A Dual Branch Multi-scale Image Dehazing Network Based on High Quality Codebook Priors

Xuehui Yin, Peixin Wu, Zehong Li, *Senior Member, IEEE*

***Abstract*—The existence of non-homogeneous haze can lead to image scene blurring, color distortion, contrast reduction, and the loss of other texture details. Current methods for removing non-homogeneous haze fail to effectively address this form of haze. We have observed that the available non-homogeneous haze datasets generally suffer from a limited number of samples (for instance, NH-HAZE-23 only contains 55 training sample pairs), and existing neural networks cannot effectively learn relevant priors for haze removal from such limited data. Inspired by the field of image generation, we propose a dual-branch multi-scale image dehazing network based on a high-quality codebook. Firstly, we utilize VQGAN to pretrain a discrete codebook reflecting the general detailed texture features of clear images on a large-scale high-quality clear image datasets, serving as a dehazing prior. Subsequently, we design a pyramid-structured encoder using dilated neighborhood attention to enhance feature extraction. And we introduce another branch to better preserve the original fine detail features of the image at different scales. Finally, we conduct extensive experiments and ablation studies to demonstrate the effectiveness of our proposed method.**

***Index Terms*—Image dehazing, codebook, a two-branch network, multi-scale, neighborhood attention.**

I. Introduction

HAZE is an aerosol system composed of numerous tiny water droplets suspended in the near-ground air, serving as one of the primary causes of image blurring, color distortion, and contrast reduction. The hazy image degradation model based on atmospheric light single scattering phenomenon[1] [2] can be represented as:

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

where $I(x)$ denotes the hazy image, and $J(x)$ is its corresponding ground truth. A is the environment lighting, and the transmission map is represented by $t(x) = e^{\beta d(x)}$ which depends on scene depth $d(x)$ and haze density coefficient β.

The outdoor visual systems such as autonomous driving, video surveillance, military reconnaissance, and remote sensing imagery are affected by hazy conditions, leading to a decrease in the accuracy of information acquisition from captured images. As the haze intensity increases and non-homogeneous haze emerges, the image quality deteriorates rapidly, resulting in color distortion, blurred features, reduced contrast, and other visual quality degradations. Consequently, it becomes challenging to identify objects and backgrounds within the image, significantly impacting the effectiveness of subsequent visual tasks such as semantic segmentation and object detection in computer vision. Therefore, it is necessary to preprocess the images to mitigate the impact of dense haze on image quality.

Although existing dehazing methods have made significant progress based on deep learning, image dehazing still remains a highly ill-posed problem that often requires the support of prior knowledge. Current dehazing methods have proposed priors such as dark channel[3] and color attenuation[4], most of which are based on empirical observations and manually set priors[5], [6]. Due to their reliance on specific scenarios and assumptions, they often perform poorly in complex scenes, resulting in sub-optimal outcomes. Furthermore, most current image dehazing methods tend to fail when dealing with dense haze areas or non-homogeneous haze in hazy images, leading to residual haze and incomplete dehazing in areas with large depth of field. This is primarily because most models assume that the haze in images is uniformly distributed, whereas in the real world, the concentration of haze may vary significantly in different locations.

To address these challenges, in this work, we propose a dual branch multi-scale image dehazing network based on a high-quality codebook. By leveraging codebook trained on high-quality clear datasets, we extract robust priors which preserve image textures and details. Through the proposed dilated neighborhood attention transformer encoder and the enhanced decoder, we accurately encode and decode the features of latent hazy images. Finally, a dual-branch network is utilized to integrate these methods, resulting in our proposed network.

The contributions of our work can be summarized as follows:

1) We employed the VQGAN generative model to train a high-quality codebook as a prior, thereby complementing robust prior knowledge and mitigating the impact of haze features in the dehazing network.

2) We designed a feature pyramid encoding module based on dilated neighborhood attention and an enhanced decoding module incorporating pixel-wise and channel-wise attention. By leveraging a dual-branch multi-scale network structure, we improved the feature extraction capability for regions with dense haze in images.

This work was supported in part by the National Natural Science Foundation of China under Grant 61701060, in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJQN202000619.

Xuehui Yin, Peixin Wu are with the School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: yinxh@cqupt.edu.cn; w1066365803@163.com).

Zehong Li is with the State Key Laboratory of Electronic Thin Films and Integrated Devices, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China, also with Shenzhen Institute for Advanced Study, UESTC, Shenzhen 518110, China, and also with Chongqing Institute of Microelectronics Industry Technology, UESTC, Chongqing 401331, China (e-mail: lizh@uestc.edu.cn).

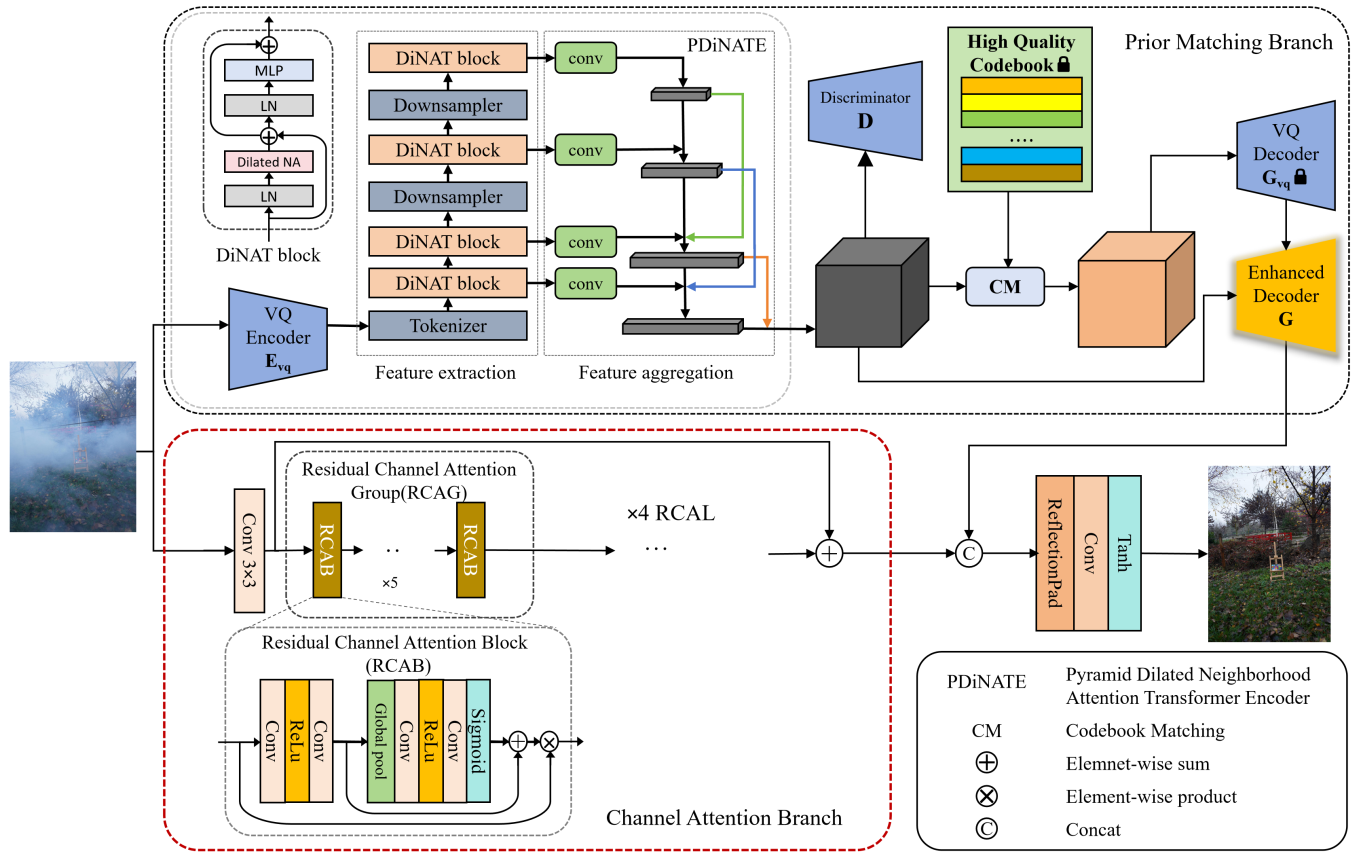

Fig. 1. An overview of our network. This model consists of two branches: prior matching branch and channel attention branch. Hazy image is processed separately by two branches, each outputting a feature map. Then, a feature fusion tail is used to fuse the feature maps of the two branches, and finally generate a hazy-free image.

3) We conducted extensive experiments to demonstrate the superiority of the proposed method compared to other existing methods.

II. RELATED WORK

A. Single Image Dehazing

In recent years, a significant amount of observations have been made on haze phenomena, and numerous prior knowledge has been proposed to assist in image dehazing. Among them, the dehazing method based on Dark Channel Prior (DCP), proposed by He et al. [3] in 2010, is widely known. Liu et al. [7] proposed the GridDehazeNet network structure, which through a unique grid-like structure and the use of attention mechanisms for multi-scale feature fusion, fully integrates low-level and high-level features. Qin et al. [8] eliminated the up-down sampling operation and proposed an end-to-end Feature Fusion Attention Network (FFA-Net) to directly restore haze-free images. The main idea of this method is to adaptively learn feature weights, giving more weight to important features. Feature attention is added after each residual block, and adaptive selection of features from each group is performed, enhancing the network's mapping capabilities. Yin et al. [9] propose a novel multiscale depth information fusion enhancement network to improve dehazing ability in scenes with large depth changes.

Transformer[10] was initially proposed for natural language processing tasks, capturing non-local interactions between words through the stacking of multi-head self-attention mechanisms and feed-forward layers. DeHamer [11] combined convolutional neural networks and Transformer for image dehazing, aggregating long-term attention in Transformer and local attention in convolutional neural network features. Dehazeformer [12] proposed an offset window partitioning scheme based on reflection padding and cropping, allowing the mask multi-head self-attention to discard part of the mask and achieve a constant window size.

However, there are currently two major challenges in applying Transformer to the image field. First is the large variation in visual entities, and the performance of Transformer may not be well in different scenes. Second is that the image resolution is high and there are many pixels, so the global self attention mechanism in Transformer leads to a large amount of computation.

In recent years, people have made a lot of achievements in heavy haze and non-homogeneous haze. Among them, TNN [13] introduced a dual-branch neural network, using Res2Net pre-trained on ImageNet [14] and Residual Channel Attention Network (RCAN) [15], and then through a learnable tail to fuse the features of the two branches, committed to solve non-homogeneous haze. FogRemoval [16] combines

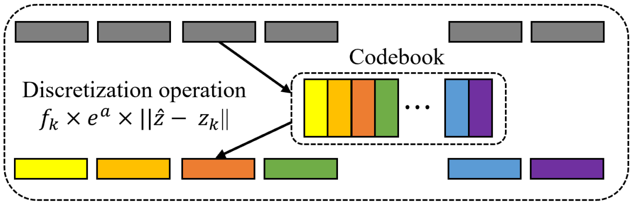

Fig. 2. The discretization operation of the codebook is shown in the figure. The feature map output by the encoder is discretized to form discrete encoded values, represented by gray rectangles. Each discrete encoded value corresponds to an embedding, and all embeddings are stored in an embedding space, which is the codebook. Using the nearest neighbor algorithm to obtain the true embedding in the codebook, represented as a colored rectangular prism, as input to the decoder.

the structural representation of ViT[17] and the features of CNN as feature regularization. It proposed a gray feature multiplier as a feature enhancement, guiding the network to learn to extract clear background information, and introduced uncertainty feedback learning, focusing on the area affected by haze. ITBDehaze [18] proposed a new network structure and a new data preprocessing method, applying RGB channel transformation on the enhanced datasets, and using Transformer as the backbone in the dual-branch network. Guo et al. [19] proposed SCANet, a network that adopts a mode of attention generation and scene reconstruction. It is an attention network capable of learning complex interactive features between non-homogeneous haze and image background.

B. Discrete Codebook

The concept of discrete codebooks can be traced back to Variational Auto Encoders (VAE)[20], which employ clever methods to constrain the encoding vector, making it conform to a standard normal distribution. Then the decoder in the trained encoder-decoder pair can recognize not only the vectors encoded by the encoder but also vectors from other standard normal distributions. However, VAE encode into continuous vectors, and Vector Quantized Variational AutoEncoders[21] believe that the image quality generated by VAE is not well because the images are encoded into continuous vectors, which better match the feature distribution of different objects in nature. To solve this problem, scholars have drawn on natural language processing to add a word embedding layer, mapping each input word to a unique continuous vector. This embedding layer is called a codebook. Subsequently, VQGAN[22] further improved this type of model and added adversarial loss during the training process. Wu et al.[23] proposed RIDCP, which uses the high-quality codebook prior trained by VQGAN in real image dehazing, and proposed a controllable high-quality prior matching operation to overcome the gap between the synthetic domain and the real domain, producing adjustable dehazing results.

III. PROPOSED METHOD

A. Network Architecture

The overall network architecture of the method we propose is shown in Fig. 1. Image data will be input into a dual-branch network, which has been successfully applied in the field of image dehazing by predecessors, demonstrating the good performance of this network. In our network structure, Branch One is called the Prior Matching Branch, the main structure of which includes a VQ encoder and decoder, discrete codebook, and the Pyramid Neighborhood Attention Encoder and Enhanced Decoder. Branch Two is called the Channel Attention Branch, which uses multiple residuals and channel convolutions to supplement Branch One. The outputs of the two branches are concatenated through a feature fusion tail composed of ReflectionPad, Conv and Tanh layer to output the final clean image.

Discrete codebook for high-quality prior: We are inspired by the latest research technology in image generation neighborhoods, VQGAN[22]. While extracting features, we use the pre-trained VQGAN to add image generation capabilities to the dehazing network, thereby helping to restore the image structure and details in heavy haze areas. The overall training is divided into two stages. The first stage requires training the VQGAN network on a high-quality clear dataset to achieve the restoration of image detail textures. In this stage, the network consists of a VQ encoder, codebook, and VQ decoder. The training goal is to obtain a codebook that stores high-quality clear image features and its corresponding VQ decoder. The VQ encoder and decoder adopt a network architecture based on UNet[24]. UNet first half is feature extraction, and the second half is upsampling. UNet has been proven to perform well in fields such as image classification and segmentation. Since the encoder's downsampling and feature refinement will lose some edge features, multiple residual structures are used. Through feature concatenation, the retrieval of edge features is achieved.

Given a high-quality image x as the input to the VQ encoder, an underlying feature map z is output. Then each pixel z_{ij} in Z is matched to the nearest element in the codebook, thereby obtaining the codebook discrete feature map z^q_{ij}. Subsequently, the discretized features are input into the decoder to obtain the processed image. The entire process can be represented as follows:

$$z_{ij} = E_{vq}(x_{ij}) \tag{2}$$

$$z^q_{ij} = M(\widehat{z}_{ij}) = arg \min_{z_k \in Z}(||\widehat{z}_{ij} - z_k||_2) \tag{3}$$

$$y_{ij} = D_{vq}(z^q_{ij}) \tag{4}$$

The features of a clear image are compressed into short vectors through discretization and stored in a codebook. The discrete codebook compresses the detailed texture features of the image, playing a crucial role in image reconstruction.

During the process of reconstructing an image using a high-quality codebook, the discrete encodings output by the encoder may have difficulty matching the high-quality codebook obtained through training on high-quality clear images due to the presence of haze. Therefore, we need to design a matching operation that utilizes a controllable distance recalculation method to reduce the problem of inconsistent data distribution caused by the domain gap between hazy and non-hazy images, thereby achieving a better reconstruction effect.

Specifically, this involves calculating the distances between the discrete encodings of the hazy image and each encoding in the codebook to find the true encoding with the smallest distance. Then, a weight function F is applied to adjust the final calculated distance, resulting in a matching formula that is expressed as follows:

$$M(z) = arg \min_{z_k \in Z} (F(f_k, \alpha) \times ||\widehat{z} - z_k||) \tag{5}$$

where f_k represents the frequency difference between the activation of the hazy image and the clear image on the codebook. The parameter α is used to adjust the degree of dehazing. z denotes the discretized features of the hazy image, while z_k represents the codebook encoding. The notation $||*||$ indicates the distance between the discretized features of the hazy image and the codebook encoding. The expression $arg \min(*)$ means to find the minimum value of the distance.

To facilitate the adjustment of matching results through the parameter α for achieving better defogging effects, we design a weight function as $F(f_k, \alpha) = f_k \times e^{\alpha}$. Subsequently, we modify the computation method of the activation frequency difference f_k in the codebook from statistical analysis to an iterative approach using network learning to obtain optimal values.

$$M(z) = arg \min_{z_k \in Z} (f_k \times e^{a} \times ||\widehat{z} - z_k||) \tag{6}$$

Regarding the solution for the parameter α, we assume the probability distribution of codebook activations for clear images as P_c, and the corresponding probability distribution for hazy images as P_h. The domain gap between the domains of clear and hazy images is transformed into the problem of finding the optimal parameter α in $F(f_k, \alpha)$ that minimizes the KL divergence between the probability distributions $P_c(x) = z_k$ and $P_h(x = z_k|\alpha)$.

The KL divergence measures the difference between two probability distributions and is commonly used in machine learning to assess the similarity of distributions. In our context, minimizing the KL divergence between P_c and P_h ensures that the codebook activations for clear and hazy images are as similar as possible, thereby reducing the domain gap and improving the accuracy of the matching process.

To solve for the optimal α, we can employ optimization techniques such as gradient descent or other numerical methods. The objective function to be minimized would be the KL divergence between $P_c(x) = z_k$ and $P_h(x = z_k|\alpha)$ weighted by the function $F(f_k, \alpha)$.

By iteratively updating α based on the gradients of the objective function, we can find the optimal value that minimizes the KL divergence, thereby improving the matching accuracy and ultimately enhancing the defogging effect.

It is worth noting that the computation of f_k and the optimization of α can be integrated into the overall training process of the network. By jointly optimizing the network parameters and the matching criteria, we can achieve better generalization and adaptability to various hazy scenes, leading to more robust and effective dehazing results.

Pyramid neighborhood attention encoder: The VQ encoder performs well when encoding clear images, but it proves insufficient when encoding images with dense haze or non-homogeneous haze. This is primarily because, in the task of dehazing, the encoder must not only extract the general structural texture features from the image but also distinguish the hazy areas within the image. Network architecture of the VQ encoder is relatively shallow, which does not allow it to adequately accomplish this task. In order to fully extract global features such as the texture and structure of hazy images, we designed an encoder based on Pyramid Neighborhood Attention in the prior matching branch. The Pyramid Neighborhood Attention is a variant of the self-attention mechanism found in the Vision Transformer[17], serving as an effective and scalable visual sliding window attention mechanism. Neighborhood Attention (NA)[25], [26] is a per-pixel operation that directs the self-attention (SA)[10] towards the nearest neighboring pixels, thus, compared to the quadratic complexity of self-attention, neighborhood attention has linear time and space complexity. Moreover, it surpasses the Vision Transformer and Swin Transformer[27] in downstream visual performance. In our designed Pyramid Neighborhood Attention encoder, four feature maps of different resolution sizes are obtained through the serialization process and two down-sampling operations. By utilizing a pyramid structure and cascading operations, the feature information from each previous layer serves as the input for the next layer, aggregating features from different levels and achieving feature reuse across different scales.

The NA operation can be expressed as:

$$NA_k = softmax(\frac{A_k}{\sqrt{d}})V_k \tag{7}$$

where A_k is the attention weight of the input with a neighborhood size of k, which is the dot product of the input query projection and its k nearest neighbor key projections. The neighboring value V_k is a matrix whose rows are projections of the k nearest neighboring values of the input, and $\sqrt{d}$ is the scaling parameter.

Enhanced decoder: The output results obtained solely through the VQ decoder tend to lack detailed information in areas with deep haze, and the image structure and texture are relatively blurred. To enhance the decoding capability of the hazy image's detailed features, we designed an enhanced decoder based on multiple attention mechanisms within the prior matching branch. By combining channel and pixel attention (CA+PA)[8], and finally passing through an enhancer block[28] based on pyramid pooling, we ensure that the detailed features across different scales are embedded into the final result.

Dual-branch network structure: In addressing the issue of poor feature extraction in dense hazy regions, we proposes to employ a channel attention branch, using multiple residual and channel convolutions to focus on the dense haze areas to achieve differentiated dehazing effects. Attention mechanisms enable the network to flexibly focus on the characteristics

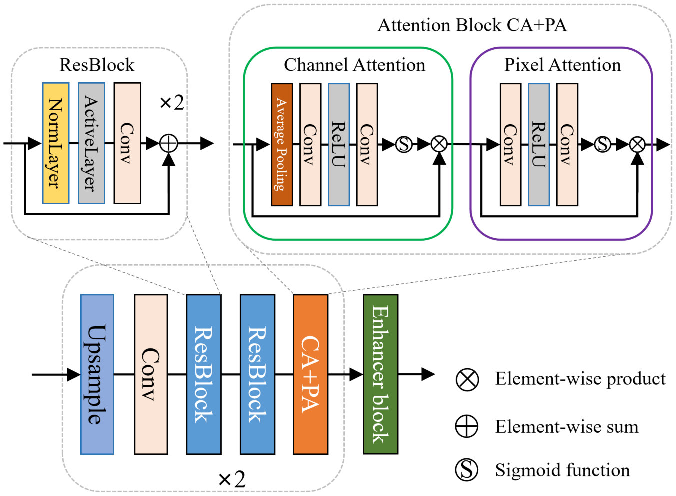

Fig. 3. Architecture of our Additional Enhanced Decoder(AED), which consists of upsampling layers, convolutions, residual blocks, channel attention and pixel attention blocks, as well as an enhancer block.

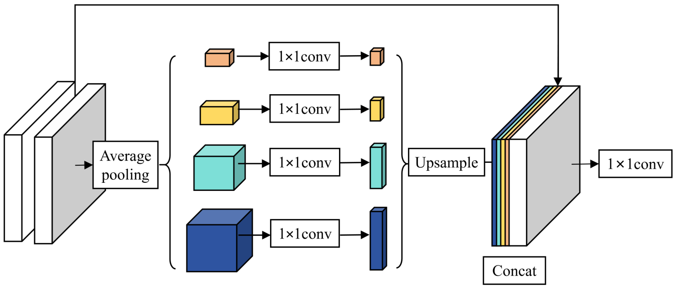

Fig. 4. Detail network structure of the Enhancer Block.

of the haze, reconstructing high-quality haze-free images. Non-homogeneous and dense haze significantly increases the brightness of its occluded areas. Paying more attention to the restoration of areas with significant brightness variations, such as the sky and snowfields, can avoid over-enhancement issues, thus improving the overall reconstruction performance of the image.

B. Loss Function

The network training proposed in this paper is divided into two stages in total. The training objective of the first stage is to make the reconstructed image as similar as possible to the original image. During training, the encoder maps the input intermediate features to a network embedding layer called the discrete codebook through discretization. The output of the encoder needs to satisfy the data distribution of the discrete codebook, and the decoder decodes the features that conform to the data distribution of the discrete codebook back into images. The training objective of the second stage is to dehaze. Firstly, the parameters of the following network layers obtained from the first stage are fixed:

1. High Quality Codebook.
2. VQ Decoder.

Then, the remaining parts of the network are trained. Therefore, different loss functions are adopted for different stages of network training in this paper.

Stage 1: High Quality Codebook Training. For the initial stage of discrete codebook training, the total training loss L_{vq} was divided into the image reconstruction loss L_{rec} and the codebook loss $L_{codebook}$, the loss is defined as

$$L_{vq} = L_{rec} + L_{codebook} \tag{8}$$

The image reconstruction loss can be further divided into the following formulas, where $\hat{x}$ is the output image and x is the input image. $|| * ||_1$ represents the L_1 loss, L_{per} is the perceptual loss, and L_{adv} is the adversarial loss

$$L_{rec} = ||\hat{x} - x||_1 + L_{per} + L_{adv} \tag{9}$$

Perceptual loss measures the perceptual similarity across the entire feature space. The function denotes the feature maps of the 3rd, 5th, and 8th convolutional layers in the pre-trained VGG16[29] network (i.e., j=3,5,8). The purpose of using this function is to capture perceptual and semantic information within the images. Perceptual loss is defined as[30]:

$$L_{perc} = \frac{1}{N}\sum_{j}\frac{1}{C_jH_jW_j}||\phi_(f_\theta(x)) - \phi_j(y)||_2^2 \tag{10}$$

Since pixel-based loss functions cannot provide sufficient supervision on small datasets, an adversarial loss is added to mitigate the shortcomings of the above losses[31].

$$L_{adv} = \sum_{n=1}^{N} -\log_D(f_\theta(x)) \tag{11}$$

where D represents the discriminator used during the training of the codebook. N indicates the number of sample data.

The codebook loss can be further divided into the following formula[32]

$$\begin{aligned} L_{codebook} = ||sg(\widehat{z}) - z^q||_2^2 + \beta||sg(z^q) - \widehat{z}||_2^2 \\ +\gamma||CONV(z^q) - \phi(x)||_2^2 \end{aligned} \tag{12}$$

where $sg[*]$ denotes stop-gradient, with $\beta = 0.25$, $\gamma = 0.1$. The last term is a semantically guided regularization item, where $CONV$ signifies a simple convolutional layer and ϕ is a pre-trained VGG19[29]. This loss function primarily measures the quantization error between the output z of the encoder and the discrete vector z_q.

Stage 2: Dehaze Training.

For the second stage of the dehazing task training. Assuming the hazy image input is denoted as x_h, and the fog-free ground truth image input as x_{gt}, with the dehazing network encoder represented by E, the training codebook's encoder by E_{vq}, and the training codebook's decoder by G_{vq}, and the enhancement decoder by G. We can obtain the intermediate features of the hazy image after processing by the encoder E, and the intermediate features of the haze-free image after processing by the encoder E_{vq}. To control the style difference between the generated images and the fog-free images during the image

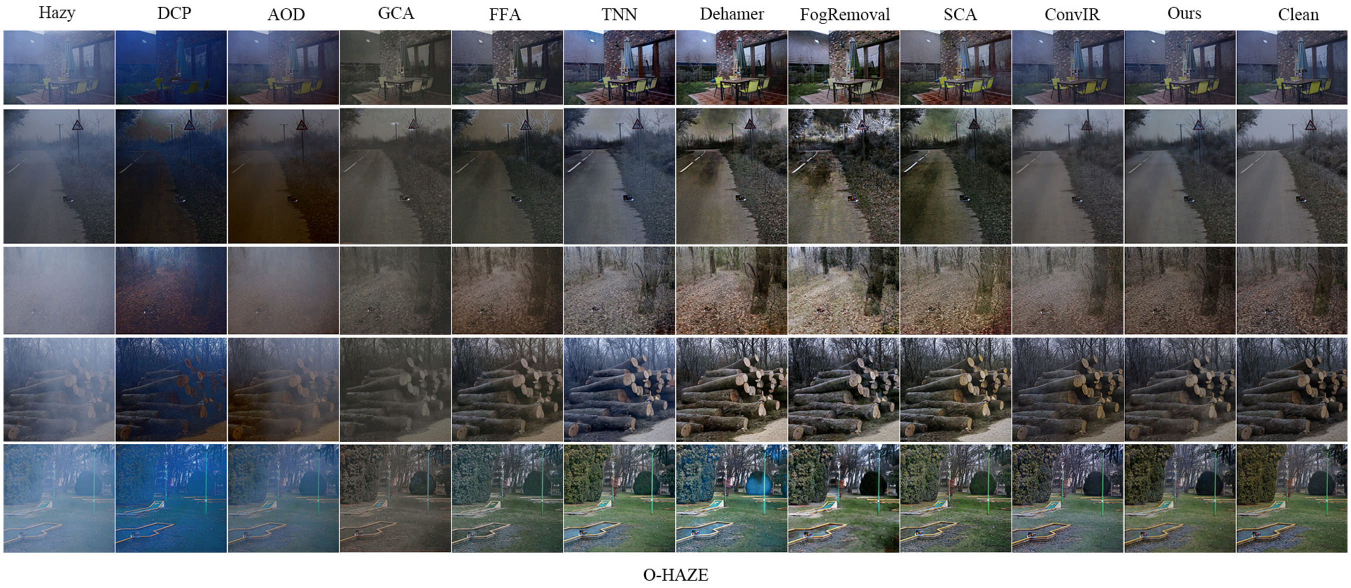

Fig. 5. Visual examples in dataset O-HAZE. From left to right are hazy images, advanced methods for comparison and our method result, and Ground Truth.

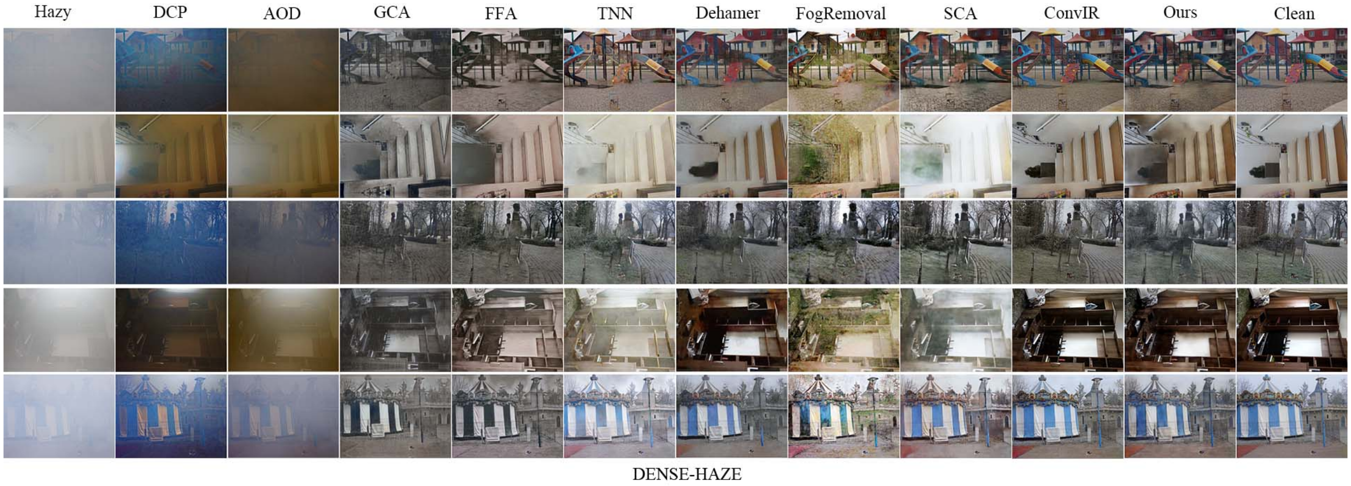

Fig. 6. Visual examples in dataset DENSE-HAZE. From left to right are hazy images, advanced methods for comparison and our method result, and Ground Truth.

generation process, we use $\Psi(\cdot)$, which is the Gram matrix, to measure the style loss[33]. A discriminator D is used to determine whether the generated features are realistic.

Therefore, the loss function for the dehazing stage can be written in the following form:

$$L_E = ||\widehat{z}_h - z_t^q||_2^2 + \lambda_{style}||\Psi(\widehat{z}_h - \Psi(z_{gt}^q)||_2^2 \\ +\lambda_{adv}\sum_i -E[D(\widehat{z}_h)] \tag{13}$$

The loss function for the remaining part of the network is as follows, where y represents the final output and ϕ denotes the pre-trained VGG16[29]. The gradient of this decoder will not be backpropagated into the encoder.

$$L_{remain} = ||y - x_{gt}||_1 + \lambda_{per}||\phi(y) - \phi(x_{gt})||_2^2 \tag{14}$$

IV. EXPERIMENTS

A. Datasets

To validate the effectiveness of our proposed model, we conducted experiments on four datasets, namely O-HAZE[36], NH-HAZE-20[37], NH-HAZE-21[38], and NH-HAZE-23[39], for the purpose of training and evaluating our model. Specifically, we selected five pairs of images from each dataset for validation and testing, while the remaining images were used for training. More detailed information can be obtained from the accompanying table II.

The O-HAZE dataset is authentic collection featuring uniform haze conditions, encompassing 45 pairs of outdoor images. These hazy images were captured under genuine atmospheric haze produced by specialized equipment, ensuring their realism and utility for benchmarking dehazing algorithms.

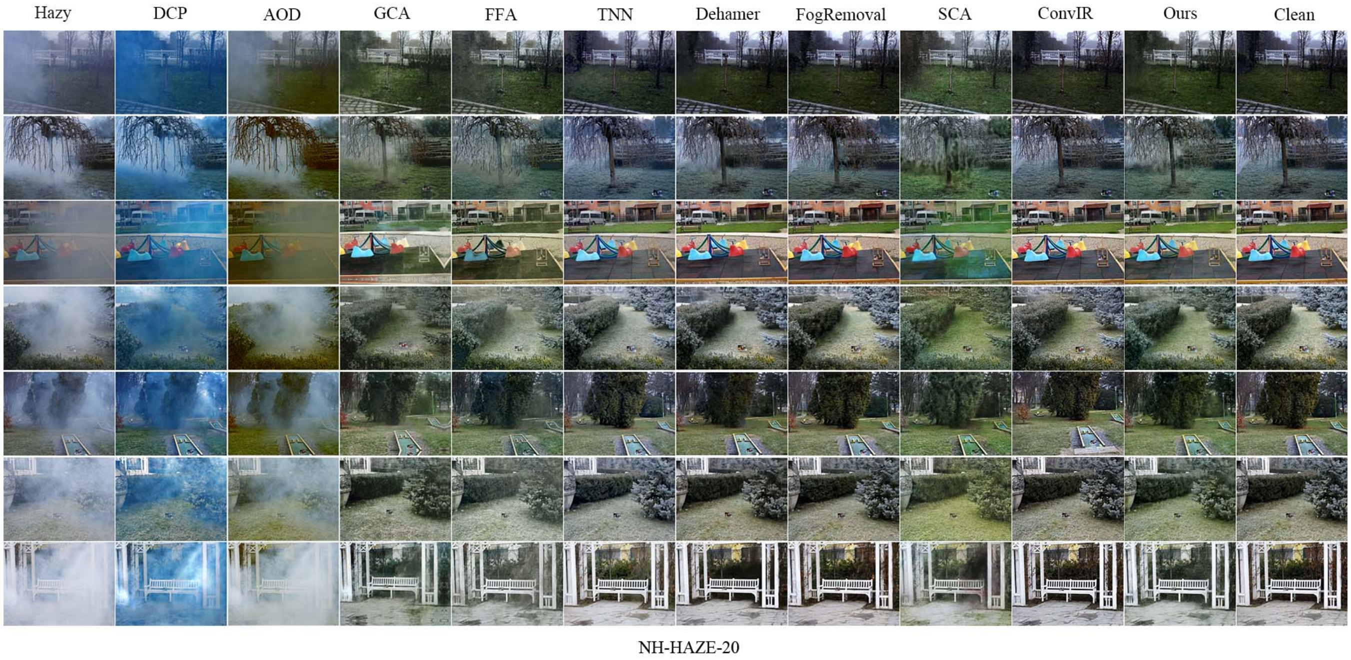

Fig. 7. Visual examples in dataset NH-HAZE-20. From left to right are hazy images, advanced methods for comparison and our method result, and Ground Truth.

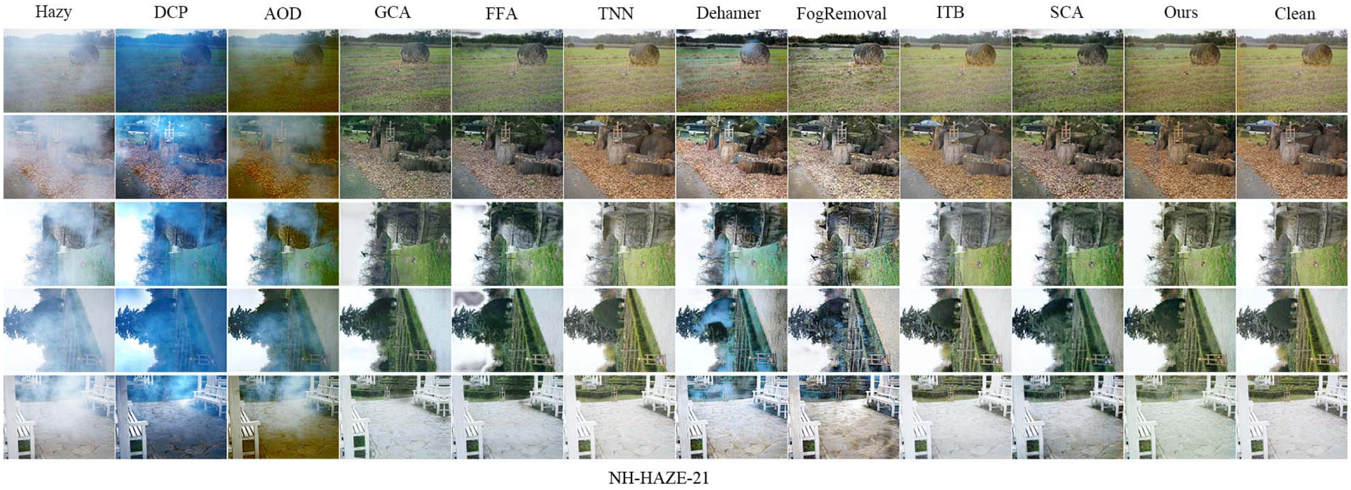

Fig. 8. Visual examples in dataset NH-HAZE-21. From left to right are hazy images, advanced methods for comparison and our method result, and Ground Truth.

The NH-HAZE dataset addresses a critical gap in the field of image dehazing: the scarcity of real-world images with non-uniform haze for reference. Unlike many synthetic datasets, NH-HAZE provides a collection of truly non-homogeneous hazy images paired with their corresponding clear counterparts. The non-homogeneous haze within the NH-HAZE dataset is introduced through the simulation of realistic hazy conditions using professional haze generators, making it a more challenging and practical dataset for evaluating dehazing methods. Its temporal evolution, marked by release dates, allows for further categorization into NH-HAZE-20, NH-HAZE-21, and NH-HAZE-23, containing 40, 25, and 55 image pairs respectively.

The DENSE-HAZE dataset introduces a novel contribution to the field by focusing on dense, uniform haze scenarios. Comprising 55 pairs of heavily hazy and their corresponding clear images across various outdoor settings, DENSE-HAZE stands out for its portrayal of dense haze conditions recorded through specialized haze-producing machinery. The resultant images are so heavily veiled that the original objects within them are nearly indiscernible, presenting a significant challenge in dehazing tasks compared to conventional datasets. This characteristic makes DENSE-HAZE particularly demanding for testing the robustness and effectiveness of dehazing algorithms under extreme conditions.

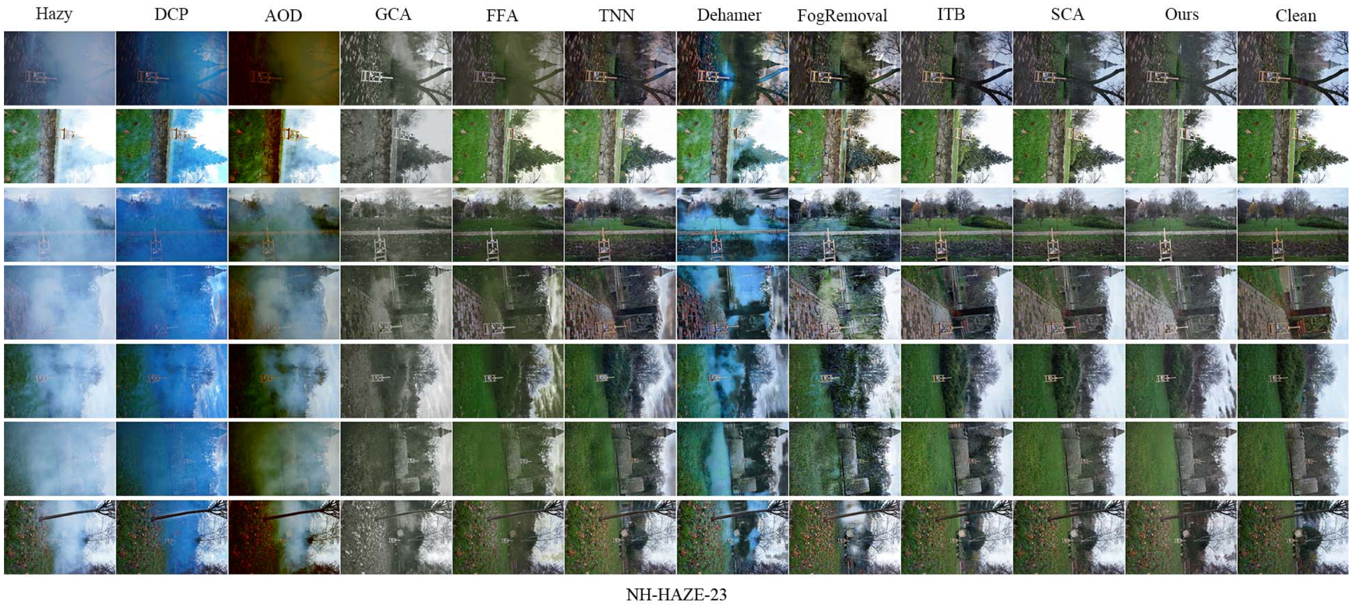

Fig. 9. Visual examples in dataset NH-HAZE-23. From left to right are hazy images, advanced methods for comparison and our method result, and Ground Truth.

TABLE I
QUANTITATIVE COMPARISON RESULTS OF VARIOUS DEHAZING METHODS AND PROPOSED METHOD ON I-HAZE, O-HAZE, DENSE-HAZE, NH-HAZE-20, NH-HAZE-21 AND NH-HAZE-23. **BOLD** INDICATES THE BEST AND UNDERLINE INDICATES THE SECOND BEST.

Dataset	Metric	DCP[3] (TPAMI10)	AOD[6] (ICCV17)	GCA[34] (WACV19)	FFA[8] (AAAI20)	TNN[13] (CVPRW21)	DeHamer[11] (CVPR22)	FogRomoval[16] (ACCV22)	SCA[19] (CVPRW23)	ConvIR[35] (TPAMI24)	ours
O-HAZE	PSNR	12.92	17.69	19.50	22.12	**25.54**	17.02	24.61	19.47	25.25	24.54
	SSIM	0.50	0.62	0.66	0.77	0.78	0.43	0.75	0.74	0.78	**0.79**
	LIPIS	0.35	0.34	0.31	0.21	**0.20**	0.29	0.28	0.39	0.33	0.24
DENSE-HAZE	PSNR	10.85	13.30	12.42	16.26	16.36	16.62	16.67	16.34	17.45	**17.79**
	SSIM	0.40	0.47	0.48	0.54	0.58	0.56	0.50	0.56	**0.64**	0.62
	LIPIS	0.79	0.77	0.62	0.51	0.50	0.49	0.46	0.58	0.42	**0.37**
NH-HAZE-20	PSNR	12.29	13.44	17.58	18.51	17.18	18.53	20.99	19.52	20.65	**21.73**
	SSIM	0.41	0.41	0.59	0.64	0.61	0.62	0.61	0.65	**0.80**	0.79
	LIPIS	0.56	0.54	0.31	0.26	0.28	0.26	0.23	0.55	0.20	**0.19**
NH-HAZE-21	PSNR	11.30	13.22	18.76	20.40	20.13	18.17	18.34	21.14	-	**21.52**
	SSIM	0.60	0.61	0.77	0.81	0.80	0.77	0.72	0.77	-	**0.89**
	LIPIS	0.50	0.49	0.24	0.15	**0.14**	0.25	0.44	0.51	-	0.20
NH-HAZE-23	PSNR	11.87	12.47	16.36	18.09	18.19	17.61	18.80	20.44	-	**20.85**
	SSIM	0.47	0.37	0.51	0.58	0.64	0.61	0.60	0.66	-	**0.79**
	LIPIS	0.62	0.56	0.49	0.33	0.27	0.35	0.55	0.36	-	**0.22**

TABLE II
THE DETAILS OF THE DATASETS USED IN OUR EXPERIMENTS

Dataset	Train set	Validation set	Image size
O-HAZE	40	5	2426×2942-5416×3592
DENSE-HAZE	50	5	1600×1200
NH-HAZE-20	50	5	1600×1200
NH-HAZE-21	20	5	1600×1200
NH-HAZE-23	35	5	4000×6000

B. Implementation Details

During training, input images were randomly cropped to a size of 256×256 and augmented through scaling, random rotation, and flipping. The Adam optimizer[40] was employed in this study, with default β_1 and β_2 values set to 0.9 and 0.99, respectively. The initial learning rate was set to 0.0001, and the batch size was set to 1. The model was implemented on an NVIDIA V100 Tensor Core using the Pytorch framework.

C. Evaluation Metrics and Competitors

Quantitative analysis was conducted using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity index (SSIM)[41], LPIPS[42] for evaluation. PSNR, measured in decibels (db), represents the ratio of the maximum power of a signal to the noise power that may affect its representation accuracy. A higher PSNR value indicates less distortion in dehazed image. SSIM evaluates image similarity from three aspects: brightness, contrast, and structure, with a value range of $[0, 1]$. A higher SSIM value indicates greater similarity between the current image and the Ground Truth. Compared to PSNR, SSIM better aligns with human visual perception in assessing image quality. LPIPS measures the difference between two images by extracting features for each image using deep learning techniques and computing their similarity based on these features.

D. Quantitative Evaluations on Benchmarks

To verify the effectiveness of the network and its generalization ability in difference scenarios, we conduct experiments for comparison. We compare our method with 9 state-of the art dehazing methods, including DCP[3], AOD[6], GCA[34], FFA[8], TNN[13], DeHamer[11], FogRemoval[16], SCA[19], ConvIR[35]. The results for the four different quantitative metrics are shown in Table I.

Fig. 5-9 present the visual comparison between our method and other state-of-the-art algorithms across six dehazing datasets. Visualization of NH-HAZE-21 and NH-HAZE-23 using ITB[18] instead of ConvIR. Using NH-HAZE-20 as an example, which is characterized by non-homogeneous haze distribution with certain regions heavily obscured, making background objects barely discernible, our approach effectively mitigates this haze. It produces a color representation and clarity closely aligned with the ground truth, without noticeable chromatic aberrations, achieving a largely dehazed effect. However, in areas with extremely dense haze, the model faces challenges in accurately estimating haze-free conditions, resulting in less precise restoration of textural details in these regions. Despite this, the perceptual effect is significantly improved, with the haze substantially cleared.

In contrast, the DCP algorithm fails to significantly reduce the haze, leaving a pronounced blue cast and severe color distortion throughout the image. While the AOD algorithm avoids obvious color distortions, its dehazing performance is limited in the context of real-world, non-uniform haze scenarios. The GCA method shows some dehazing effect but suffers from severe detail loss in the post-dehazed images, particularly in areas with high haze concentration, where residual white haze remains. Although FFA, TNN, and Dehamer address uneven haze to some extent, they still struggle with detail preservation and color accuracy.

FogRemoval produces superior results in comparison, but ITB processed images exhibit slight color distortions. SCA generally performs well in dehazing but tends to over-smooth regions with fine details, such as densely packed branches, leading to perceptual distortions. In contrast, our method consistently outperforms these alternatives, visually aligning more closely with haze-free images. Most of the haze is effectively removed, and critical textural details are preserved.

E. Ablation Study

To assess the contribution of each component in our proposed method, we conducted a series of ablation studies. The ablation experiments were conducted on the NH-HAZE-20 dataset, with results summarized in Table III. And experiments involved the following configurations:

1) Base: Consists of VQ Encoder, VQ Decoder and RSTB[44].

2) Base+CB: Includes the codebook(CB).

3) Base+PDiNAT+CB: Replaces RSTB with the proposed PDiNAT.

4) Base+PDiNAT+CB+CAB: Incorporates the channel attention branch(CAB).

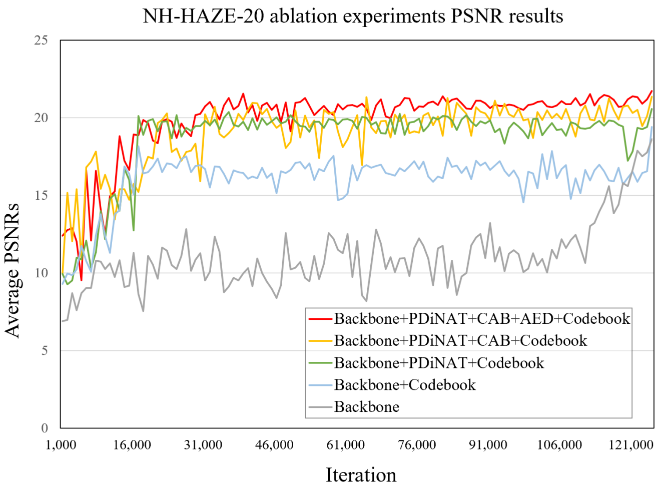

Fig. 10. The ablation experiments on the NH-HAZE-20 dataset show the PSNR variation curves for different experimental configurations as the number of iterations increases.

5) Base+PDiNAT+CB+CAB+AED: Adds the additional enhancement decoder(AED).

Effectiveness of CodeBook. The high-quality codebook introduced in this paper effectively compresses the detailed texture features of the image by reconstructing the clear image, encapsulating high-quality prior knowledge. When the codebook is integrated into the Base model, the overall performance improves, with a PSNR increase of 0.8 dB, an SSIM enhancement of 0.01, and an LPIPS reduction of 0.06.

Effectiveness of PDiNAT. The proposed PDiNAT module enables the model to fully extract local texture and structural features from hazy images. By utilizing a pyramid structure, it aggregates features from different levels, facilitating multi-scale feature reuse. This approach effectively preserves the texture features of hazy images, thereby assisting the subsequent codebook feature matching operation. Experimental results demonstrate that the inclusion of PDiNAT leads to a further performance improvement: PSNR increases by 1.15 dB, SSIM improves by 0.07, and LPIPS decreases by 0.09, compared to the model with the Base configuration.

Effectiveness of CAB. The CAB module employs a classical channel attention branch commonly used in the dehazing domain, which emphasizes areas with significant brightness variation, such as non-homogeneous fog and dense haze regions. This focus helps mitigate over-enhancement issues and improves the overall image reconstruction performance.

Effectiveness of AED. The AED module introduces an enhanced decoder based on multiple attention mechanisms, which prioritizes the detailed features of hazy images. This enhancement strengthens the network's decoding capability, contributing to more accurate and refined reconstruction of hazy scenes.

As observed, the model's performance improves progressively as more components are integrated. Fig. 10 illustrates the PSNR performance of networks with different modules at various iterations. As shown, the application of the modules proposed in this paper leads to a significant improvement in

TABLE III
ABLATION STUDY ON THE PROPOSED NETWORK PERFORMANCE AFTER GRADUALLY ADD MODULES. THE SCORES ARE EVALUATED ON NH-HAZE-20 DATASET. **BOLD** INDICATES THE BEST AND UNDERLINE INDICATES THE SECOND BEST.

Index	Method	PSNR	SSIM	LPIPS
1	Base	18.61	0.61	0.37
2	1 + CB	19.41	0.62	0.31
3	2 + PDiNAT	20.56	0.69	0.22
4	3 + CAB	21.36	0.78	0.20
5	4 + AED	**21.73**	**0.79**	**0.19**

PSNR, indirectly validating the effectiveness of the proposed components.

V. CONCLUSION

In this paper, we propose a two-branch neural network for non-homogeneous dehazing via high quality codebook and prove its strong power in various dehazing tasks. Subsequently, we proposed a feature pyramid encoder based on dilated neighborhood attention, which can effectively extract features of haze distribution in images through multi-scale feature maps. In order to enhance the decoding capability of the network, we have designed an enhanced decoder that utilizes multiple attention and enhancement blocks to restore the structure and detailed features of the image. Finally, extensive experiments have shown that our method has significant advantages in non-homogeneous datasets.

REFERENCES

[1] E. McCartney, "Optics of the atmosphere: scattering by molecules and particles," 1976.

[2] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *International Journal of Computer Vision*, vol. 48, pp. 233–254, 2002.

[3] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[4] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Josa*, vol. 61, no. 1, pp. 1–11, 1971.

[5] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[6] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4770–4778.

[7] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7314–7323.

[8] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 908–11 915.

[9] X. Yin, G. Tu, and Q. Chen, "Multiscale depth fusion with contextual hybrid enhancement network for image dehazing," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5812–5820.

[12] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, 2023.

[13] 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*, 2021, pp. 193–202.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[15] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301.

[16] Y. Jin, W. Yan, W. Yang, and R. T. Tan, "Structure representation network and uncertainty feedback learning for dense non-uniform fog removal," in *Asian Conference on Computer Vision*. Springer, 2022, pp. 155–172.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[18] Y. Liu, H. Liu, L. Li, Z. Wu, and J. Chen, "A data-centric solution to nonhomogeneous dehazing via vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1406–1415.

[19] Y. Guo, Y. Gao, W. Liu, Y. Lu, J. Qu, S. He, and W. Ren, "Scanet: Self-paced semi-curricular attention network for non-homogeneous image dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1884–1893.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[22] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 873–12 883.

[23] R.-Q. Wu, Z.-P. Duan, C.-L. Guo, Z. Chai, and C. Li, "Ridcp: Revitalizing real image dehazing via high-quality codebook priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 282–22 291.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[25] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6185–6194.

[26] A. Hassani and H. Shi, "Dilated neighborhood attention transformer," *arXiv preprint arXiv:2209.15001*, 2022.

[27] 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 International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[28] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8160–8168.

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

[30] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[32] C. Chen, X. Shi, Y. Qin, X. Li, X. Han, T. Yang, and S. Guo, "Real-world blind super-resolution via feature matching with implicit high-resolution priors," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1329–1338.

[33] M. W. Gondal, B. Schölkopf, and M. Hirsch, "The unreasonable effectiveness of texture transfer for single image super-resolution," in *Proceedings of the European Conference on Computer Vision*. Springer, 2019, pp. 80–97.

[34] 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 *2019 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2019, pp. 1375–1383.

[35] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Revitalizing convolutional network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: A dehazing benchmark with real hazy and haze-free outdoor images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 754–762.

[37] C. O. Ancuti, C. Ancuti, F.-A. Vasluianu, and R. Timofte, "Ntire 2020 challenge on nonhomogeneous dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 490–491.

[38] ——, "Ntire 2021 nonhomogeneous dehazing challenge report," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 627–646.

[39] C. O. Ancuti, C. Ancuti, F.-A. Vasluianu, and Timofte, "Ntire 2023 hr nonhomogeneous dehazing challenge report," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2023, pp. 1808–1825.

[40] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[43] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[44] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 1833–1844.