Two Stage Dehazing Framework for Dense and Non-Homogeneous Dehazing

Wei Song, Yichang Gao, Jiahao Xiong, Hualiang Lin, Dong Li*, Yun Zhang*
Guangdong University of Technology, Guangzhou, China
{*sonwe, 2112204059, 2112304352, 2112304356* }*@mail2.gdut.edu.cn*, {*dong.li, yz*}*@gdut.edu.cn*,

Abstract

In real-world environments, haze often causes a decrease in visibility, leading to potentially severe consequences. Although current methods based on the assumption of homogeneous haze density have achieved commendable results, dehazing techniques for non-homogeneous density haze still fall short in terms of visibility restoration and color accuracy. We observe that although single-stage methods have made significant strides, a multi-stage enhancement can further improve dehazing in terms of both visibility and color restoration. In this paper, we propose a two-stage dehaze framework, named Two Stage Dehazing Framework. Our approach consists of a DehazeNet, which does not require specifying a particular model for dehazing and can accept a hazy image as input, producing an clear image of the same dimensions as the original. Two such DehazeNet are sequentially connected to form the final serial DehazeNet. Moreover, to better approximate the output image to real-world scenes, we propose the Multi-Scale Attention Head. Our method achieved third place in NTIRE 2024 Dense and NonHomogeneous Dehazing Challenge, demonstrating outstanding performance metrics in the Peak Signal to Noise Ratio (PSNR), the Structral Similarity Index (SSIM), and the Mean Opinion Score (MOS). Related code will be acailable on code.

1. Introduction

Visibility degradation due to atmospheric conditions, such as haze, significantly impacts various applications, ranging from autonomous driving [12] to outdoor surveillance systems [9]. While dehazing techniques aim to mitigate these effects by restoring the original appearance of images, the variability and unpredictability of the haze density in different scenes add complexity to the dehazing process. To address this issue, recently, a significant amount of research has been conducted in the field of computer vision targeting non-homogeneous dehazing [1–5, 24, 40].

Traditional dehazing methods [6, 8, 13, 22] often rely on the assumption of a globally homogeneous haze density, which simplifies the model, but does not accurately reflect real-world conditions where the haze density can vary significantly within a single image. Homogeneous dehazing methods are primarily based on the atmospheric scattering model [28], which can be formulated:

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

In this study, I and J denote the hazy image and its corresponding haze-free image, respectively. The term x signifies the pixel coordinates. The global atmospheric light is represented by A, and $t(x)$, the transmission map, is determined by the atmospheric scattering coefficient β and the scene depth $d(x)$, given by the equation:

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

Upon assuming the atmospheric scattering model holds true, the task of image dehazing simplifies to estimating the transmission map $t(x)$ and the atmospheric light A. However, it is important to note that this model presumes a homogeneous hazy scene. Consequently, traditional dehazing methods based on the atmospheric scattering model paradigm is inadequate for tackling dehazing in non-homogeneous conditions.

Recent advancements in the field have increasingly leveraged deep learning techniques [7, 10, 17, 19–21, 31, 32, 41] to encapsulate the intricate interplay between light and atmospheric haze. Although these models demonstrate superior performance over traditional dehazing techniques, they are predominantly designed as single-stage networks. Despite their advancements, these networks often struggle to effectively handle highly variable haze densities.

To address these shortcomings, some recent methodologies adopt a dual-branch network structure [10, 24, 40]. While these dual-branch approaches have indeed demonstrated commendable results, their integration strategies are

*Corresponding author.
†This work was supported by the Guangdong Basic and Applied Basic Research Foundation No. 2021A1515011867

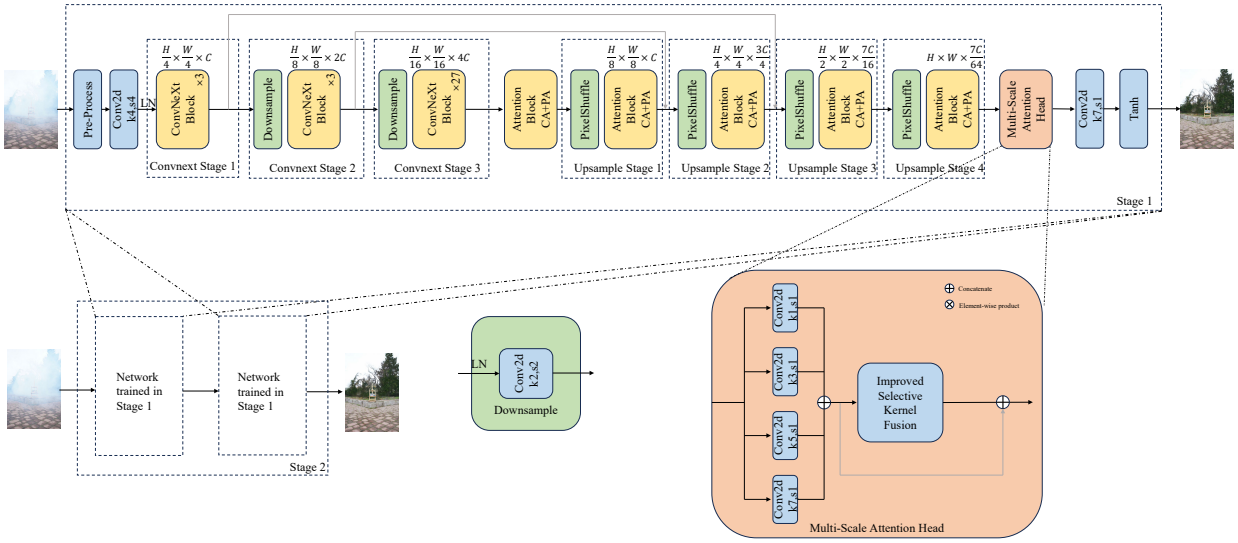

Figure 1. The network structure of our proposed method. The top part of the image is the dehazeing network we trained in the stage 1, the bottom left part of the image is the serial network trained in the stage 2, the bottom right part of the image is our proposed multi-scale attention head.

predominantly limited to the latter stages of processing. Such an approach does not fully exploit the potential of feature fusion between branches, leading to only marginal improvements. In the prevalent dual-branch network structures within the domain, it is common to observe a primary, more capable model working in conjunction with a secondary, less powerful one. This observation leads to a compelling inquiry: if the final stage involves merely combining the outputs of a robust and a weaker model, then one could contemplate the potential benefits of employing two high-performance models. In our experiments, we found that single-stage networks can remove most of the haze. However, they may not perform well in restoring color fidelity and texture details. Therefore, we propose dedicating the first model to executing an initial, comprehensive dehazing process, while the second model can focus on refining the output to restore and enhance image details. Adopting this strategy could improve the quality of the dehazed image significantly. It leverages the collective strengths of two advanced models in a sequential yet collaborative manner.

To address these challenges, we present the Two StageDehazing Framework, a two-stage dehazing framework that improves the state-of-the-art in image dehazing. Within this framework, we define a two-stage training approach that does not require specifying a particular model for dehazing. This flexible framework allows for superior dehazing performance compared to both single-stage versions of the network and dual-branch structures composed of the same network. Here, we refer to this unspecified dehazing network as DehazeNet. For best challenge performance, we incorporate a knowledge-prior encoder ConvNeXt-xLarge [26] and a multi-layer hybrid attention decoder of FFA-Net [30] as DehazeNet to effectively handle dense and non-homogeneous haze density. The use of a two-stage architecture enables refined image processing, resulting in superior visibility restoration and enhanced color accuracy. In addition, the Multi-Scale Attention Head is proposed to align the dehazed output with real-world visual perceptions, addressing a common criticism of existing dehazing methods. Our contributions are as follows:

- We introduce a Two Stage Dehazing Framework primarily designed to excel in dehazing. Based on this framework, we achieve superior dehazing performance compared to both single-stage versions of the network and dual-branch structures composed of the same network.

- To improve the restoration of image details and colors, our framework includes a Multi-Scale Attention Head for the final feature map output. This head integrates multi-scale convolutions [33] and Selective Kernel (SK) Fusion Attention [32] to capture an enriched set of spatial and channel information from the last feature.

- We conduct extensive experiments and ablation studies to justify the overall design and demonstrate its competitive performance.

2. Related Works

Single Image Dehazing. The task of single image dehazing, which presents a challenge in the domains of Computer Vision and Image Processing, has been extensively studied based on the haze model. Physical prior-based dehazing methods primarily depend on the physical scattering model, and to ensure performance, these methods require reasonable assumptions and understanding of hazy images to obtain accurate estimations of the transmission map and atmospheric light intensity in ASM modeling [27]. Furthermore, dark channel prior (DCP) [13], haze-lines [6], color-lines [8], rank-one prior [22], and color attention prior (CAP) [44] are also frequently employed in physical prior-based dehazing. However, due to the constrained applicability of the foundational assumptions, physical-based dehazing methods tend to be instability. With the rapid advancement of deep learning technology, deep learning has gained widespread application in the field of image dehazing in recent years. In the initial stages of deep learning-based approaches, ASM remains influential. For instance, DehazeNet [7] proposes to use a convolution network (CNN) model to predict the medium transmission map, subsequently utilizing it within ASM to produce a dehazed image. In addition, AOD-Net [17] concurrently estimates the atmospheric light and transmission map to generate the reconstructed image. Ren et al. [31] employ a fusion-based approach utilizing a multi-scale architecture in their framework for generating haze-free images. Zhang et al. [39] introduce a densely connected pyramid dehazing network (DCPDN), which predicts the transmission map through an edge-preserving densely connected encoder-decoder structure integrated with a multilevel pyramid pooling module.

Framework in Non-Homogeneous Dehazing. Integrating diverse models has been proven to effectively enhance the capabilities of neural networks. Within the context of non-uniform dehazing, dual-stream architectures are widely adopted. The DWT-FFC-GAN [40], comprising a DWT-FFC frequency branch and a prior knowledge branch, learns complementary information across branches to achieve superior generalization. Similarly, the model proposed by [24] consists of a Transfer Learning Branch and a Data Fitting Branch, whereas the model proposed by [37] a DWT branch and a ResNet Branch. Differently, TransER[14] introduces a two-stage deep network composed of two independent deep neural networks: the TransConv Fusion Dehaze (TFD) model, capable of generating two pseudo haze-free images, and the Lightweight Ensemble Reconstruction (LER) network, which merges the outputs of the two TFDs to produce the final haze-free image. Distinct from these methodologies, our approach employs a two-stage network with a two-stage training process.

Multi-scale Feature Fusion. Multi-scale feature fusion plays an important role in the field of computer vision. It is widely used to improve the performance of image processing tasks by making full use of the characteristics of images at different scales. One kind of work is to obtain images with different resolutions through multiple down-sampling operations, and then extract features separately, which is also called pyramid structure. It is often used in segmentation, object detection and other tasks, such as feature pyramid network [18] (FPN) and its variants[23, 36, 38]. Another kind is to use different sizes of convolution kernels to realize different sizes of receptive fields and achieve multi-scale feature extraction, such as InceptionNets [33–35]. Our proposed multi-scale attention head is similar to the basic module of Inception v1, which uses multi-scale convolution to extract features at different scales and Selective Kernel (SK) Fusion to fuse multi-scale features.

3. Proposed Method

Building upon our findings, our training procedure and network architecture present distinct variations from previous methods. Our approach employs a two-stage training process and utilizes a sequential network structure, as shown in Fig. 1.

3.1. Two-Stage Training Process

In the initial stages of our experimentation, we implement a single-model approach for dehazing. While iterative enhancements to the model yield improved dehazing results, it is apparent that some areas with denser haze remain uncleared, and color restoration in certain regions is suboptimal. According to current analyses of image restoration and the effectiveness of single-stage approaches on dense haze, we believe that it is possible to refine the single-stage results to achieve better outcomes.

Therefore, we propose a Two Stage Dehazing Framework. We begin by establishing a network with fundamental dehazing capabilities, DehazeNet, which does not require specifying a particular model for dehazing and can accept a hazy image as input, producing an clear image of the same dimensions as the original. In the first training stage, we train a single DehazeNet to equip it with basic dehazing proficiency. During this stage, we preserve the parameters that perform optimally on the validation set. In the second stage, two DehazeNets are serially connected, each preloaded with the best parameters from the first stage, to process a hazy image successively, resulting in the final dehazed image. The mean output of both DehazeNets is computed during this training process, and the loss is calculated by comparing it to a haze-free image.

3.2. ConvNeXt-xLarge Encoder

For the best challenge performance, we build on the work of DWT-FFT-GAN [40] and incorporate the prior knowledge branch as our DehazeNet. Our network follows the same

structure as DWT-FFT-GAN, including the first three stages of ConvNeXt [26], pre-trained weights from ImageNet1k, and the decoder from FFA-Net [30].

3.3. Multi-Scale Attention Head

Previous approaches often use a single convolution layer and Tanh function to finalize their outputs. However, this approach is inadequate in achieving optimal imaging results as a single convolution layer is insufficient in effectively representing different image scales. To address this issue, we propose a Multi-Scale Attention Head.

The decoder outputs features represented by Y, a tensor with dimensions $B \times C \times H \times W$. In this representation, B represents the batch size, C represents the feature channels, and H and W denote the height and width of the input image, respectively. The multi-scale head [33] produces four feature sets, denoted as $[X1, X2, X3, X4]$, using convolution kernels of different sizes $[1, 3, 5, 7]$. These sets are then fused using a modified SKFusion [32] process to create the aggregated feature X_{SK}. This modification involves replacing the original Squeeze and Excitation mechanism with the Efficient Channel Attention mechanism for improved feature integration, i.e. Improved SKFusion in Figure 1. Finally, the four feature sets $[X1, X2, X3, X4]$ and X_{SK} are concatenated to form the final feature set, which is then processed by a 7×7 convolution with padding of 3 and activated by a Tanh function.

3.4. Loss Functions

This section explores our loss functions of perceptual loss, GAN Loss [43], Multi-scale Structure Similarity (MS-SSIM) Loss and Smooth L1 Loss [11].

Perceptual Loss: Given the ConvNeXt encoder's difficulty in capturing global image details, an alternative approach was chosen using a swinv2-model Perceptual loss. This method calculates L1 losses for features at each downsampling stage, aiming to preserve a global receptive field in input patches. The perceptual loss can be defined as Eq. (3):

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

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.

GAN Loss: Previous work uses GAN loss [43] as a loss to enhance the detail of the recovered image. However, we notice that most of the discriminators employed in GAN-loss use a simple stack of convolutional layers, but we find that weaker models do not compete well against powerful generators, so we tried the simpler DenseNet [16], the powerful SwinV2 [25], and ConvNeXt [26]. In conclusion, it is discovered that while SwinV2 and ConvNeXt create a dual-stream discriminator that enhances the image, the training cost is too high. Therefore, DenseNet201 is ultimately chosen as the discriminator model. The formulation of GAN loss can be defined as:

$$\mathcal{L}_{\text{GAN}} = \sum_{n=1}^{N} -logD(f_\theta(x)), \tag{4}$$

where $f_\theta(x)$ denotes the dehazed image. $D()$ representsthe discriminator.

MS-SSIM Loss: The SSIM index for a given pixel i is defined as follows:

$$\begin{aligned} SSIM(i) &= \frac{2\mu_D\mu_C + T_1}{\mu_D^2 + \mu_C^2 + T_1} \cdot \frac{2\sigma_{DC} + T_2}{\sigma_D^2 + \sigma_C^2 + T_2} \\ &= l(i) \cdot s(i) \end{aligned} \tag{5}$$

where T_1 and T_2 represent two negligible constants to stabilize division with small denominators, D and C denote two windows of fixed size centered on the corresponding pixel in the reconstructed and clear images, respectively. By applying Gaussian filtering, the means μ_D, μ_C, standard deviations σ_D, σ_C, and the covariance σ_{DC} are computed. The MS-SSIM loss, as specified in Equation 5, accounts for multiple scales of structural similarity, where S denotes the total number of scales, and α and β are predefined parameters contributing to the loss calculation.

$$\mathcal{L}_{\text{MS-SSIM}} = 1 - \prod_{s=1}^{S} (l^{\alpha}(i) \cdot cs_s^{\beta_s}(i)) \tag{6}$$

This formulation elucidates the role of MS-SSIM in quantifying image quality through a comprehensive assessment of luminance, contrast, and structure at varying scales.

Smooth L1 Loss: The smooth L1 Loss can be calculated using Eq. (7) and Eq. (8), 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.

$$\mathcal{L}_{\text{smooth-L1}} = \frac{1}{N} \sum_{x=1}^{N} \sum_{i=1}^{3} f(I_i^{gt}(x) - \tilde{I}_i(x)) \tag{7}$$

where

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

Total loss: The total loss is the weighted sum of the above four components with predefined weights:

$$\begin{aligned} \mathcal{L}_{total} =& \mathcal{L}_{\text{smooth-L1}} + 0.5\mathcal{L}_{\text{MS-SSIM}} + 0.2\mathcal{L}_{\text{Perceptual}} \\ &+ 0.0005\mathcal{L}_{\text{GAN}} \end{aligned} \tag{9}$$

Figure 2. Our results on NTIRE 2024 dehazing challenge, achieving the outstanding performance in terms of PNSR, SSIM and LPIPS.

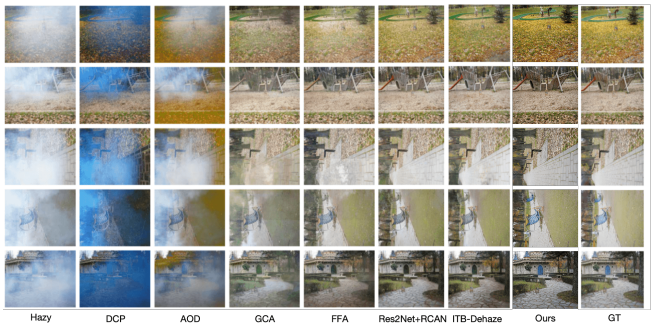

Figure 3. Qualitative evaluation on the NH-HAZE20 datasets

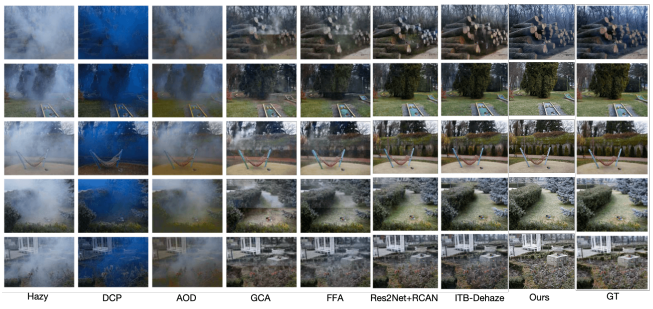

Figure 4. Qualitative evaluation on the datasets NH-HAZE21

4. Experiments

In this section, we first introduce the dataset used for our experiments, followed by the experimental setup, and then we present the ablation experiments we conducted. We then compared our results with some State-of-the-art Methods. Finally, we summarize our method's results and effects of NTIRE2024 HR Non-Homogeneous Challenge.

4.1. Datasets

NH-HAZE20 & NH-HAZE21. In the NTIRE2020 [2] and NTIRE2021 [3] challenges, NH-HAZE20 and NH-HAZE21 datasets were introduced, featuring non-homogeneous haze patterns. The images in these datasets are sized at 1600 $\times$ 1200 pixels. NH-HAZE20 comprises 55 samples and NH-HAZE21 comprises 25 samples. In this two dataset, we adopted the Data-Centric data pre-processing approach introduced by [24], that reduces the distribution gaps between the target dataset and the augmented one.

NH-HAZE23 & NH-HAZE24. Continuing the tradition of non-homogeneous haze styles, NTIRE2023 [4] introduces 50 image pairs, each boasting a significantly higher resolution of 4000 $\times$ 6000 pixels. The larger image size necessitates more extensive training data and increased computational resources. As the ground truth images for the 5 validation and 5 test samples have not been publicly released, we use only the 40 training pairs for evaluation outside the challenge server. Similar to NH-HAZE23, we can only use 40 image pairs of 4000 $\times$6000 pixels from NH-HAZE24 in NTIRE2024. Through experiments, it is found that the color distribution of NH-HAZE23 is basically the

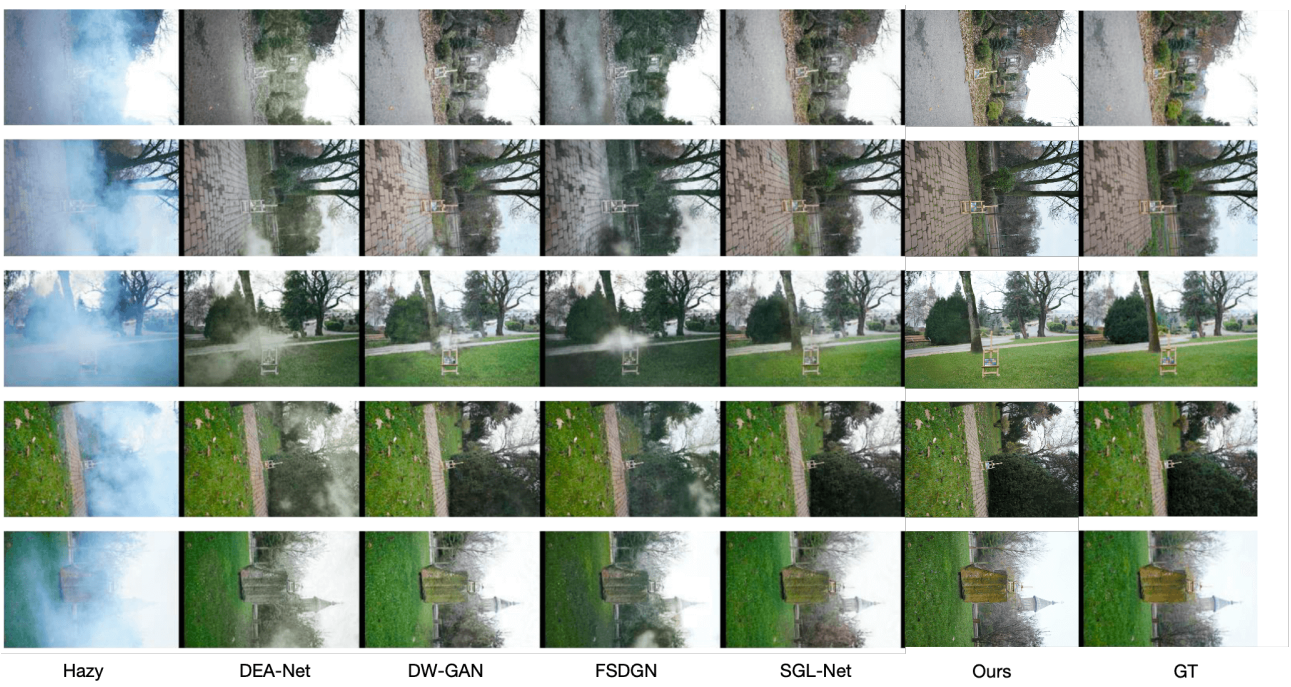

Figure 5. Qualitative evaluation on the datasets NH-HAZE23

same as that of NH-HAZE24, which means that we did not apply data pre-processing on NH-HAZE23.

4.2. Experimental Settings

In terms of dataset division, we take a total of 160 image pairs of NH-HAZE20, NH-HAZE21, NH-HAZE23 and NH-HAZE24 as the training set. In order to improve the proportion of NH-HAZE24, we resample NH-HAZE24, that is, the training set consists of 200 image pairs. In order to select the best model during the training process, we used the previous dehazing method to process the training set to select 20 samples with poor performance, that is hard samples, as the validation set.

First-Stage Training The input image is randomly cropped to a size of 256×256 and augmented by several data augmentation strategies, including random rotations of 90 degrees, 180 degrees, 270 degrees, horizontal flips, and vertical flips. In the first stage, we train the ConvNeXt-xLarge [26] as DehazeNet with AdamW [29] optimizer and batch size 3 for 4000 epochs to give it a basic dehazing capability. We set the base learning rate to 1e-4 and adopt a warmup schedule for 60 epochs and cosine schedule decay.

Second-Stage Training In the second stage, we connect two DehazeNet to obtain a serial network, the initial parameters of each DehazeNet is loaded from best checkpoint from the first stage. And the parameters of two model is not shared. Then we fine-tune the cascade model for 400 epochs to enhance its dehazing capability. Different from the first stage is that we reduce batchsize and LR to 1 and 5e-5, respectively. We adopt a sliding window-based testing strategy, specifically: 1) we use a 2000×2000 sliding window with 1500 steps to traverse the whole image to obtain 12 patches; 2) For the patches beyond the boundary, we create a border around the image like a photo frame using mirror reflection of the border elements; 3) All the patches are input to the network to obtain the output, and the overlapping parts are averaged to combine all the outputs to obtain the final result.

4.3. Ablation Study

In this section, we conduct comprehensive ablation experiments to verify the necessity of each component in the proposed method. Our ablation study is comprised of two components: The first part involves testing on the NH-HAZE24 validation set, where we iteratively add or modify components to the final model to demonstrate the relative importance of each component. The second part conducts tests on the NH-HAZE24 test set to validate the efficacy of our proposed two-stage training strategy.

Our ablation study is structured on foundational , resulting in the design of six distinct networks for the first part: (1) ConvNeXt-xLarge Encoder + FFA-Net Decoder + MultiScale Attention Head; (2) w/ more training epochs: increaing the number of epochs from 800 to 3200 in our experimental setup; (3) w/ total loss, building upon (2) by incorporating additional loss components beyond L1 Loss

Table 1. Results of Ablation Study.

Methods	PSNR	SSIM	FLOPs(G)
(1) ConvNeXt-xLarge Encoder + FFA-Net Decoder + MultiScale Attention Head	22.09	0.7061	-
(2) w/ more training epochs	22.74	0.7121	-
(3) w/ total loss	**23.15**	**0.7200**	-
(4) First stage	22.32	0.7173	99.93
(5) Two stage	**22.60**	**0.7268**	199.87
(6) Perform consecutive inferences using the first stage best network	15.67	0.5731	99.93

into the Total Loss calculation.

In the second part, we applied the conclusions drawn from the first part to the test set using network (1), thereby obtaining the results for (4) First Stage. We preserved the optimal parameters from the first stage and sequentially connected two identical networks, each preloaded with these optimal parameters from the first stage, thereby achieving the results presented in (5) Two stage. Comparing item (4) with (5) demonstrates the effectiveness of our proposed two-stage training strategy in Sec. 3.1. (6) Perform consecutive inferences using the first stage best network: Our findings indicate that utilizing a two stage approach with two networks produces different results compared to performing consecutive inferences using the same network. Furthermore, it is evident that the dehazing performance is notably worse when the same network is used for consecutive inferences.

Conventional wisdom might lead one to assume that utilizing two models, thereby having more parameters, would invariably result in superior performance. However, this is not necessarily the case. During the development of our final framework, we experimented with a variety of approaches, including increasing the number of layers in the FFA-Net decoder, constructing a dual-branch structure with two networks in parallel, and experimenting with our framework without pretraining and specific loss settings. These attempted solutions were not included in the ablation study table because they either failed to achieve the fundamental dehazing capability or required computational resources that were unfeasible to sustain. To explore how to train a large model composed of two cascaded sub-models, we analyzed the training strategies for this architecture. Specifically, we examined scenarios including the absence of pretrained model loading and the lack of supervision on the output from the first model, as illustrated in Figure 6. The Peak Signal to Noise Ratio (PSNR) metrics observed during the validation phase of the training process indicate that utilizing a pre-trained model and applying supervision to the outputs of the first model significantly enhance both the training efficiency and the performance metrics for dehazing tasks.

Table 2. MultiScale Attention Head Results of Ablation Study.

Methods	PSNR	SSIM
(1) Unetpp + Conv1x1 Head	18.76	0.6723
(2) Unetpp + MultiScale Head	19.84	0.6756
(3) Unetpp + MultiScale SKFusion Attention Head	20.34	0.6821
(4) Unetpp + MultiScale Attention Head	**21.57**	**0.6899**

To conduct the ablation study, we utilize Unetpp[42] with seresnext101[15] as encoder for our DehazeNet, maintaining the same training settings as previously described. The rationale for selecting this particular model is its effectiveness as our initial baseline during the early stages of the competition.

Multi-scale Kernel Design. In the design of the Multi-Scale Attention Head, critical investigation into the multi-scale kernel structure is detailed in Section 3.3 and summarized in Table 3. Initial experiments utilizing solely 3×3 kernels demonstrated inadequate performance, attributed to their limited capability in dehazing. Subsequently, an enhanced kernel architecture is implemented, featuring sizes ranging from 1×1 to 11×11 with an incremental stride of 2. This configuration yielded the highest performance improvements. Extended experimentation with exclusively larger kernels 7×7 led to a performance decrement of 0.41.

MultiScale Attention Head Design. We further explore the integration of SKFusion[32] and Inception-like[33] architectures within the Multi-Scale Attention Head, as detailed in Section 3.3 and quantitatively assessed in Table 2. Typically, a 1×1 kernel convolution serves as the baseline head for this task; however, its performance is notably limited, primarily due to deficiencies in color restoration capabilities. Our experiments compare three configurations: a standard 1×1 kernel convolution, a Multi-Scale Head, and a Multi-Scale Attention Head. The results clearly demonstrate that the Multi-Scale Attention Head surpasses the other designs in performance, confirming its effectiveness in enhancing model capabilities.

Table 3. Kernel Size of MultiScale Attention Head Results of Ablation Study.

Kernel Design	PSNR	SSIM	FLOPs(G)	Params(M)
(1) (1, 3, 5, 7)	**21.57**	**0.6899**	81.737	84.797
(2) (3, 5, 7, 9)	21.39	0.6877	87.106	84.89
(3) (5, 7, 9, 11)	21.34	0.6831	94.621	84.993
(4) (7, 7, 7, 7)	21.06	0.6811	89.253	84.911
(5) (3, 3, 3, 3)	20.74	0.6749	**78.515**	**84.748**

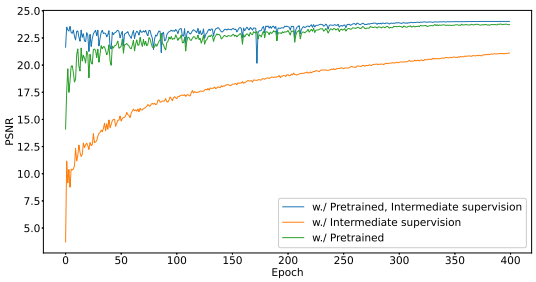

Figure 6. Validation PSNR bewtween different setting in two stage training. Pretrained: Indicates that the training of the two-stage model utilized the optimal parameters obtained from the first stage of training. Intermediate Supervision: Denotes that the output predictions from the preceding sub-model in the sequential architecture were also subject to loss-based supervision.

4.4. Comparison with State-of-the-art Methods

In our experiments, we compare our results with those of ITB-Dehaze[24] and SGL-Net[37], shown in Figure 3, 4, 5, as detailed in their respective papers. It's important to note that we did not retrain their models using our training set. Instead, we opted to use their published prediction results for comparison. This approach ensures that the comparison is fair and unbiased, as the showcased images used for demonstration were not included in our training dataset.

4.5. NTIRE2024 HR NonHomogeneous Challenge

We list the comparison of PSNR, SSIM, Learned Perceptual Image Patch Similarity (LPIPS) and Mean Opinion Score (MOS) of the proposed method with other methods in the competition according to [5] in Table 4. Our model is the third place of the challenge and is one of the top perceptual quality approaches in terms of PSNR (22.60) and SSIM (0.7268). The test results of our proposed model are shown in Fig. 2, which demonstrates the advanced ability of the model to effectively remove haze, producing visually appealing and structurally consistent outputs.

4.6. Limitations and Discussion

In our participation in the NTIRE 2024 Dense and Non-Homogeneous Dehazing Challenge, we strategically selected the DehazeNet constructed from ConvNeXt-xLarge, known for its robust dehazing capabilities, to achieve a higher PSNR score. While the implementation of our Two Stage Dehazing Framework on this DehazeNet yielded competitive results, we faced a limitation in timing. The results of the competition were announced too late for us to conduct a comprehensive evaluation of the performance across different types of DehazeNets.

Table 4. The average PSNR, Structural Similarity(SSIM), Learned Perceptual Image Patch Similarity (LPIPS) , MOS of top10 methods over NTIRE 2024 HR Non-Homogeneous Dehazing Challenge dataset. The best scores of each metrics are shown in bold, and the second scores of each metrics are shown in underline.

Team name	PSNR	SSIM	LPIPS	MOS
USTC-Dehazers	**22.94**	**0.729**	0.352	**6.315**
Dehazing R	<u>22.84</u>	0.725	0.347	<u>5.96</u>
ITB Dehaze	22.32	0.714	0.334	5.705
TTWT	21.93	0.714	0.334	5.675
DH-AISP	21.90	0.714	0.402	5.81
BU-Dehaze	21.68	0.709	<u>0.327</u>	5.22
RepD	21.78	0.706	0.333	4.83
PSU Team	20.54	0.632	**0.267**	5.31
xsourse	21.66	0.695	0.449	5.28
Team Woof	22.60	<u>0.726</u>	0.381	5.79

5. Conclusion

In this paper, we introduce an effective solution for addressing the challenge of dehazing in conditions characterized by deep and non-uniform fog. Our methodology comprises two primary components. Initially, we propose a Two Stage Dehazing Framework. The first stage involves the preliminary training of DehazeNet to attain a basic level of dehazing capability. The optimal parameters from DehazeNet are then saved and applied in the second stage, enabling the serially connected DehazeNets to both achieve effective dehazing performance. During the second stage, the models are fine-tuned to realize the expected dehazing results. Furthermore, we introduced the Multi-Scale Attention Head to enhance the network's ability for texture and color restoration. Our approach achieved competitive results in the NTIRE 2024 Dense and Non-Homogeneous Dehazing Challenge. For future work, we aim to investigate the operational mechanism of this framework further. Specifically, we will explore whether connecting two networks serially provides the same benefits as increasing the depth of a single network. Moreover, we plan to investigate the performance of this framework on other low-level vision tasks. This exploration will help us understand the broader applicability and potential of our proposed framework in tackling various challenges in the field of computer vision.

References

[1] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 444–445, 2020. 1

[2] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2020 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 490–491, 2020. 5

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2021 nonhomogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–646, 2021. 5

[4] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Han Zhou, Wei Dong, Yangyi Liu, Jun Chen, Huan Liu, Liangyan Li, et al. Ntire 2023 hr nonhomogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1808–1825, 2023. 5

[5] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2024 dense and nonhomogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 1, 8

[6] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1674–1682, 2016. 1, 3

[7] 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. 1, 3

[8] Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1):1–14, 2014. 1, 3

[9] Catarina Fontes, Ellen Hohma, Caitlin C Corrigan, and Christoph Lütge. Ai-powered public surveillance systems: why we (might) need them and how we want them. *Technology in Society*, 71:102137, 2022. 1

[10] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–212, 2021. 1

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[12] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020. 1

[13] 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. 1, 3

[14] Trung Hoang, Haichuan Zhang, Amirsaeed Yazdani, and Vishal Monga. Transer: Hybrid model and ensemble-based sequential learning for non-homogenous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1670–1679, 2023. 3

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 7

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[17] 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. 1, 3

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[19] 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 (CVPR)*, pages 5831–5840, 2022. 1

[20] 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.

[21] 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 Workshops*, pages 430–431, 2020. 1

[22] 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. 1, 3

[23] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[24] Yangyi Liu, Huan Liu, Liangyan Li, Zijun Wu, and Jun Chen. A data-centric solution to nonhomogeneous dehazing via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1406–1415, 2023. 1, 3, 5, 8

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 4

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. 2, 4, 6

[27] William Edgar Knowles Middleton. *Vision through the atmosphere*. University of Toronto Press, 1952. 3

[28] William Edgar Knowles Middleton. Vision through the atmosphere. In *geophysik ii/geophysics ii*, pages 254–287. Springer, 1957. 1

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[30] 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*, pages 11908–11915, 2020. 2, 4

[31] 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. 1, 3

[32] 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, 4, 7

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2, 3, 4, 7

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[35] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 3

[36] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[37] Hao-Hsiang Yang, I Chen, Chia-Hsuan Hsieh, Hua-En Chang, Yuan-Chun Chiang, Yi-Chung Chen, Zhi-Kai Huang, Wei-Ting Chen, Sy-Yen Kuo, et al. Semantic guidance learning for high-resolution non-homogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1447–1455, 2023. 3, 8

[38] Nianyin Zeng, Peishu Wu, Zidong Wang, Han Li, Weibo Liu, and Xiaohui Liu. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 3

[39] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2018. 3

[40] 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*, pages 1894–1903, 2023. 1, 3

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

[42] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 7

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 4

[44] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing*, 24(11):3522–3533, 2015. 3