SDCNet:Spatially-Adaptive Deformable Convolution Networks for HR NonHomogeneous Dehazing

Yidi Liu Xingbo Wang Yurui Zhu Xueyang Fu* Zheng-Jun Zha
University of Science and Technology of China
{liuyidi2023, wxb864557356, zyr}@mail.ustc.edu.cn {xyfu, zhazj}@ustc.edu.cn

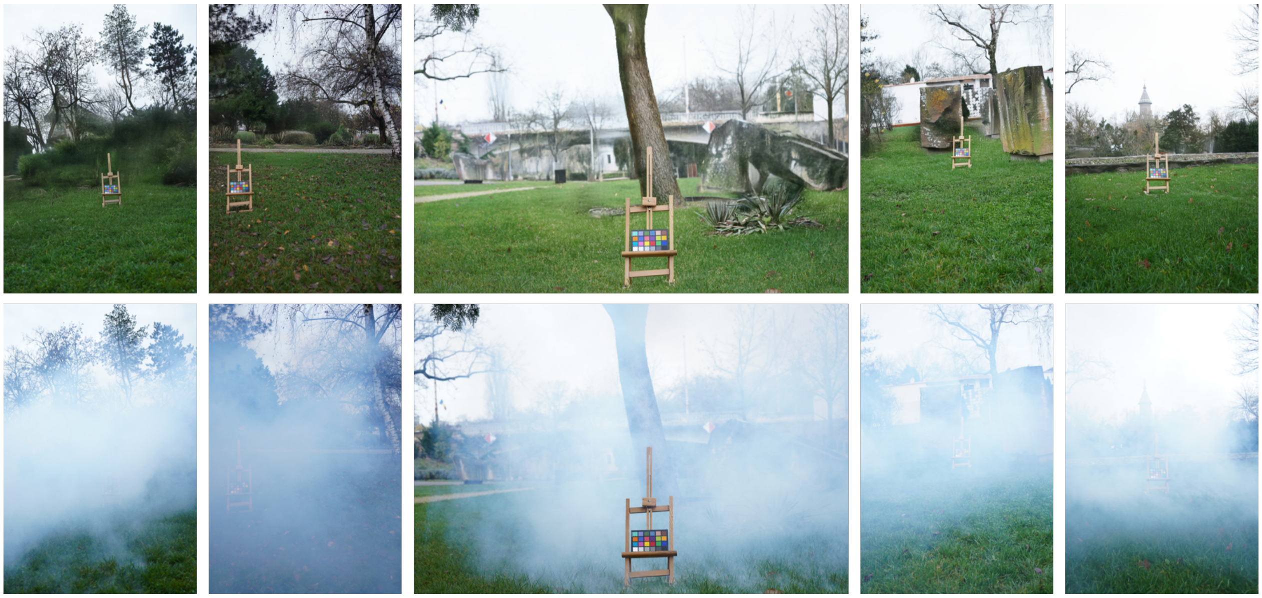

Figure 1. Our results on the NTIRE 2024 Dense and NonHomogeneous Dehazing Challenge, achieving the best performance in terms of PSNR, SSIM, and MOS. Ultimately, our solution was emerged as the champions of this challenge

Abstract

*In recent years, the field of image dehazing has garnered increasing attention. Many deep learning models have demonstrated exceptional capabilities in removing homogeneous haze, yet they often perform suboptimally when faced with the challenge of non-homogeneous dehazing. One of the primary issues is that these models are trained under conditions of homogeneous haze, which does not align with the characteristics of real-world haze scenarios. non-homogeneous haze typically leads to structural distortion and color shifts in images. Another contributing factor is the limited scale of datasets available for non-homogeneous dehazing, which hampers the training of robust models.To address these challenges, we have designed a Spatially-Adaptive Deformable Convolution Networks (**SDCNet**). The first branch of our model incorporates a high-level prior model that serves as an encoder for extracting high-level features from the image. The second branch is composed of a lightweight network specifically tailored to extract low-level features from hazy images. Our model fuses the information from both branches and combines progressive training as well as dynamic data augmentation strategies to obtain visually pleasing dehaze results. Extensive ablation studies have been conducted, substantiating the effectiveness and feasibility of our proposed methodology. Furthermore, in the NTIRE 2024 Dense and NonHomogeneous Dehazing Challenge, we achieved the best performance in terms of PSNR, SSIM, and MOS.*

1. Introduction

In recent years, deep learning has gained rapid development and wide application [36][9][15][31][11][1]. In the current landscape of image dehazing research, nonhomogeneous dehazing has emerged as a prominently investigated subtask. Traditional deep learning models excel in removing nonhomogeneous haze, yet they often exhibit suboptimal performance when confronted with nonhomogeneous dehazing challenges. One of the primary issues lies in the fact that these models are trained under conditions of homogeneous haze, which diverges from real-world haze scenarios. Nonhomogeneous haze typically induces structural distortion and color shifts in images, which can degrade image quality and undermine the judgments of intelligent systems such as tracking [24], satellite remote sensing [20], and object detection [23]. This underscores the significance of image dehazing as a crucial low-level visual task, prompting the development of numerous approaches to tackle this challenge [5, 7, 8, 10, 12–14, 22].

Among these methods, some are developed based on the early proposed atmospheric scattering model (ASM) [18]. This model utilizes Eq.(1) to establish the relationship between hazy images and their clear counterparts, where I and J represent the hazy image and its clear counterpart, respectively, and x indicates the pixel position. A denotes the global atmosphere light, and $t(x)$ is the transmission map determined by the atmosphere scattering parameter β and the scene depth $d(x)$ as described in Eq.(2)

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

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

Since the ASM model is based on the assumption of haze homogenization, it is not intended for nonhomogeneous dehazing.Therefore,as regards the latter, mostly models are used to learn image-to-image mappings directly [1–3]. However, such methodologies often necessitate extensive data to facilitate CNNs in learning these mappings effectively. Given the scarcity of data, a vast number of two-branch architectures have been proposed to introduce a pretrained prior [7, 14, 35].

Our model is based on a two-branch architecture, first for transfer learning branch, we utilized the Flash InternImage [28]. Upon comparison with notable classification networks such as Convnext [17], Swin Transformer [16], VMamba [15], and UniRepLKNet [6], we observed that Flash InternImage, which incorporates Deformable Convolution v4 (DCNv4), demonstrates superior and more rapid long-range modeling capabilities, along with adaptive spatial aggregation. This improved speed and efficiency substantially enhance the network's dehazing capabilities.

Considering the high resolution (6000*4000) of the data for this challenge, for fine-detail extraction branch, we use a lightweight model, Spatially-Adaptive Feature Modulation(SAFMN) [25]. This decision stemmed from a profound understanding of SAFMN's superiority in feature fusion. By introducing selective attention mechanisms, dynamically fuses features from different levels and enhances the model's perception of crucial information.

Another challenge of this task is the scarcity of data samples, which often leads the model to encounter overfitting issues, although the problem is somewhat mitigated by the introduction of the two-branch architecture, it still restricts the model performance from further improvement.We use the method in [27] to introduce synthetic haze data , and propose a dynamic data enhancement strategy to control the ratio of synthetic data to real data. The above strategy effectively alleviates the dilemma of few training samples.

Compared to the traditional VGG perceptual loss, we introduce EfficientVit-SAM [33] as a feature extractor to construct a novel enhanced perceptual loss. This loss reduces the output haze residue to a greater extent.

In summary, our contributions are outlined as follows:

1. We introduced Flash InternImage [28] and SAFMN [25] into the two-branch architecture, achieving a favorable balance between performance and efficiency.

2. We proposed a dynamic data augmentation strategy to enhance model generalization and an enhanced perceptual loss to improve the visual quality of the output.

3. We presented extensive experimental results and comprehensive ablation analyses to illustrate the superiority of the dual-branch framework in addressing the challenges posed by nonhomogeneous dehazing.

2. Related Works

2.1. Single Image Dehazing.

In the study of single-image dehazing, some classical methods are primarily based on physical models of images and statistical techniques. These methods typically model the imaging process and features of hazy images, attempting to estimate the degree of haze and global illumination conditions within the image. Subsequently, they perform dehazing processing based on the estimated results. However, these methods often require significant prior knowledge and handcrafted features, and they make strict assumptions about the input images, making them difficult to adapt to different scenes and complex situations[1].

In the early stages of deep learning-based methods, ASM was commonly employed. For instance, DehazeNet [5]was devised, which employed a CNN model to estimate the medium transmission map and then utilized it through ASM to obtain the dehazed image. Subsequently, AOD-Net [10] was introduced, which concurrently estimated atmospheric light and transmission maps to generate the restored image. These methods typically necessitated substantial prior knowledge and manually designed features, and they imposed stringent assumptions on the input images, making it challenging to adapt to various scenes and complex sce-

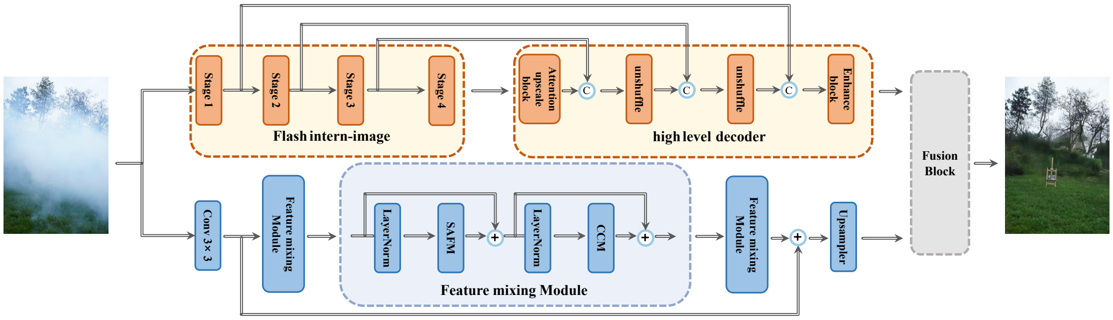

Figure 2. An overview of our network. The model consists of two branches. The transfer learning branch is composed by Swin Transformer based model. The data fitting branch consists of residual channel attention groups.

narios. Early networks often employed CNNs to estimate parameters in the degradation model. Later on, in order to minimize errors more effectively, researchers proposed some end-to-end networks to directly map hazy images to clean images [7, 35]. Many of the aforementioned methods performed quite well on synthesized ideal data, but exhibited poor performance when confronted with real-world non-homogeneous haze.

2.2. Transfer Learning.

Transfer learning aims to enhance the capabilities of a target model on a specific task by leveraging knowledge from related but different tasks, thus reducing the reliance on large volumes of data in the target domain [21, 38]. Some existing methods utilize extensive prior knowledge acquired through pre-trained models such as ImageNet, which is employed to aid image restoration tasks. For instance, in the latest two relevant competitions in NTIRE2023 [3], the champions and runners-up respectively utilized the Swin Transformer [14] and ConvNeXt [35] models as foundational blocks for knowledge transfer, effectively mitigating overfitting issues. In contrast, the Flash-InternImage [28] model we utilize is based on deformable convolutions (DCNv4), enabling the model to possess the required large receptive field while also allowing for adaptive spatial aggregation based on input and task information. By reducing the inductive bias of traditional CNNs, DCNv4 can learn more powerful and robust patterns from extensive data.Meanwhile, to the best of our knowledge, there remains considerable unexplored territory for DCNv4 in the field of image reconstruction.

3. Proposed Method

In this section, we begin by outlining the architecture of our proposed network, depicted in Fig. 2, which integrates Fine-detail Extraction Branch, and Transfer Learning branch.

3.1. Network Framework

Numerous techniques featuring dual branches have demonstrated significant achievements in the NTIRE 2020,2021 and 2023 NonHomogeneous dehazing challenge [1–3]. Inspired by these observations, we devised a neural network with two branches, as illustrated in Fig. 2. We obtain outputs at the original resolution separately from two branches, and then feed them into a fusion module with a simple design. This module consists solely of an 11×11 convolutional layer followed by a tanh activation function.

Transfer Learning branch. The purpose of this branch is to leverage the pretrained network to provide more prior information for the few-sample data. Firstly, we introduce a transfer learning branch based on the Flash intern-image_base [28] backbone and utilize its pretrained weights on ImageNet. This backbone integrates the Deformable Convolution v4 (DCNv4) operator, resulting in significant improvements in both speed and accuracy.

Drawing inspiration from the dual-branch design used in previous dehazing competitions, we employ the Flash intern-image to extract features from input images. Specifically, we utilize the outputs from its first four stages to form a multi-scale encoder [7, 14].

In the decoder, we adopt the same design as [14]. Multiple upsampling layers are utilized, with each containing a pixel-shuffle block and an attention module. The pixel-shuffle blocks are introduced to reduce the computational load and gradually restore the size of the feature maps to the original resolution. Meanwhile, attention blocks enable our model to discern dynamic hazy patterns.

Fine-detail Extraction Branch. For the low-level branch, we aim to learn more bottom-level features of the image through this branch. The structure of this branch is primarily modified from [25], as illustrated in Fig. 3.

Compared to self-attention mechanisms or large-kernel convolutions, SAFM serves as a lightweight alternative to learn long-range dependencies from multi-scale feature rep-

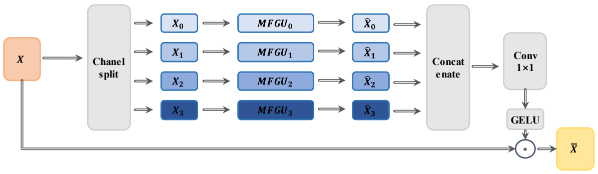

Figure 3. Network architecture of the SAFM module.

resentations, thereby better exploring more useful features for HR image reconstruction.

SAFM applies a feature pyramid to generate spatially-adaptive feature modulation attention maps. To reduce model complexity and obtain a pyramid-style feature representation, the normalized input features undergo a channel-wise splitting operation, resulting in four components. A 3×3 depthwise convolution processes the first component, while the remaining parts are fed into a multi-scale feature generation unit. Since SAFM aims to select discriminative features when learning non-local interactions, adaptive Maxpooling is applied to the input features to generate multi-scale features. This procedure can be formulated given the input feature X as:

$$\begin{aligned}&[X_0, X_1, X_2, X_3] = \mathrm{Split}(X),\\&\hat{X}_0 = \mathrm{DW-Conv}_{3\times3}\left(X_0\right),\\&\hat{X}_i = \uparrow_p \left(\mathrm{DW-Conv}_{3\times3}\left(\downarrow_{\frac{p}{2^i}}\left(X_i\right)\right)\right), 1 \leq i \leq 3,\end{aligned} \tag{3}$$

where,Split$(\cdot)$ denotes the channel split operation, DW-Conv3×3$(\cdot)$ represents a 3×3 depth-wise convolution, $\uparrow_p(\cdot)$ signifies upsampling features to the original resolution p via nearest interpolation for efficient implementation, and $\downarrow_{p/2^i}$ indicates pooling the input features to the size of $p/2^i$.

Subsequently, the short- or long-range features are combined by concatenating them along the channel dimension and being subjected to a 1×1 convolution. Following this, normalization is applied using a GELU non-linearity to estimate the attention map, which is utilized to adaptively modulate the input X through element-wise product.This procedure can be formulated as:

$$\begin{aligned}\hat{X} &= \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left(\left[\hat{X}_0, \hat{X}_1, \hat{X}_2, \hat{X}_3\right]\right)\right),\\\bar{X} &= \phi(\hat{X}) \odot X,\end{aligned} \tag{4}$$

where Concat$(\cdot)$ represents the concatenation operation, and Conv1×1$(\cdot)$ signifies the 1×1 convolution. $\phi()$ denotes the GELU function, and $\odot$ denotes the element-wise product.

To also incorporate local contextual information and enable channel mixing, a compact convolutional channel mixer (CCM) based on FMBConv [26] is introduced. The CCM comprises a 3×3 convolution to encode local contexts and double the channels for mixing, followed by a 1×1 convolution to reduce channels back to the original dimension. A GELU function is applied for non-linear mapping.The SAFM and CCM can be formulated as:

$$\begin{aligned}Y &= \mathrm{SAFM}(\mathrm{LN}(X)) + X,\\Z &= \mathrm{CCM}(\mathrm{LN}(Y)) + Y,\end{aligned} \tag{5}$$

where LN$(\cdot)$ represents the LayerNorm layer, with X, Y, and Z denoting the intermediate features. The incorporation of additional residual learning aims to stabilize the training process and learn high-frequency details, thereby facilitating high-quality image reconstruction.More details can be found in [25].

3.2. Dynamic Data Augmentation

Given the limited dataset in this challenge, ensuring that the model learns robust generalization from few data samples is a worthwhile consideration. To address the data-hungry issue, we follow the approach proposed in [27] and implement a dynamic data augmentation strategy. The specific process of the haze synthesis model proposed in [27] is as follows:

$$I(x) = JPEG\left(\mathcal{P}\left(J(x)^{\gamma} + \mathcal{N}, e^{\beta \hat{d}(x)}, A + \Delta A\right)\right). \tag{6}$$

- For simulating poor light conditions commonly found in hazy weather, we adjusted the brightness adjustment factor γ to range from 1.3 to 1.7, and introduced Gaussian noise distribution $\mathcal{N}$.
- The transmission map, a crucial parameter in our degradation model, is controlled by $beta \in [0.8, 1.7]$. Furthermore, to utilize more accurate depth estimation, we replaced the previous RA-depth method [19] with the latest Depth Anything [30] method for depth estimation. The results indicate that Depth Anything [30] exhibits stronger generalization performance and zero-shot capability, thereby providing more accurate depth estimation results for subsequent haze synthesis tasks.
- To introduce diversity in hazy images, we considered the color bias of atmospheric light, represented by a three-channel vector $\Delta A \in [-0.025, 0.025]$, with A ranging from 0.8 to 1.0.
- Since the hazy images in the training dataset do not exhibit prominent JPEG artifacts, we opt not to use JPEG noise augmentation as proposed in [27].

We use clean samples from the training dataset as backgrounds for generating synthetic data. A comparison of the synthetic data example with the real data is shown in the Fig. 4, it can be seen that although we have tried to mimic the style of real hazy conditions as much as possible in the synthesized results, there is still a significant gap. The fundamental reason is that the haze synthesis model can only generate homogeneous haze, whereas the haze in this challenge is non-homogeneous. Therefore, to mitigate

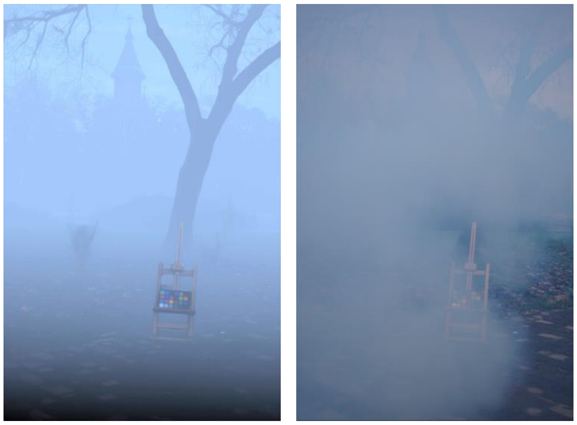

Figure 4. The comparison figure between synthetic hazy data (left part) and real hazy data(right part).

the data-hungry issue using synthesized data while preventing further interference with the model's understanding of non-homogeneous haze, we propose a dynamic data augmentation strategy.

In addition, for training our model, we employ the progressive training setting from Restormer [31], which is an efficient training method that gradually increases patch size and decreases batch size during the training process. Therefore, considering this training approach, we propose to increase the amount of synthesized data in the early stages of network training and gradually decrease it as the training progresses. Moreover, since the patch size is smaller at the beginning of training, the gap between synthesized and real data is relatively small, which can better address the issue of network overfitting. Thus, we utilize this dynamic data augmentation strategy to control the proportion of synthesized data added, gradually reducing it to zero as the progressive training strategy proceeds.

3.3. Loss Functions

We designate the ground truth image as I_{gt}, and refer to the hazy image and the dehazed image as I_{hazy} and I_{res}, respectively. To represent our proposed method and the discriminator, we use G and D respectively.Regarding the design of the loss function, we mainly refer to [14].

Smooth L1 Loss. The smooth L1 Loss is computed using Eq. (7) and Eq. (8), where N denotes the total number of pixels, $I_{\text{gt}}^{(i)}(x)$ and $I_{res}^{(i)}(x)$ represent the intensity of pixel x in the i-th channel of the ground truth image and the dehazed image, respectively.

$$\mathcal{L}_{\text{smooth-L1}} = \frac{1}{N} \sum_{x=1}^{N} \sum_{i=1}^{3} f\left(I_{\text{gt}}^{(i)}(x) - I_{res}^{(i)}(x)\right), \tag{7}$$

where

$$f(\gamma) = \begin{cases} 0.5\gamma^2 & \text{if } |\gamma| < 1 \\ |\gamma| - 0.5 & \text{otherwise} . \end{cases} \tag{8}$$

Enhanced Perceptual Loss. In previous image restoration tasks, pretrained VGG-16 on ImageNet is commonly used as a feature extractor to compute perceptual loss. However, in this challenge, most images exhibit heavy haze distribution, resulting in residual haze remaining in the reconstructed images. The feature extraction capability of VGG-16 is insufficient to fully capture these perceptual differences. Therefore, we attempt to replace the backbone of the perceptual loss with a backbone possessing stronger feature extraction capabilities and a larger-scale visual pretraining. We believe that such a model could better guide dehazing models to enhance their perceptual capabilities, representing an indirect form of ”knowledge distillation”. Moreover, it would only be used during the training phase, thus not impacting the model's inference efficiency.

This brings us to SAM [9], a visual large model trained on a massive dataset for segmentation tasks, endowed with powerful feature extraction capabilities and generalization. It could be a suitable candidate as a backbone for computing perceptual loss. Although SAM is designed for segmentation tasks, we can still extract its intermediate feature maps to compute perceptual loss between clean and hazy images.

However, directly using SAM would encounter a series of training efficiency issues due to its slow inference speed, greatly slowing down the training time of dehazing models. To address these issues, there are currently a series of works focused on accelerating SAM [33, 34]. For EfficientViT-SAM [33], the method leverages EfficientViT to accelerate SAM. Specifically, EfficientViT-SAM retains the prompt encoder and mask decoder architecture of SAM while replacing the image encoder with EfficientViT, thereby striking a balance between speed and performance. Therefore, we adopt EfficientViT-SAM as the feature extractor for computing perceptual loss. The calculation of perceptual loss is as follows:

$$\mathcal{L}_{\text{EPL}} = \sum_{j=1}^{6} \frac{1}{C_j H_j W_j} \left\| \phi_j\left(I_i^{\text{res}}\right) - \phi_j\left(I_i^{gt}\right) \right\|_2^2 , \tag{9}$$

where φ_j denotes the activation of the j-th stage including the final fused output in the Efficient-VIT, and C_j, W_j, and H_j represent the channel, width, and height of the corresponding feature map. Subsequent ablation experiments can demonstrate its impact on both objective metrics and subjective perception.

MS-SSIM Loss. We incorporate the Multi-scale Structural Similarity (MSSSIM) [7] into our loss function. Initially, we compute the SSIM for pixel i using Eq. (10):

$$\begin{aligned} \text{SSIM}(i) &= \frac{2\mu_D\mu_C + T_1}{\mu_D^2 + \mu_C^2 + T_1} \cdot \frac{2\sigma_{DC} + T_2}{\sigma_D^2 + \sigma_C^2 + T_2} \\ &= l(i) \cdot s(i), \end{aligned} \tag{10}$$

In the given context, T_1 and T_2 represent two small constants, while D and C denote two fixed-size windows centered at the current pixel in the reconstructed image and the clear image, respectively. Following the application of Gaussian filters, we can compute the means μ_D, μ_C, standard deviations σ_D, σ_C, and covariance σ_{DC}. The Multi-Scale Structural Similarity (MS-SSIM) loss is defined in Eq. (11), where S represents the total number of scales, and α and β are default parameters.

$$\mathcal{L}_{\mathrm{MS-SSIM}} = 1 - \prod_{s=1}^{S} \left(l^{\alpha}(i) \cdot c_s^{\beta_s}(i) \right). \tag{11}$$

Adversarial Loss. To address the limitations of pixel-wise loss functions in offering adequate supervision, particularly when training on a small dataset, we incorporate the adversarial loss [37]:

$$L_{adv} = \sum_{n=1}^{N} -\log D\left(I_{ref}\right), \tag{12}$$

where I_{ref} denotes the dehazed image. $D()$ represents the discriminator.

Focal-frequency-loss. We use focal frequency loss to reduce the gap between ground truth and dehazed images. This loss allows the model to adaptively focus on difficult-to-synthesize frequency components by reducing the weight of easily synthesizable components. This objective function serves as a complement to existing spatial loss and provides significant resistance to loss of important frequency information due to inherent biases of neural networks.The calculation of FFL loss is as follows:

$$\begin{gathered} F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}, \\ w(u,v) = \left|F_r(u,v) - F_f(u,v)\right|^{\alpha}, \\ \mathrm{FFL} = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} w(u,v) \left|F_r(u,v) - F_f(u,v)\right|, \end{gathered} \tag{13}$$

where $F(u,v)$ represents the discrete 2D Fourier Transform, $w(u,v)$ is the weight for the spatial frequency at (u,v), and α is the scaling factor for flexibility. In our experiments, we set α to 1.

Total Loss. The total loss used for supervising the training of our proposed method is formulated as Eq. (14):

$$L_{\text{total}} = L_{\text{smooth-L1}} + \lambda_1 L_{\text{MS-SSIM}} + \lambda_2 L_{\text{EPL}} + \lambda_3 L_{\text{adv}} + \lambda_4 L_{\text{FFL}}, \tag{14}$$

where $\lambda_1 = 0.2$, $\lambda_2 = 0.01$, $\lambda_3 = 0.002$, $\lambda_4 = 0.0005$ and $\lambda_5 = 0.001$ are hyperparameters for each loss function.

4. Experiments

In this section, we initially present the dataset utilized in our study. Subsequently, we delineate the experimental settings and evaluation metrics employed, as well as the experiments and ablation studies conducted. We will then compare our approach both qualitatively and quantitatively with state-of-the-art models. Finally, we showcase the results of our experiments in the NTIRE 2024 Dense and Non-Homogeneous Dehazing Challenge.

4.1. Datasets

The NH-HAZE24 dataset[4], continuing the legacy of its predecessors, includes 50 image pairs similar to the NH-HAZE 23 dataset [3], where 40 images are used as training set, 5 as validation set and 5 as test set. Each image boasts a high resolution of 4000 × 6000 pixels. We also utilized the train data from NTIRE 2020 [1], 2021 [2], 2023 [3] as an additional dataset for augmentation.

4.2. Implementation Details

We implement our proposed network via the PyTorch 1.8 platform. Adam optimizer with parameters β_1 = 0.9 and β_2= 0.99 is adopted to optimize our network.Additionally, motivated by [31], we introduce the progressive training strategy. The training phase of our network could be divided into two stages:

Initial Training Stage. We use progressive training strategy at first. We start training with patch size 128×128 and batch size 10 for 4000 epoch. The patch size and batch size pairs are updated to [(192^2, 8),(256^2, 6),(320^2, 6),(384^2, 4)] at epoch [4000,7000,9500,10500]. The initial learning rate is 1.2e-4,we employed a cosine annealing learning rate decay strategy, gradually reducing the learning rate to 2.5e-6. For data augmentation, we use our data augmentation mentioned above. The initial ratio of synthetic data is set to 0.5, and it has been updated to [0.4,0.2,0.1,0] at [4000,7000,95000,10500] respectively. The first stage performs on the NVIDIA 4090 device. We obtain the best model at this stage as the initialization of the second stage.

Finetune Training Stage. We start training with patch size 512×512 and batch size 2. The initial learning rate is 1e-5 and changes with Cosine Annealing scheme to 1e-7, including 800 epoch in total. We use the entire real data without any data augmentation technologies. Exponential Moving Average (EMA) is applied for the dynamic adjustment of model parameters. The second stage performs on the NVIDIA 4090 device.

4.3. Ablation Study

We conducted numerous ablation experiments to validate the effectiveness of our proposed approach, with performance metrics evaluated based on PSNR and SSIM on the NTIRE 2024 validation set.

The effectiveness of the dual-branch approach. First, we conducted comparative experiments to validate the effectiveness of the dual-branch approach. The results are shown in Tab. 1. Firstly, we compared the performance of the individual low-level branches. Both results were unsatisfactory,

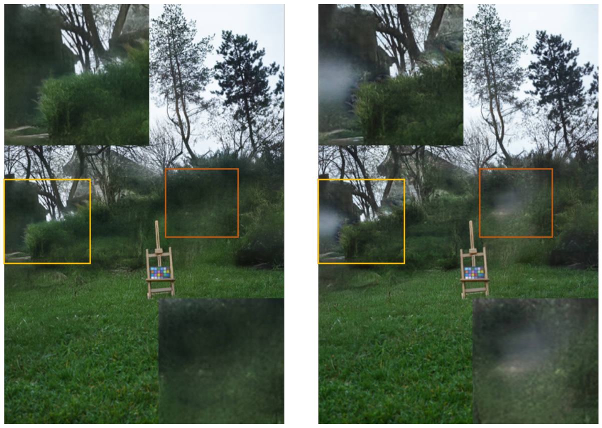

Figure 5. The comparison figure between the use of EPL loss(left part)and PL loss(right part).

but SAFMN [25]exhibited some advantages over RCAN [32]. Next, we compared the performance of the individual high-level branches. It can be observed that the performance of the high-level branch is improved compared to the low-level branch. We attribute this improvement to the pretrained weights, which help alleviate overfitting to some extent in the case of few samples. Meanwhile, we observe a further performance improvement of Flash InternImage [28] compared to InternImage [29]. Finally, when we combined both branches (Flash InternImage + SAFMN), we observed a significant improvement in performance. This further validates the effectiveness of the dual-branch approach in the context of few samples.

Table 1. The effectiveness of the dual-branch approach

Methods	PSNR	SSIM
Single RCAN branch	16.23	0.571
Single SAFMN branch	17.63	0.602
Single InternImage branch	19.36	0.642
Single Flash InternImage branch	20.45	0.674
Ours	**23.28**	**0.719**

The effectiveness of the loss terms. As the effectiveness of $L_{\text{smooth-L1}}$, L_{adv} and $L_{\text{MS-SSIM}}$ loss functions in non-homogeneous dehazing has been extensively validated in numerous works and reports, we refer to these loss functions as base-loss. Due to space constraints, our ablation experiments primarily focus on investigating the effects of L_{PL}, L_{EPL}, and L_{FFL} loss functions. Here, L_{PL} represents the perceptual loss using VGG-16.

As shown in Tab. 2, L_{FFL}can significantly improve objective metrics by supervising the frequency consistency between outputs and ground truth, effectively complementing spatial information. Also perceived loss L_{PL} as well as L_{EPL} can lead to performance gains. Looking at the objective metrics it seems that L_{EPL} only gives a marginal gain compared to L_{PL}, but higher objective metrics do not always represent cleaner haze removal. As illustrated in Fig. 5, upon comparison, we observe that after incorporating L_{EPL}, we achieve significantly cleaner dehazing results, substantially reducing the residue left in the dehazed images. However, if both perceptual losses are used at the same time, the performance will be relatively degraded, so we use L_{EPL} as the perceptual loss.

Table 2. The effectiveness of the loss terms

	base-loss	L_{FFL}	L_{PL}	L_{EPL}	PSNR	SSIM
1	✓				22.52	0.691
2	✓	✓			23.08	0.714
3	✓	✓	✓		23.24	0.716
4 (Ours)	✓	✓		✓	**23.28**	**0.719**
5	✓	✓	✓	✓	23.19	0.715

The effectiveness of the training strategy. We perform ablation experiments on the two training strategies employed, progressive training and dynamic data augmentation strategy. The results are shown in Tab. 3, with progressive training compared to fixed patch training , the performance goes up and the training is more efficient.

Meanwhile, regarding data augmentation, for fixed data augmentation, we fix the synthetic data ratio at 0.3, while the dynamic data augmentation synthetic data ratio gradually decreases from 0.5 to 0 with the training epoch.From the results, it can be seen that fixed data augmentation can only bring a marginal improvement, while dynamic data augmentation can bring a large performance gain. Therefore, we combine the two training strategies to obtain further improved results.

Table 3. The effectiveness of the training strategy

Methods	PSNR	SSIM
Fixed patch training (baseline)	22.73	0.682
Progressive training	22.91	0.690
Fixed data augmentation	22.76	0.681
Dynamic data augmentation	23.12	0.708
Ours	**23.28**	**0.719**

4.4. Comparisons with State-of-art Models

We consider two SOTA models, DWT-FFC [35], and ITB-Dehaze [14], which have the best composite score and the best objective metric for NITRE 2023 [3], respectively. We trained our method as well as the SOTA method following exactly the same training setup and compared them on the NITRE 2024 test set. Objective metrics are shown in Tab. 4, we've gotten a significant boost in objective metrics.

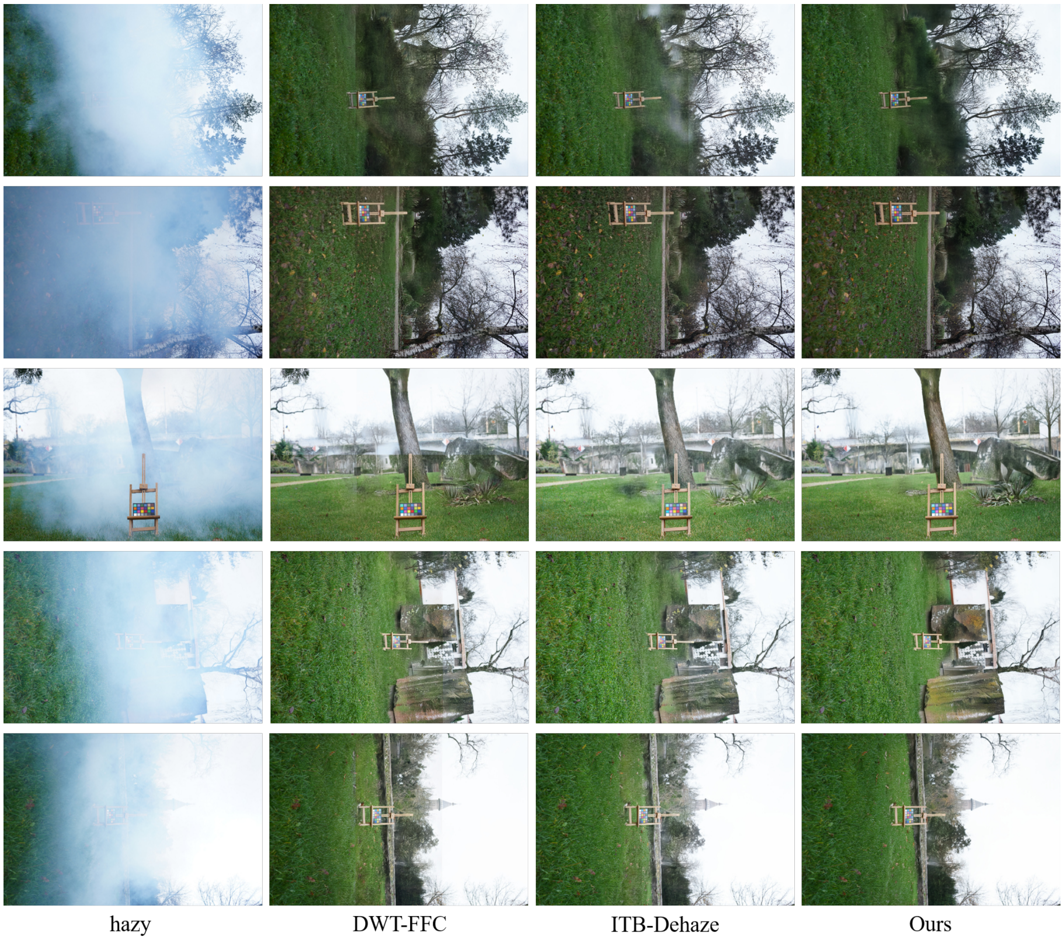

Figure 6. Comparisons with State-of-art Model on NTIRE 2024 test set

Meanwhile, the visual comparison is shown in Fig. 6, where our method similarly achieves the best in terms of perceptual quality, with our results having the least amount of haze residuals as well as boundary artifacts.

Table 4. Comparisons with State-of-art Model on NTIRE 2024 test set

Methods	PSNR	SSIM
ITB-Dehaze [14]	22.23	0.721
DWT-FFC [35]	22.08	0.718
Ours	**22.94**	**0.729**

4.5. NTIRE 2024 Dense and Non-Homogeneous Dehazing Challenge

Our method won the NTIRE 2024 [4] championship, and we simultaneously achieved the best in three metrics: PSNR(22.94), SSIM(0.729), and MOS(6.315). Our test results are shown in Fig. 1, yielding visually pleasing results.

5. Conclusion

In this paper, we propose a novel dual-branch network approach to address the issue of non-homogeneous fog removal. Our method leverages a lightweight model to learn the mapping between hazy and clean images, while introducing a prior knowledge branch to complementarily address overfitting on small-scale datasets, thereby enhancing the model's generalization capability. A dynamic data augmentation strategy is also proposed to further address data scarcity as well as an enhanced perceptual loss to improve the visual quality of the output. Extensive empirical evaluations demonstrate the impressive performance of our proposed model in real-world scenarios. The model surpasses state-of-the-art techniques, exhibiting outstanding fidelity and perceptual quality.

References

[1] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Jing Liu, Wu, et al. Ntire 2020 challenge on nonhomogeneous dehazing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2029–2044, 2020. 1, 2, 3, 6

[2] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Minghan Fu, Liu, et al. Ntire 2021 nonhomogeneous dehazing challenge report. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 627–646, 2021. 6

[3] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Han Zhou, Wei Dong, Yangyi Liu, Jun Chen, Yangyi Liu, Huan Liu, Li, et al. Ntire 2023 hr nonhomogeneous dehazing challenge report. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1808–1825, 2023. 2, 3, 6, 7

[4] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Han Zhou, Wei Dong, Yangyi Liu, Jun Chen, Yangyi Liu, Huan Liu, Li, et al. Ntire 2024 dense and nonhomogeneous dehazing challenge report. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 6, 8

[5] 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. 2

[6] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition, 2024. 2

[7] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 203–212, 2021. 2, 3, 5

[8] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2010. 2

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 5

[10] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4770–4778, 2017. 2

[11] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8222–8232, 2023. 1

[12] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2022. 2

[13] Jing Liu, Haiyan Wu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Trident dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[14] Yangyi Liu, Huan Liu, Liangyan Li, Zijun Wu, and Jun Chen. A data-centric solution to nonhomogeneous dehazing via vision transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1406–1415, 2023. 2, 3, 5, 7, 8

[15] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024. 1, 2

[16] 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)*, pages 10012–10022, 2021. 2

[17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 2

[18] W. E. K. Middleton. *Vision Through the Atmosphere*. University of Toronto Press, 1952. 2

[19] He Mu, Hui Le, Bian Yikai, Ren Jian, Xie Jin, and Yang Jian. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *ECCV*, 2022. 4

[20] Weiping Ni, Xinbo Gao, and Ying Wang. Single satellite image dehazing via linear intensity transformation and local property analysis. *Neurocomputing*, 175:25–39, 2016. 2

[21] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 3

[22] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2018. 2

[23] Vishwanath A. Sindagi, Pranav Oza, Ravi Yasarla, and Vishal M. Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *European Conference on Computer Vision*, pages 763–780, 2020. 2

[24] Dilbag Singh and Vijay Kumar. A comprehensive review of computational dehazing techniques. *Archives of Computational Methods in Engineering*, 26(5):1395–1413, 2019. SN - 1886-1784. 2

[25] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13144–13153, 2023. 2, 3, 4, 7

[26] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 4

[27] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22282–22291, 2023. 2, 4

[28] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications, 2024. 2, 3, 7

[29] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. *arXiv preprint arXiv:2401.06197*, 2024. 7

[30] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 4

[31] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2022. 1, 5, 6

[32] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 7

[33] Zhuoyang Zhang, Han Cai, and Song Han. Efficientvit-sam: Accelerated segment anything model without performance loss. *arXiv preprint arXiv:2402.05008*, 2024. 2, 5

[34] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. 5

[35] Han Zhou, Wei Dong, Yangyi Liu, and Jun Chen. Breaking through the haze: An advanced non-homogeneous dehazing method based on fast fourier convolution and convnext. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1895–1904, 2023. 2, 3, 7, 8

[36] Jiaying Zhu, Dong Li, Xueyang Fu, Gang Yang, Jie Huang, Aiping Liu, and Zheng-Jun Zha. Learning discriminative noise guidance for image forgery detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7739–7747, 2024. 1

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. 6

[38] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. 3