Breaking Through the Haze: An Advanced Non-Homogeneous Dehazing Method based on Fast Fourier Convolution and ConvNeXt

Han Zhou[1], Wei Dong[2], Yangyi Liu[1], Jun Chen[1]
[1]Department of Electrical and Computer Engineering, Mcmaster University, Hamilton, Canada
[2]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
zhouh115@mcmaster.ca, wdong1745376@gmail.com, {liu5, chenjun}@mcmaster.ca

Abstract

Haze usually leads to deteriorated images with low contrast, color shift and structural distortion. We observe that many deep learning based models exhibit exceptional performance on removing homogeneous haze, but they usually fail to address the challenge of non-homogeneous dehazing. Two main factors account for this situation. Firstly, due to the intricate and non uniform distribution of dense haze, the recovery of structural and chromatic features with high fidelity is challenging, particularly in regions with heavy haze. Secondly, the existing small scale datasets for non-homogeneous dehazing are inadequate to support reliable learning of feature mappings between hazy images and their corresponding haze-free counterparts by convolutional neural network (CNN)-based models. To tackle these two challenges, we propose a novel two branch network that leverages 2D discrete wavelete transform (DWT), fast Fourier convolution (FFC) residual block and a pretrained ConvNeXt model. Specifically, in the DWT-FFC frequency branch, our model exploits DWT to capture more high-frequency features. Moreover, by taking advantage of the large receptive field provided by FFC residual blocks, our model is able to effectively explore global contextual information and produce images with better perceptual quality. In the prior knowledge branch, an ImageNet pretrained ConvNeXt as opposed to Res2Net is adopted. This enables our model to learn more supplementary information and acquire a stronger generalization ability. The feasibility and effectiveness of the proposed method is demonstrated via extensive experiments and ablation studies. The code is available at https://github.com/zhouh115/DWT-FFC.

1. Introduction

As a natural phenomenon, haze usually heavily reduces the visibility, resulting in blurred hazy images with low contrast, color shift and structural distortion. Various intelligent applications, like object detection [41] and autonomous driving [45], need to operate normally in hazy conditions, which necessitates the restoration of missing information from hazy images. As a consequence, image dehazing has been studied extensively in the field of computer vision recently [9, 19, 23, 27, 39, 47–50, 54, 57].

Early methods for image dehazing [21, 26] are primarily developed based on the atmospheric scattering model (ASM) [35] to establish the correspondence between hazy images and hazy-free images. This model can be described by Eq. (1) below:

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

Here I and J represent the hazy image and its clear counterpart, respectively; x indicates the pixel position; A denotes the global atmosphere light; $t(x)$ is the transmission map, which is determined by the atmosphere scattering parameter β and the scene depth $d(x)$ as follows:

$$t(x) = e^{-\beta d(x)}. \tag{2}$$

It is clear that assuming the validity of ASM, image dehazing boils down to estimating $t(x)$ and A [19, 21, 34]. Unfortunately, this model is only applicable to idealized homogeneous haze. As such, ASM based methods cannot handle non-homogeneous dehazing tasks.

In recent several years, inspired by its remarkable success for classification, object detection and other vision tasks [16, 17, 55], deep learning has also been brought to bear upon single image dehazing [3–6, 10, 13, 28, 29, 48, 50]. In principle, with end-to-end supervised training, deep learning based dehazing methods are no longer confined by the ASM framework. Indeed, they are shown to be able to handle complex and non-homogeneous hazy images to a certain extent [5, 6].

Deep learning based dehazing methods often rely on the availability of large training data. However, it is very difficult and even impossible to acquire big volumes of image pairs in the real world [15]. The lack of sufficient training data has become a hindrance to the development of

non-homogeneous dehazing methods. To cope with limited training data and alleviate over-fitting, some recent methods [5, 6, 15] resort to pretrained models, e.g., Res2Net pertrained on ImageNet [12, 16], for transfer learning [37]. However, these methods do not take advantage of the state-of-the art models, such as Vision Transformer [14] and its follow-ups [18, 53], Swin Transformer [30] and ConvNeXt [31, 46], which achieve superior performance compared to Res2Net on ImageNet. As such, there is a potential to improve the dehazing performance by utilizing more powerful pretrained models.

Besides, most existing methods are incapable of recovering high-frequency components, such as edges and fine textures. DW-GAN [15] leverages discrete wavelet transform (DWT) to extract high-frequency features in the downsampling phase and passes them through the upsampling process. However, this method still has difficulty in dealing with severe hazy areas.

Therefore, two problems need to be addressed in order to realize high-quality dehazing. 1) Including those released by NTIRE [1, 2], most datasets for non-homogeneous dehazing are small-scale ones, which are not sufficient for CNN-based models to learn the mapping between hazy images and its corresponding hazy-free images. Although adopting a pretrained model can alleviate the over-fitting problem caused by limited training data, the network should be designed carefully to maximize its ability to acquire prior knowledge. 2) The complicated and non-uniform haze patterns pose significant challenges to image restoration, especially regarding dense haze areas [6, 15].

In consideration of the aforementioned two problems, we propose a two-branch generative adversarial network, with each branch designed to address one problem. The first branch aims to learn the color and structure mapping from hazy to clean images. It consists of three DWT blocks and three FFC [11, 43] residual blocks. DWT blocks are used to acquire high-frequency knowledge and structure details [15, 22, 42], and FFC residual blocks provide a wide mixed receptive field that covers an entire image by simultaneously exploiting spectral and spatial information [32, 43]. Due to the larger receptive field, FFC residual blocks enable the encoder to capitalize on the global context and thus improve the perceptual quality of the recovered image, which is especially crucial for high-resolution non-homogeneous dehazing. The second branch serves the purpose of transfer-learning [37, 44, 58]. Specifically, we employ the first three layers of a pretrained ConvNeXt to build the encoder due to its outstanding classification performance. In general, pretrained models can successfully adapt to different tasks than what they were originally trained for [6]. Compared to Res2Net, ConvNeXt performs much better on ImageNet with the incorporation of several key components of vision transformer and the preservation of the advantage of convolutional network [14, 31]. For the decoder of the second branch, pixel-shuffle layers and channel/pixel wise attention blocks [25] are employed to gradually recover the image to its initial resolution. Then, a fusion operation [6] is utilized to aggregate the outputs of the two branches. Finally, a discriminator guided by the adversarial loss [15] is introduced to ensure the perceptual quality of the final reconstruction.

The main contributions of our work are as follows: 1) We introduce FFC residual blocks and combine them with 2D discrete wavelet transform to tackle non-homogeneous dehazing. 2) We employ the pretrained ConvNeXt model to leverage the prior knowledge to cope with the small-scale dataset problem and verify the effectiveness of this approach. 3) We conduct extensive experiments and ablation studies to justify the overall design and demonstrate its competitive performance.

2. Related Works

Single Image Dehazing. Single image dehazing is a challenging task in computer vision and image processing, as it involves removing the unwanted atmospheric haze from a single input image. Over the years, several effective methods that can be divided into two main categories, $i.e.$, physical-based methods and CNN-based solutions, have been proposed to tackle this problem. Physical-based methods mainly depend on ASM [33, 35] and the hand-crafted priors, like dark channel prior [19], color attenuation prior [38], non-local prior [8]. However, owing to the limited applicability of the underlying assumptions, physical-based methods tend to be not very robust.

With the rapid advancement of deep learning, the past few years witnessed its wide applications in single image dehazing. Early deep learning based methods still utilize ASM. For instance, DehazeNet [9] designs CNN model to estimate the medium transmission map, then uses it to obtain a dehazed image via ASM. Later, AOD-Net [23] estimates the atmospheric light and transmission map simultaneously to generate the recovered image. Recently, various deep learning models have been proposed to directly map hazy images to their clean counterparts without resorting to ASM. For example, GCANet [10] introduces a gated context aggregation network to remove the grid artifacts and fuse the feature representations of different levels. Qin et al. [49] propose FFA-net which handles different features and pixel regions adaptively in order to enhance flexibility via groups of channel attention and pixel attention mechanisms. MSBDN [13] employs a multi-scale boosted decoder to gradually recover the dehazed images. Most previous works assume a homogeneous distribution of haze, which is often not representative of real-world scenarios and can lead to significant performance degradation in scenes with dense or non-homogeneous haze. To handle real-world hazy images, Trident Dehazing Network (TDN) [5], which

consists of details refinement sub-net, encoder-decoder subnet and haze density map generation sub-net, has been proposed in the NTIRE 2020 NonHomogeneous dehazing challenge. In the NTIRE 2021 NonHomogeneous dehazing challenge [6], TDN is surpassed by DW-GAN [15] which employs DWT to extract low-frequency and high-frequency features, which are beneficial to the dehazing task. Besides, [48] introduces a novel dehazing approach based on contrastive learning that leverages both positive and negative image information. All the aforementioned methods, except for DW-GAN, are spatial-domain-centric dehazing methods, which do not directly exploit the characteristics of haze degradation in the frequency domain [50].

Frequency Domain Learning. In recent years, there has been a growing research attention to frequency domain learning in pursuit of effective spectral features. Specifically, [56] exploits wavelet-based representations that facilitate high-resolution image restoration. [52] decomposes images into low-frequency and high-frequency bands via discrete wavelet transform and extract features for each band. In addition, Chi et al. [11] propose FFC as a non-local operation unit that concurrently enlarges the receptive field by learning spatial and spectral features. Subsequently, [43] achieves excellent performance in the large mask inpainting task by employing FFC as the main convolution operation and the proposed method exhibits strong generalization abilities. [50] unveils the connection between haze degradation and the frequency property and designs a dual-branch network that guide the dehazing process spatially and spectrally.

The above-mentioned methods have demonstrated that frequency domain information can be leveraged to significantly enhance the performance of image restoration methods. As such, We borrow the idea of wavelet-based decomposition and incorporate non-local FFC operation units into our deep learning network architecture. By this, our model can effectively utilize spectral information while maintaining meaningful texture details and ensuring the consistent structures of the final recovered high-resolution image. We shall verify the superiority of our proposed model over the existing ones through quantitative evaluation.

Transfer Learning. Transfer learning aims to enhance the capability of target models on specific tasks by leveraging the knowledge acquired from related but distinct tasks, which weakens the need for large volumes of data on target domain [37, 44, 58]. Some existing methods employ substantial prior knowledge obtained through pre-training on ImageNet to assist image restoration tasks. As an example, the champions of both NTIRE dehazing challenges in 2020 and 2021 [5, 6] utilize the pretrained Res2Net model as the fundamental block for knowledge transfer, successfully alleviating the over-fitting problem caused by small-scale training datasets. Despite its impressive capabilities, Res2Net exhibits limited efficacy in dehazing, particularly for high-resolution images with non-uniform haze. We intend to substantiate this through our experimental analysis. Furthermore, it is worth noting that even in the classification area, the performance of CNNs, like Res2Net, has been overshadowed by that of Vision Transformers [14]. Liu et al. [31] redesign the standard ResNet [20] by mimicing Vision Transformers through the incorporation of several key components that contribute to the remarkable performance of transformers. The resulting architecture, named ConvNeXt, outperforms Swin Transformers [30] and challenges the widely held belief that Vision Transformers are more accurate and efficient than CNNs. However, to the best of our knowledge, ConvNeXt-based image restoration is still a largely unexplored territory.

3. Proposed Method

Within this part, we first describe the details of our proposed network (shown as Fig. 1) based on DWT, FFC residual block and pretrained ConvNeXt parameters. Then, we introduce DWT and FFC residual blocks in detail, and analyze their significance to the whole network, respectively. Finally, we discuss the loss functions utilized to train our model.

3.1. Network Framework

Many methods with two-branches have shown great success in NTIRE 2020 and 2021 NonHomogeneous dehazing challenge [5, 6]. In observing that, we design a two branch neural network (shown as Fig. 1).

DWT-FFC branch. Inspired by [6], we construct our DWT-FFC frequency branch as a encoder-decoder network to learn the feature mapping between hazy images and clear images, and we leverage massive skip connections at each feature scale. Besides conventional convolution techniques, we adopt discrete wavelet transform (DWT) to achieve feature extraction. By using DWT, both high-frequency and low-frequency features can be detected by our model (Sec. 3.2 provides detailed explanation). As indicated in Fig. 2, low-frequency representations are concatenated with common convolution output, whereas high-frequency features are transferred to the up-sampling module to recover the hazy-free image gradually. In order to make the reconstructed image more realistic and perceptual, as shown in Fig. 3, we propose to utilize fast Fourier convolution residual blocks (FFC, details can be found in Sec. 3.3) to utilize the spatial and spectral information for dehazing.

However, the model sorely with DWT-FFC frequency branch cannot achieve plausible performance for non-homogeneous challenge due to the lack of large volumes of training data. Therefore, we further introduce the second branch with a strong network pretrained on large datasets to acquire prior knowledge.

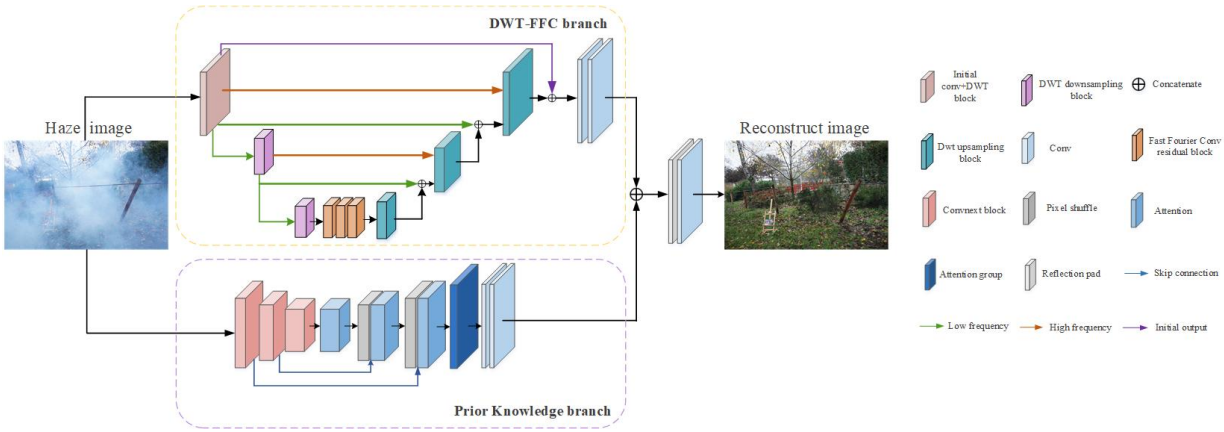

Figure 1. The network structure of our proposed method.

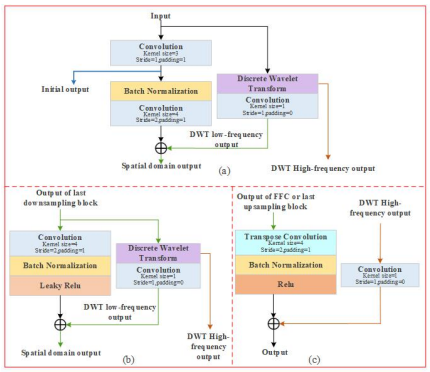

Figure 2. The illustration of Connections between Conventional Convolution and DWT in DWT-FFC branch. Green lines represent DWT low-frequency features, brown lines denote DWT high-frequency features.

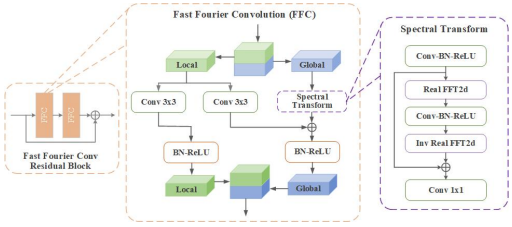

Figure 3. The network structure of FFC residual block. A FFC residual block has two FFC units, each unit contains one spectral transform.

Prior Knowledge Branch. Our second branch, prior knowledge branch, is designed to bring additional information from large dataset image classification task to the current non-homogeneous dehazing challenge. Specifically, we design this branch similar to the U-Net [36], and utilize the ImagNet pretrained ConvNeXt as the backbone of the encoder. We only adopt the first three stages of ConvNeXt. In this way, the model with the pretrained weights can perform better than that with random initial parameters. For the decoder, inspired by [49], several upsampling layers are employed and each upsampling layer contains a pixel-shuffle block and a attention module. Here, pixel-shuffle blocks are introduced to decrease the computational burden and make the size of the feature maps gradually recover to the original resolution, attention blocks enable our model to identify the dynamic hazy patterns. Moreover, as shown in Fig. 3, skip connection technique is used to enhance the generalization ability. Finally, final recovered clear images are generated by combining the outputs of each branch via a simple and effective fusion [51].

3.2. Discrete Wavelet Transform

2D discrete wavelet transform has one low-pass filter (f_{LL}) and three high-pass filters (f_{LH}, f_{HL}, f_{HH}), and these filters are equivalent to convolution operations with fixed parameters and stride 2. For example, in Haar DWT introduced in our method, $f_{LL} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $f_{LH} = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$, $f_{HL} = \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}$, $f_{HH} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. After convolving with each filter, the input can be decomposed into four sub-components, *i.e.*, x_{LL}, x_{LH}, x_{HL} and x_{HH}. Specifically, x_{LL} can be expressed as Eq. (3), where i and j donates the index of pixel, and the expressions of the other three sub-bands are similar to that of x_{LL}.

$$\begin{aligned} x_{LL} = x(2i-1, 2j-1) + x(2i-1, 2j) \\ + x(2i, 2j-1) + x(2i, 2j) \end{aligned} \tag{3}$$

Besides, we combine x_{LL} with conventional convolution outputs at each feature scale so that our proposed model can learn spatial and frequency information.

3.3. Fast Fourier Convolution

Fast Fourier convolution, based on the channel-wise fast Fourier transform, has a large receptive field as the entire image or feature map. FFC splits the input into two parallel branches: $i)$ *local branch* applies two conventional convolutions in parallel, and $ii)$ *global branch* simultaneously utilizes one conventional convolution and one spectral transform to account for the global context. Specifically, there are three steps in spectral transform:

a) For an input tensor, the $Real\ FFT2d$ is applied and the real and imaginary parts are concatenated:

Real FFT2d: $\mathbb{R}^{H\times W\times C} \rightarrow \mathbb{C}^{H\times \frac{W}{2}\times C}$,

ComplexToReal: $\mathbb{C}^{H\times \frac{W}{2}\times C} \rightarrow \mathbb{R}^{H\times \frac{W}{2}\times 2C}$;

b) ReLU activation, batch normalization and convolution operation are applied in the frequency domain:

ReLU $\cup$ *BN* $\cup$ *Conv1*$\times$*1*:$\mathbb{R}^{H\times \frac{W}{2}\times 2C} \rightarrow \mathbb{R}^{H\times \frac{W}{2}\times 2C}$;

c) The inverse FFT transform is applied to recover the spatial structure:

RealToComplex: $\mathbb{R}^{H\times \frac{W}{2}\times 2C} \rightarrow \mathbb{C}^{H\times \frac{W}{2}\times C}$,

Inverse Real FFT2d: $\mathbb{C}^{H\times \frac{W}{2}\times C} \rightarrow \mathbb{R}^{H\times W\times C}$.

As shown in Fig. 3, FFC leverages the fusion operation to combine the outputs of the local and global branch. Finally, two FFC units form a FFC residual block by connecting the input to the output. In our proposed method, we adopt three FFC residual blocks in total to make our network remove the haze from the perspective of considering the entire image or feature map.

3.4. Loss Function

We define the ground truth image as I^{gt}, and we denote the hazy image and the dehazed image as I^{hazy} and $\tilde{I}$, respectively. We utilize G and D to represent our proposed method and the discriminator.

Smooth L1 Loss. The smooth L1 Loss can be calculated using Eq. (4) and Eq. (5), where N denotes the total number of pixels, $I_i^{gt}(x)$ and $\tilde{I}_i(x)$ represent the strength of the pixel x in the i-th channel of the ground truth image and of the dehazed image.

$$\mathcal{L}_{\text{smooth}-\text{L1}} = \frac{1}{N}\sum_{x=1}^{N}\sum_{i=1}^{3} f(I_i^{gt}(x) - \tilde{I}_i(x)) \tag{4}$$

where

$$f(\gamma) = \begin{cases} 0.5\gamma^2 & \text{if } |\gamma| < 1 \\ |\gamma| - 0.5 & \text{otherwise} \end{cases} \tag{5}$$

Perceptual Loss. We introduce the VGG-16 [40] pretrained on ImageNet as the loss network ϕ. The perceptual loss can be defined as Eq. (6):

$$\mathcal{L}_{\text{PL}} = \sum_{j=1}^{3} \frac{1}{C_j H_j W_j} \left\| \phi_j(G(I_i^{hazy})) - \phi_j(I_i^{gt}) \right\|_2^2 \tag{6}$$

where ϕ_j denotes the activation of the j-th layer in the backbone network, and C_j, W_j and H_j represent the channel, width and height of the corresponding feature map. The ablation study can demonstrated that the perceptual loss appears to be crucial to our proposed method.

MS-SSIM Loss. Multi-scale Structure similarity (MS-SSIM) [15] is introduced in our loss function. We first calculate the SSIM for pixel i using Eq. (7):

$$\begin{aligned} SSIM(i) &= \frac{2\mu_D\mu_C + T_1}{\mu_D^2 + \mu_C^2 + T_1} \cdot \frac{2\sigma_{DC} + T_2}{\sigma_D^2 + \sigma_C^2 + T_2} \\ &= l(i) \cdot s(i) \end{aligned} \tag{7}$$

where T_1 and T_2 denote two small constants, D and C are two fixed size windows centered at current pixel in the reconstructed image and in the clear image, respectively. After applying Gaussian filters, we can compute the means μ_D, μ_C, standard deviations σ_D, σ_C and covariance σ_{DC}. The MS-SSIM loss is described as Eq. (8), where S represents the total number of scales, α and β are default parameters.

$$\mathcal{L}_{\text{MS}-\text{SSIM}} = 1 - \prod_{s=1}^{S} (l^{\alpha}(i) \cdot cs_s^{\beta_s}(i)) \tag{8}$$

Adversarial Loss. The adversarial loss is calculated as Eq. (9), where $D(G(I^{hazy}))$ represents the possibility that the recovered image is considered as a ground truth image by the discriminator [17].

$$\mathcal{L}_{\text{adv}} = 1 - \sum_{s=1}^{N} -logD(G(I^{hazy})) \tag{9}$$

Total Loss. The total loss used to supervise the training of our proposed method is shown as Eq. (10) :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{smooth}-\text{L1}} + \alpha\mathcal{L}_{\text{MS}-\text{SSIM}} + \beta\mathcal{L}_{\text{PL}} + \gamma\mathcal{L}_{\text{adv}} \tag{10}$$

where $\alpha = 0.2$, $\beta = 0.01$, $\gamma = 0.0005$ are the hyper-parameters for each loss function.

4. Experiments

In this section, we first introduce the datasets. Then, we discuss the experiment settings and evaluation criteria. Besides, we present the ablation study for our network. Moreover, we compare our method with state-of-art dehazing methods quantitatively and qualitatively. Finally, we introduce our dehazing results in NTIRE 2023 High-Resolution NonHomogeneous Dehazing Challenge.

4.1. Datasets

We conduct experiments utilizing RESIDE (Indoor Training Set, ITS) [24], NH-HAZE [2], NH-HAZE2 [6], and HD-NH-HAZE [7]. ITS encompasses more than ten thousands of training samples, and its Synthetic Objective Testing Set (SOTS) can be used for testing. NH-HAZE contains a total of 55 images. we employ the official testing data for model evaluation while the remaining are utilized for model training. NH-HAZE2 consists of a mere 25 training data, 5 validation data, and 5 test data. Given the fact that both the validation and test data remain undisclosed, we extract the first 20 images from the training data for training,

Methods	L_1	L_{PL}	L_{ssim}	L_D	PSNR	SSIM
(1) Single DWT branch	✓	✓	✓	✓	18.35	0.6491
(2) Single FFC branch	✓	✓	✓	✓	19.03	0.6834
(3) Single DWT-FFC branch	✓	✓	✓	✓	19.38	0.6991
(4) Prior knowledge branch	✓	✓	✓	✓	20.35	0.7161
(5) DWT-FFC branch+Res2Net	✓	✓	✓	✓	21.67	0.7301
(6) Ours	✓	✓	✓	✓	**22.20**	**0.7458**
(7) ours	✓	✓	✓		22.07	0.7427
(8) ours	✓	✓			21.81	0.7344
(9) ours	✓				21.71	0.7299

Table 1. Results of Ablation Study. The first raw is to study the necessity of each component of the network; The second raw is to illustrate the rationality of loss function used for training. Figures in **bold** denote the best results.

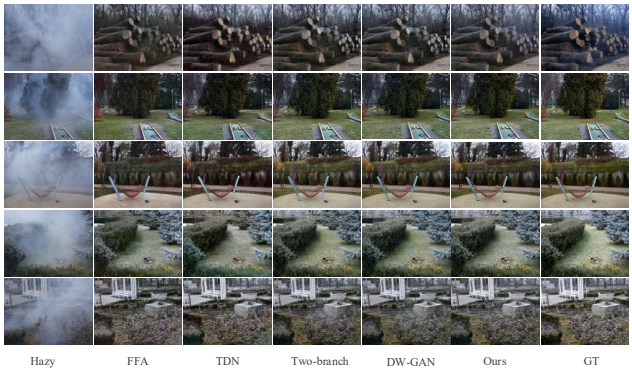

Figure 4. Comparison results on NH-HAZE.

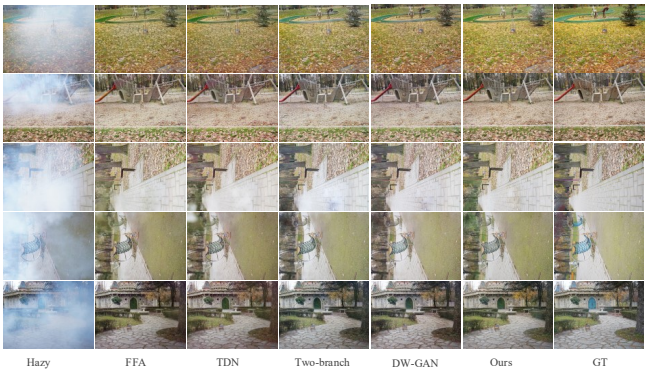

Figure 5. Comparison results on NH-HAZE2.

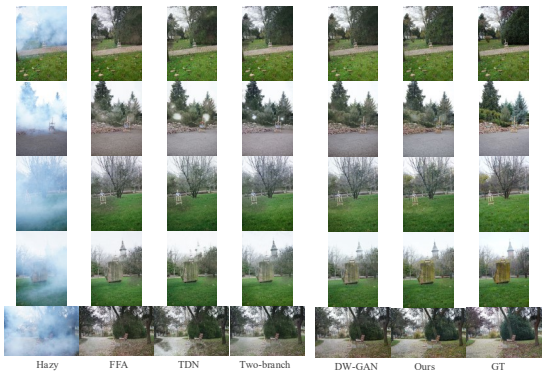

Figure 6. Comparison results on HD-NH-HAZE. Our results are nearly closed to the ground truth and are highly authentic with little noticeable dissonance.

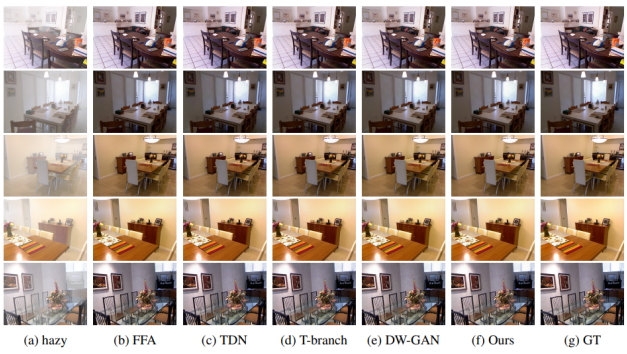

Figure 7. Comparison results on ITS. All methods generate visually pleasing dehazed images.

and utilize the remaining 5 images as testing samples. The HD-NH-HAZE dataset is characterized by dense and non-uniformly distributed hazy scenes. It comprises 40 training data, 5 validation data, and 5 testing data, each with a resolution of 6000×4000. Considering the validation and testing sets are not publicly available at present, we utilize images 1-35 as training data while images 36-40 are reserved for evaluation purposes. Besides, we have utilized gamma correction to build a combined dataset with 120 image pairs based on NH-HAZE, NH-HAZE2 and HD-NH-HAZE dataset (More details can be found in Sec. 4.5). We use the same 5 images with HD-NH-HAZE for testing, and train models using the rest 115 images.

Methods	ITS		NTIRE20		NTIRE21		HD-NH-HAZE		Combined Dataset	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FFA	36.55	0.9888	19.50	0.6441	20.56	0.8106	20.23	0.7103	20.43	0.7112
TDN	34.87	0.9804	20.73	0.6727	20.44	0.8014	20.06	0.7132	20.30	0.7140
Two-branch	<u>37.76</u>	<u>0.9905</u>	21.55	0.7149	21.75	0.8277	21.15	0.7323	21.39	0.7327
DW-GAN	36.33	0.9866	<u>21.69</u>	<u>0.7161</u>	<u>22.13</u>	<u>0.8314</u>	<u>21.52</u>	<u>0.7325</u>	<u>21.68</u>	<u>0.7371</u>
ours	**37.86**	**0.9907**	**22.64**	**0.7298**	**22.82**	**0.8738**	**22.20**	**0.7458**	**22.26**	**0.7469**

Table 2. Quantitative comparisons over ITS, NH-HAZE, NH-HAZE2, HD-NH-HAZE and Combined Dataset. Figures in **bold** denote the best results, and results with <u>double underline</u> represent the second best.

4.2. Experimental Settings

To diversify the training data for our model, we incorporate random cropping of patches with a size of 384×384, accompanied by random rotation at 90, 180, or 270 degrees, vertical flip and horizontal flip. The optimization process is facilitated by the Adam optimizer, with default values of β_1 and β_2 (0.9 and 0.999, respectively). Our training method features a specialized decay strategy, which initiates with a learning rate of e^{-4} and undergoes a decay of 0.5 times at 3000, 5000, and 8000 epochs, ultimately concluding after 10000 epochs. The discriminator uses the same optimizer and training strategies. To execute the experiments, we utilize one RTX 2080 Ti GPU. To undertake a quantitative assessment of our model efficacy, we employ two commonly used metrics, namely the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). Due to our limited GPU resource, our trained model cannot process an entire image from HD-NH-HAZE, we apply a block-based testing strategy to tackle this problem: 1) we first split the image into 9 blocks and the size of each block is 1600×2432; 2) we then output the dehazed results for each block using our trained weights; 3) we finally integrate these outputs into the final result of the same size of the hazy input by averaging the overlapping regions.

4.3. Ablation Study

We conduct comprehensive ablation studies, which aim to establish the indispensability of each component in our proposed methodology. Consistent with the principles of ablation, we devise and construct six distinct networks, each featuring a unique combination of modules. These networks serve to demonstrate the relative significance of each module and underscore the importance of the individual components in our overall framework. (1) Single DWT branch: only using three DWT downsampling and three DWT upsampling without FFC. (2) Single FFC branch: solely utilizing FFC modules, without the inclusion of DWT downsampling blocks or high-frequency skip connections. (3) Single DWT-FFC branch: combining DWT downsampling and upsampling with FFC modules as a single branch. (4) Prior knowledge branch: only using the transfer learning branch to restore defogging images. (5) DWT-FFC frequency branch and Res2Net: The only distinction between this particular network and our final methodology lies in the encoder of the transfer learning branch. (6) Ours: Two-branches consists of DWT-FFC frequency branch and ConvNext based prior knowledge branch.

By comparing (3) to (1) and (2) in Tab. 1 respectively, we can conclude that both DWT and FFC can help improve the dehazing performace in terms of PSNR and SSIM, and even FFC plays a more prominent role than DWT. This observation is consistent with our discussion for DWT (Sec. 3.2) and FFC (Sec. 3.3). Moreover, we can observe that the performance of our proposed two-branch network is far superior to that of any single branch (by comparing (6) to (3) and (4), respectively). This finding validate our initial intention for designing dual-branch network. Finally, we compare our model to another two-branch network in which the pretrained Res2Net weights are introduced. The increased PSNR and SSIM between (6) and (5) in Tab. 1 can demonstrate the effectiveness of pretrained ConvNeXt weights. Besides, to verify the rationality of the total loss function used in training, we train our model with several combinations of loss functions introduced in Sec. 3.4. As shown (6)-(9) in Tab. 1, our model can achieve the best performance using the loss function defined in Eq. (10), discarding any loss could worsen our model in terms of PSNR and SSIM.

4.4. Comparisons with State-of-art Models

In this section, we compare our proposed method with state-of-art models quantitatively and qualitatively. These methods include the winner solution in NTIRE 2020 NonHomogeneous Dehazing Challenge (TDN [27]), the winner solution (DW-GAN) and a two branch dehazing method via ensemble learning (abbreviated as Two-branch in this paper) in NTIRE 2021 NonHomogeneous Dehazing Challenge and FFA [49].

Quantitative Comparison Results. The comparison results are provided in Tab. 2. The performance of our method

Figure 8. The test result of our method on NTIRE 2023 HR NonHomogeneous Dehazing Challenge.

greatly exceeds that of other models on all datasets. Specially, our model surpasses the second-ranked method by an average of 0.725 dB for PSNR across these four non-homogeneous datasets. On the other hand, on ITS, the superority of our method compared to other models is not obvious (just 0.09 dB higher than Two-branch in terms of PSNR). The reason for this phenomenon lies on the number of training data and complexity of hazy patterns. ITS is a large-scale synthetic dataset. Considering the availability of training data that exceeds 100 times than other four datasets, the contribution of the pretained ConvNeXt weights to the whole network has been attenuated. Besides, the haze in ITS is not as complicated as that in other datasets, the advantages of DWT and FFC residual blocks which are designed to address non-homogeneous and dense haze cannot be fully exploited.

Qualitative Comparison Results. In this section, we compare the performance of our model to other methods from the perceptual and visual standpoint. The comparison results are shown as Fig. 7, Fig. 4, Fig. 5 and Fig. 6. For ITS, all methods can generate dehazed images that are visually pleasing and quite close to the ground truths. Large volumes of training data and relatively simple haze patterns enable these advanced models learn the feature and color mapping from hazy to clear images. For non-homogeneous datasets, the employment of distinct methods results in discernible visual discrepancies. Specifically, on NH-HAZE, blurred results generated by FFA and TDN indicate the limited ability of these two methods for non-homogeneous dehazing task, while Two-branch and DW-GAN methods suffer from a certain degree of chromatic and structural displacement. On NH-HAZE2, TDN performs comparably bad due to the presence of haze in the results, FFA, Two branch and DW-GAN methods can produce results that are highly close to the ground truth, but the unpleasing color shift still exists in some area, like the grass in the first test image and the road in the last image. With respect to HD-NH-HAZE dataset, there are several noticeable defects in the images processed by FFA, TDN and Two-branch method: the existence of white holes that are both meaningless and highly discordant. DW-GAN shows its image recovering ability based on extremely hazy images, however, the results contain several blurred local regions and exhibits noticeable fragmentation caused by block-based testing strategies. Generally, over all datasets, even though our method may produce artifacts or cannot remove haze on some small areas, like the trees in the first test image of NH-HAZE2 and in the last picture of HD-NH-HAZE, the color and structural details in vast majority of image regions are very nearly closed to the ground truth, and the results are highly authentic with little noticeable dissonance.

4.5. NTIRE2023 HR NonHomogeneous Challenge

Combined Dataset based on NH-HAZE, NH-HAZE2 and HD-NH-HAZE. The challenge only provides 40 training samples, which are not abundant for the training process. In order to increase the number of training data and improve the generalization ability, we apply channel-wise gamma correction for NH-HAZE and NH-HAZE2, respectively. Specifically, in each channel of these two datasets, the gamma correction with a unique gamma parameter has been utilized to make the average intensity of this channel much closer to that of HD-NH-HAZE. By achieving this, we assume these two datasets share similiar haze patterns with HD-NH-HAZE and the network trained on the combined datasets may perform better than that trained sorely on HD-NH-HAZE. The results in Tab. 2 can indicate this tendency and verify the effectiveness of this pre-processing method.

Performance on NTIRE 2023 High-Resolution Non-Homogeneous Dehazing Challenge. According to the report [7], our model is the winner of the challenge and is one of the top perceptual quality approaches in terms of PSNR (22.87) and SSIM (0.71). The test results of our model are illustrated in Fig. 8, which indicates our model features advanced capabilities in effectively removing haze, yielding visually appealing outputs with consistent structure.

5. Conclusions

In this paper, we propose a novel two branch network for high-resolution image dehazing. DWT-FFC frequency branch takes advantage of DWT to extract intricate high-frequency features while simultaneously employing FFC as a wide receptive field to enhance the perceptual quality of the output. The prior knowledge branch serves as providing supplementary information to address over-fitting on small-scale datasets, thereby improving the model generalization capability. Extensive empirical evaluations conclusively indicate that our model exhibits impressive performance on both synthetic datasets and real-world scenes. Moreover, our model surpasses the latest state-of-the-art techniques with superior fidelity and perceptual quality.

References

[1] C.O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte. A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE International Conference on Image Processing (ICIP)*, pages 1014–1018, 2019. 2

[2] C.O. Ancuti, C. Ancuti, and R. Timofte. A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE international conference on image processing (ICIP)*, pages 1014–1018, 2019. 2, 5

[3] C. Ancuti, C.O. Ancuti, R. Timofte, et al. Ntire 2018 challenge on image dehazing: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 891–901, 2018. 1

[4] C.O. Ancuti, C. Ancuti, R. Timofte, et al. Ntire 2019 image dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2241–2253, 2019. 1

[5] C.O. Ancuti, C. Ancuti, F.A. Vasluianu, R. Timofte, et al. Ntire 2020 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 1, 2, 3

[6] C.O. Ancuti, C. Ancuti, F.A. Vasluianu, R. Timofte, et al. Ntire 2021 nonhomogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1, 2, 3, 5

[7] 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. 5, 8

[8] D. Berman, T. Treibitz, and S. Avidan. Non-local image dehazing. In *CVPR*, pages 1674–1682, 2016. 2

[9] B. Cai, X. Xu, K. Jia, C. Qing, and D.Tao. An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 1, 2

[10] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua. Gated context aggregation network for image dehazing and deraining. In *IEEE winter conference on applications of computer vision*, pages 1375–1383, 2019. 1, 2

[11] L. Chi, B. Jiang, and Y. Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 2, 3

[12] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[13] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M. Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2157–2167, 2020. 1, 2

[14] A. Dosovitskiy, A. Kolesnikov L. Beyer, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3

[15] M. Fu, H. Liu, Y. Yu, J. Chen, and K. Wang. Dw-gan:adiscrete wavelet transform gan for nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 203–212, 2021. 1, 2, 3, 5

[16] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. Toll. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021. 1, 2

[17] I. Goodfellow, P. A. Jean, M. Mehdi, B. Xu, W.F. David, O. Sherjil, C. Aaron, and B. Yoshua. Generative adversarial networks. *Association for Computing Machinery*, 63(11):139–144, 2020. 1, 5

[18] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 2

[19] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2016. 1, 2

[20] K. He, X. Zhang, S.Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[21] M. Ju, Z. Gu, and D. Zhang. Single image haze removal based on the improved atmospheric scattering model. *Neurocomputing*, 260:180–191, 2017. 1

[22] L.Dai, X. Liu, C. Li, and J. Chen. Awnet: Attentive wavelet network for image isp. In *European Conference on Computer Vision Workshop (ECCVW)*, 2020. 2

[23] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. Single image haze removal using dark channel prior. In *CVPR*, pages 4770–4778, 2017. 1, 2

[24] B. Li, w. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2019. 5

[25] J. Li, Z. Pan, Q. Liu, and Z.Wang. Stacked u-shape network with channel-wise attention for salient object detection. *IEEE TRANSACTIONS ON MULTIMEDIA*, 23:1397–1409, 2020. 2

[26] Y. Li, Q. Miao, J. Song, Y. Quan, and W. Li. Single image haze removal based on haze physical characteristics and adaptive sky region detection. *Neurocomputing*, 182:221–234, 2016. 1

[27] J. Liu, H. Wu, Y. Xie, Y. Qu, and L. Ma. Trident dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 1, 7

[28] X. Liu, Y. Ma, Z. Shi, and J. Chen. Griddehazenet: Attention based multi-scale network for image dehazing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1

[29] X. Liu, Z. Shi, Z. Wu, J. Chen, and G. Zhai. Griddehazenet+: An enhanced multi-scale network with intra-task knowledge transfer for single image dehazing. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 1

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3

[31] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 2, 3

[32] Z. Lu, J. Jiang, J. Huang, G. Wu, and X. Liu. Glama: Joint spatial and frequency loss for general image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1301–1310, 2022. 2

[33] E.J. McCartney. *Optics of the atmosphere: scattering by molecules and particles*. 1976. 2

[34] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan. Efficient image dehazing with boundary constraint and contextual regularization. In *Proceedings of the IEEE international conference on computer vision*, pages 617–624, 2013. 1

[35] W. E. K. Middleton. *Vision through the atmosphere*. University of Toronto Press, 1952. 1, 2

[36] P. Fischer O. Ronneberger and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, page 234–241, 2015. 4

[37] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 2, 3

[38] J. Mai Q. Zhu and L. Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE TRANSACTIONS ON MULTIMEDIA*, 24(11):3522–3533, 2015. 2

[39] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M. Yang. Gated fusion network for single image dehazing. In *CVPR*, pages 3253–3261, 2018. 1

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

[41] V. A. Sindagi, P. Oza, R. Yasarla, and V.M. Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *ECCV*, pages 763–780, 2020. 1

[42] X. Song, D. Zhou, W. Li, H. Ding, Y. Dai, and L. Zhang. Wsamf-net: Wavelet spatial attention-based multi-stream feedback network for single image dehazing. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):575–588, 2023. 2

[43] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 2, 3

[44] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on transfer learning. In *In Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks*, pages 270–279, 2018. 2, 3

[45] H. Wang, Y. Xu, Z. Wang, Y. Cai, L. Chen, and Y. Li. Centernet-auto: A multi-object visual detection algorithm for autonomous driving scenes based on improved centernet. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–11, 2023. 1

[46] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, and S. Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. 2023. 2

[47] H. Wu, J. Liu, Y. Xie, Y. Qu, and L. Ma. Knowledge transfer dehazing network for nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 1

[48] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2022. 1, 3

[49] Y. Bai X. Xie X. Qin, Z. Wang and H. Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11908–11915, 2020. 1, 2, 4, 7

[50] H. Yu, N. Zheng, M. Zhou, J. Huang, Z. Xiao, and F. Zhao. Frequency and spatial dual guidance for image dehazing. In *ECCV*, pages 181–198, 2022. 1, 3

[51] Y. Yu, H. Liu, M. Fu, J. Chen, X. Wang, and K. Wang. A two-branch neural network for non-homogeneous dehazing via ensemble learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 193–202, 2021. 4

[52] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14114–14123, 2021. 3

[53] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, Francis E.H. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 2

[54] H. Zhang and V.M. Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 1

[55] Z. Zhao, P. Zheng, S. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019. 1

[56] Z. Zhong, T. Shen, Y. Yang, Z. Lin, and C. Zhang. Joint sub-bands learning with clique structures for wavelet domain super-resolution. *Advances in neural information processing systems*, 31, 2018. 3

[57] Z. Zhou and Z. Shi. Cggan: a context-guided generative adversarial network for single image dehazing. *IET Image Processing*, 14(15):3982–3988, 2020. 1

[58] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2, 3