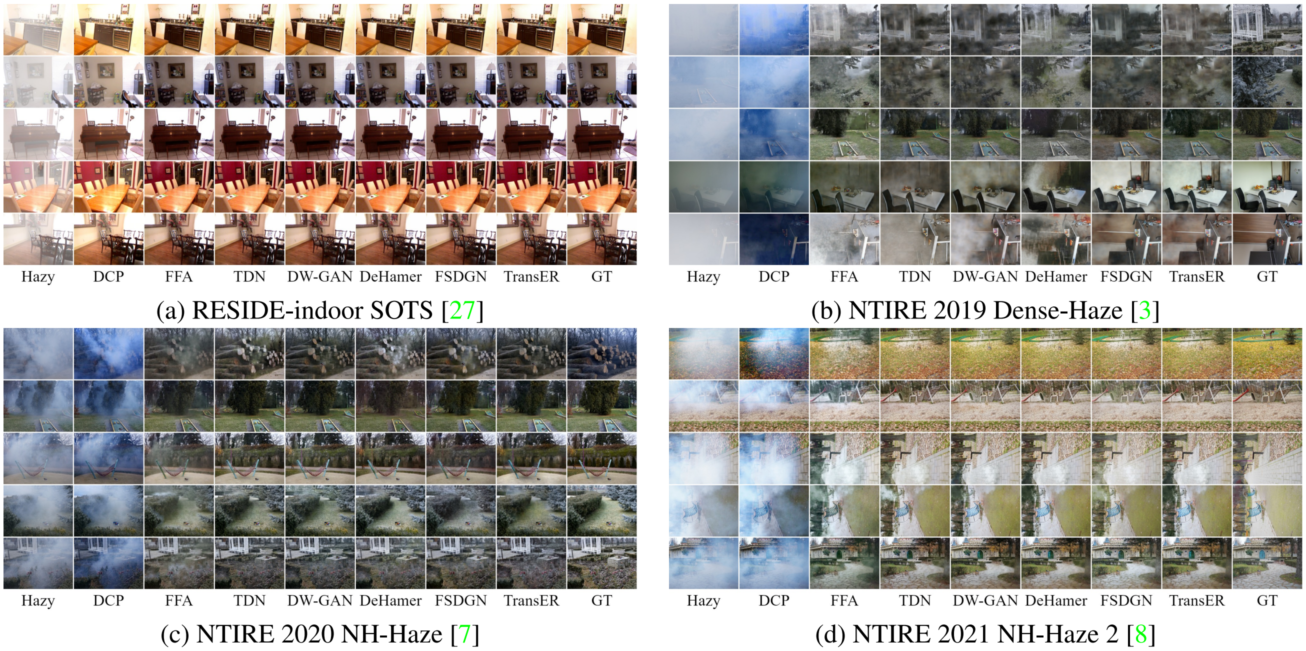
## 5.1. Ablation Study

To ensure fairness, we conduct ablation studies on the NH-Haze 3 validation dataset provided in NTIRE 2023 [9], which includes 5 hazy inputs without ground truth images. After completing the experiments, we uploaded 5 reconstructed haze-free images to the validation server and report the received results in Tables 1 and 2.

**Effects of TCD module, LER, TFD, and TRN models**. In Table 1, we show quantitative results between 7 methods. We use the notation TFD (without TCD) to indicate that instead of using the proposed TCD block, we utilize a modified vision transformer [44] for the TFD model of our network. It is clear that the use of TRN can help the LER model in reconstructing the haze-free images, hence significantly enhancing the PSNR and SSIM. Furthermore, our two-stage method performs much better than those only with one stage (either LER or TFD), indicating the effectiveness of our design. Our full model achieves a PSNR of $21.28$ dB and an SSIM of $0.693$, the scores demonstrate that the effective combination of the three models with the TCD module leads to a notable performance improvement.

**Effects of different loss terms**. We also analyze the effect of removing different loss terms from the total loss

[1] Please find implementation details at https://github.com/trungpsu1210/TransER



(a) RESIDE-indoor SOTS [27]

(b) NTIRE 2019 Dense-Haze [3]

(c) NTIRE 2020 NH-Haze [7]

(d) NTIRE 2021 NH-Haze 2 [8]

Figure 4. The qualitative results on different testing datasets.

Table 1. Ablation study results of TransER's architecture.

| **Method** | **Results** | |
|---|---|---|
| | PSNR | SSIM |
| LER | 19.53 | 0.662 |
| TFD (without TCD) | 20.01 | 0.675 |
| TFD | 20.96 | 0.685 |
| TFD (without TCD) + LER | 20.14 | 0.667 |
| TFD + LER | 20.95 | 0.684 |
| TFD (without TCD) + LER + TRN | 21.16 | 0.688 |
| TransER (ours) | **21.28** | **0.693** |

Table 2. Ablation studies for different loss function terms.

| **Loss terms** | | | | | **Results** | |
|---|---|---|---|---|---|---|
| $\mathcal{L}_{L_1}$ | $\mathcal{L}_p$ | $\mathcal{L}_{ssim}$ | $\mathcal{L}_{std}$ | $\mathcal{L}_{KD}$ | PSNR | SSIM |
| ✓ | | | | | 20.37 | 0.671 |
| ✓ | ✓ | | | | 20.68 | 0.675 |
| ✓ | ✓ | ✓ | | | 20.85 | 0.688 |
| ✓ | ✓ | ✓ | ✓ | | 20.95 | 0.684 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **21.28** | **0.693** |

functions $\mathcal{L}_{S_I}$ and $\mathcal{L}_{S_{II}}$ in Eq. (3) and (4) respectively. We can observe that the total customized losses increase the performance. Specifically, the perceptual loss, MS-SSIM, and knowledge distillation losses have a significant impact on both PSNR and SSIM results, while the standard deviation loss has some effect on the PSNR results.

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

This section illustrates the comparisons between TransER with the state-of-the-art (SOTA) methods on the datasets introduced in Section 4. These SOTA methods compared in our experiments consist of one prior-based image dehazing method, DCP [22], and five learning-based methods, namely FFA [39], TDN [30], DW-GAN [18], DeHamer [13], and FSDGN [52]. TDN and DW-GAN are the winner methods in NTIRE 2020 and NTIRE 2021 NonHomogeneous Dehazing Challenge.

Table 3 compares the quantitative results of different methods, which indicates our TransER achieves the best and second best performances on both synthetic and real-world datasets. Notably, by utilizing point-wise and depth-wise separable convolution [24] to efficiently aggregate information and transform features, TransER achieves the best parameter-performance trade-off, compared to the other SOTA approaches, especially the champion models [18,30]. For visual quality, in Figure 4 we observe that most methods generate pleasing images on synthetic dataset, except for DCP, which produces images with color distortion for both two types of data. Especially, on real-world datasets, although other learning-based methods can generate better results, there are still obvious visual problems such as low brightness and blurry borders. In contrast, TransER generates visually pleasing results, which are close to the ground truth images in terms of color and object details.

Table 3. Quantitative comparisons between TransER and state-of-the-art methods over RESIDE-indoor [27], Dense-Haze [3], NH-Haze [7] and NH-Haze 2 [8]. The best results are in bold, and the second-best results are underlined.

| Method | Venue | RESIDE-indoor | | NTIRE 2019 | | NTIRE 2020 | | NTIRE 2021 | | # Parameters |
|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | |
| DCP [22] | CVPR 2009 | 16.62 | 0.817 | 10.85 | 0.404 | 12.29 | 0.411 | 11.30 | 0.605 | - |
| FFA [39] | AAAI 2020 | 36.39 | 0.989 | 16.26 | 0.545 | 18.51 | 0.637 | 20.40 | 0.806 | 4.68 M |
| TDN [30] | CVPRW 2020 | 34.59 | 0.975 | 15.50 | 0.508 | 20.44 | 0.668 | 20.23 | 0.762 | 46.18 M |
| DW-GAN [18] | CVPRW 2021 | 35.94 | 0.986 | 16.49 | 0.591 | 21.51 | 0.711 | **21.99** | 0.856 | 56.73 M |
| DeHamer [13] | CVPR 2022 | 36.63 | 0.988 | 16.62 | 0.560 | 20.66 | 0.684 | 20.72 | 0.815 | 132.45 M |
| FSDGN [52] | ECCV 2022 | **38.63** | 0.990 | 16.91 | 0.581 | 19.99 | 0.731 | 20.05 | 0.786 | 2.73 M |
| TransER (ours) | - | 37.24 | **0.992** | **17.03** | **0.597** | **21.64** | **0.743** | 21.59 | **0.871** | **2.60** M |

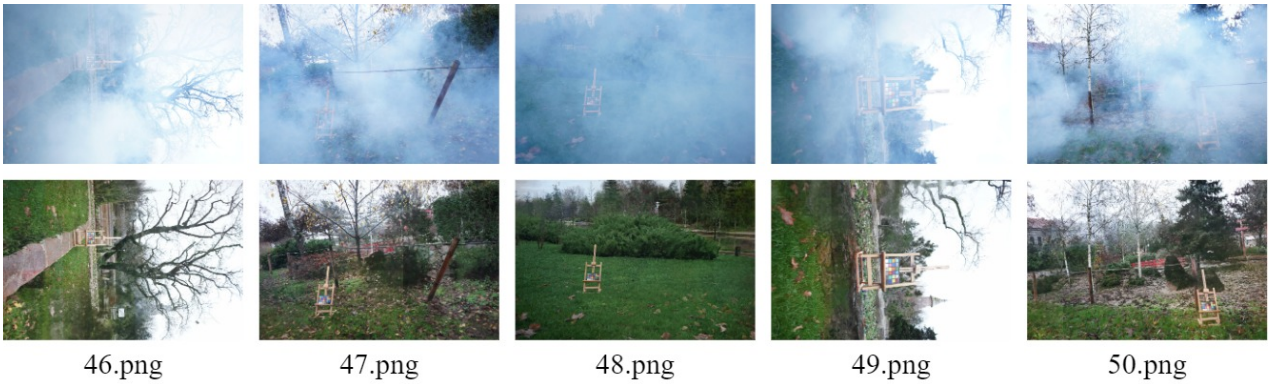


Figure 5. Test results of high resolution non-homogeneous dehazing NTIRE 2023 challenge [9] over 5 hazy images. The first row shows the original hazy inputs and the second row illustrates dehazing results using TransER.

### 5.3. NTIRE 2023 Dehazing Challenge

We did not use any pre-trained models or extra datasets during the NTIRE 2023 dehazing challenge [9]. Moreover, we only employed the TransER pipeline without any self-ensemble/model-ensemble or pre/post-processing methods to boost performance. The dehazed images we generated, presented in Figure 5, clearly show the removal of haze in the 47.png, 48.png, and 50.png cases. In Table 4, we compare TransER with other leading methods, and the results are verified and reported in the final ranking table [9] by the organizers. As shown, TransER achieved good performance in terms of PSNR, SSIM, and LPIPS, while using significantly fewer parameters and low inference time. These results validate the effectiveness of TransER.

## 6. Conclusion

This study introduces a two-stage learning-based approach for single image dehazing. In the first stage, inspired by the physical haze model, our proposed method jointly estimates the global atmospheric light, transmission map, and haze-free scene directly using the novel TransConv Dehaze module, which can extract both global and local information in parallel. The second stage is a simple and lightweight model inspired by gate mechanism architecture, ensemble learning, and knowledge distillation, to generate the final haze-free output. Experimental results on various synthetic and real-world datasets with dense and non-homogeneous haze demonstrate the superior performance of our TransER method over state-of-the-art alternatives. Notably, our method achieves good results on the high-resolution non-homogeneous NTIRE 2023 dataset, outperforming many competing methods.

Table 4. Comparison between our network against other top performing methods participating in the competition over the test set provided in NTIRE 2023. Results are reported in [9].

| | IR-SDE | DWT-FFC-GAN | TransER |
|---|---|---|---|
| PSNR ↑ | 20.83 | 22.87 | 22.01 |
| SSIM ↑ | 0.61 | 0.71 | 0.70 |
| LPIPS ↓ | 0.406 | 0.346 | 0.384 |
| # Parameters | 78 M | 373 M | 2.6 M |
| Runtime | 5.0 s | 23.3 s | 0.72 s |

## References

[1] Codruta Orniana Ancuti, Cosmin Ancuti, and Philippe Bekaert. Effective single image dehazing by fusion. In *2010 IEEE International Conference on Image Processing*, pages 3541–3544, 2010. 2

[2] Codruta O. Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE International Conference on Image Processing*, 2019. 1

[3] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte. Ntire 2019 image dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2241–2253, 2019. 3, 6, 7, 8

[4] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1798–1805, 2020. 1

[5] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 867–8678, 2018. 1

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

[7] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2020 challenge on nonhomogeneous dehazing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2029–2044, 2020. 2, 5, 6, 7, 8

[8] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2021 nonhomogeneous dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 627–646, 2021. 2, 6, 7, 8

[9] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1, 2, 3, 6, 8

[10] Dana Berman, Tali Treibitz, and Shai Avidan. Non-local image dehazing. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* ., pages 1674–1682, 2016. 2

[11] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., USA, 1995. 3

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

[13] Saeed Anwar Runmin Cong Wenqi Ren Chongyi Li Chun-Le Guo, Qixin Yan. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Patter Recognition*, 2022. 2, 3, 7, 8

[14] Thomas G. Dietterichl. Ensemble learning. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 405–408. MIT Press, 2002. 2, 3

[15] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* ., pages 2154–2164, 2020. 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3

[17] Raanan Fattal. Dehazing using color-lines. volume 34, New York, NY, USA, 2014. ACM. 2

[18] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. Dw-gan: A discrete wavelet transform gan for non-homogeneous dehazing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 203–212, 2021. 2, 3, 7, 8

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, and Ozair. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[20] Tiantong Guo, Venkateswararao Cherukuri, and Vishal Monga. Dense '123' color enhancement dehazing network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1, 2

[21] Tiantong Guo, Xuelu Li, Venkateswararao Cherukuri, and Vishal Monga. Dense scene information estimation network for dehazing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1, 2, 3

[22] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1956–1963, 2009. 2, 7, 8

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. cite arxiv:1503.02531Comment: NIPS 2014 Deep Learning Workshop. 2, 3

[24] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, and Weyand. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 7

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* ., pages 2261–2269, 2017. 4

[26] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4780–4788, 2017. 3

[27] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2019. 2, 6, 7, 8

[28] Zhuwen Li, Ping Tan, Robby T. Tan, Danping Zou, Steven Zhiying Zhou, and Loong-Fah Cheong. Simultaneous video defogging and stereo reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition .*, pages 4988–4997, 2015. 2

[29] Jun Liu, Ryan Wen Liu, Jianing Sun, and Tieyong Zeng. Rank-one prior: Real-time scene recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[30] Jing Liu, Haiyan Wu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Trident dehazing network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1732–1741, 2020. 2, 3, 7, 8

[31] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7313–7322, 2019. 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[33] IIya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 6

[34] IIya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[35] E. J. MCCartney. Optics of the atmosphere. scattering by molecules and particles. *Wiley Series in Pure and Applied Optics, New York: Wiley*, 1976. 1, 2, 3

[36] Kareem Metwaly, Xuelu Li, Tiantong Guo, and Vishal Monga. Nonlocal channel attention for nonhomogeneous image dehazing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1842–1851, 2020. 1, 2

[37] Vishal Monga. Handbook of convex optimization methods in imaging science. 2017. 6

[38] Adam Paszke, Sam Gross, and Soumith Chintala. 6

[39] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2, 3, 4, 7, 8

[40] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, pages 8152–8160, 2019. 3

[41] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multiscale convolutional neural networks. In *European Conference on Computer Vision*, 2016. 3

[42] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks large-scale image recognition. 2015. 5

[44] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 1, 2, 3, 4, 6

[45] Yuda Song, Yang Zhou, Hui Qian, and Xin Du. Rethinking performance gains in image dehazing networks. *arXiv preprint arXiv:2209.11448*, 2022. 1, 2, 4, 5

[46] Ketan Tang, Jianchao Yang, and Jue Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2995–3002, 2014. 2

[47] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions, 2021. 3

[48] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition .*, pages 17683–17693, June 2022. 3

[49] Haiyan Wu, Jing Liu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Knowledge transfer dehazing network for nonhomogeneous dehazing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1975–1983, 2020. 3

[50] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the Conference on Computer Vision and Pattern Recognition .*, pages 10551–10560, June 2021. 2, 3

[51] Zhang-Shu Xiao. A multi-scale structure similarity metric for image fusion qulity assessment. In *2011 International Conference on Wavelet Analysis and Pattern Recognition*, pages 69–72, 2011. 5

[52] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *European Conference on Computer Vision*, page 181–198, 2022. 7, 8

[53] Mingzhao Yu, Venkateswararao Cherukuri, Tiantong Guo, and Vishal Monga. Ensemble dehazing networks for nonhomogeneous haze. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1832–1841, 2020. 1, 2, 3

[54] Yankun Yu, Huan Liu, Minghan Fu, Jun Chen, Xiyao Wang, and Keyan Wang. A two-branch neural network for nonhomogeneous dehazing via ensemble learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 193–202, 2021. 2, 3

[55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 3

[56] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 2017. 5

[57] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing*, 2015. 2