A Data-Centric Solution to NonHomogeneous Dehazing via Vision Transformer

Yangyi Liu[1], Huan Liu[1], Liangyan Li[1], Zijun Wu[2] and Jun Chen[1]

[1]McMaster University, Hamilton, Canada
[2]China Telecom Research Institute, Shanghai, China
{*liu5, lil61, chenjun*}@*mcmaster.ca*, *liuh127@outlook.com*, *wuzj12@chinatelecom.cn*

Figure 1. Our results on NTIRE 2023 dehazing challenge, achieving the best performance in terms of PNSR, SSIM and LPIPS.

Abstract

Recent years have witnessed an increased interest in image dehazing. Many deep learning methods have been proposed to tackle this challenge, and have made significant accomplishments dealing with homogeneous haze. However, these solutions cannot maintain comparable performance when they are applied to images with non-homogeneous haze, e.g., NH-HAZE23 dataset introduced by NTIRE challenge. One of the reasons for such failures is that non-homogeneous haze does not obey one of the assumptions that is required for modeling homogeneous haze. In addition, a large number of pairs of non-homogeneous hazy image and the clean counterpart is required using traditional end-to-end training approaches, while NH-HAZE23 dataset is of limited quantities. Although it is possible to augment the NH-HAZE23 dataset by leveraging other non-homogeneous dehazing datasets, we observe that it is necessary to design a proper data-preprocessing technique that reduces the distribution gaps between the target dataset and the augmented one. This finding indeed aligns with the essence of data-centric AI. With a novel network architecture and a principled data-preprocessing approach that systematically enhances data quality, we present an innovative dehazing method. Specifically, we apply RGB-channel-wise transformations on the augmented datasets, and incorporate the state-of-the-art transformers as the backbone in the two-branch framework. We conduct extensive experiments and ablation studies to demonstrate the effectiveness of our proposed method. The source code is available at https://github.com/yangyiliu21/ntire2023_ITBdehaze.

1. Introduction

Recent years have witnessed an increased interest in image dehazing, which is categorized as one of the sub-tasks in image restoration. Haze naturally exists all over the world, and has become more frequent due to the climate change. This common atmospheric phenomenon has drawn significant attention because of its potential risks to traffic safety, as both the human observation and computer vision models are prone to fail in hazy scenes. These make image dehazing an important low-level vision task, and many methods have been proposed to tackle this challenge [10, 15, 18, 21–23, 29, 30, 35–37, 39, 42].

Among them, many neural network based approaches [10, 11, 21, 29, 39, 42] show remarkable performance in

handling image dehazing problem. Specifically, benefiting from powerful network modules and vast training data, the end-to-end approaches deliver promising results. However, as the distribution of haze becomes more complicated and non-homogeneous, many of them fail to achieve satisfying results. The reason for such failures is because the thickness of the non-homogeneous haze is not determined entirely by the depth of the background scene.

Although researchers have made tremendous efforts collecting data with non-homogeneous haze, e.g., the NH-HAZE datasets [6–8], the quantity is still limited. A common belief is that models are prone to encounter the overfitting problems when training a deep neural network from scratch with such small datasets. A naive solution is to combine all the available non-homogeneous haze datasets together to form a relatively larger dataset. However, due to the differences between datasets caused by a variety of factors, such as color distortion, objects complexity and camera capability, it has been shown that a direct combination actually compromises the dehazing performance on individual datasets [22]. It remains a serious challenge to find a robust solution to cope with the practical situation where both the quality and quantity of the available data limited.

To address the above-mentioned problems, we adopt the two-branch framework consisting of state-of-the-art backbone networks, with a novel data-preprocessing transformation applied on the NH-HAZE datasets from previous years. Motivated by the idea of data-centric AI that machine learning has matured to a point that high-performance model architectures are widely available, while approaches to engineering datasets have lagged [1, 27], we put much effort in engineering the data. Inspired by the promising performance of gamma correction [15, 37], we propose a simple yet effective RGB-channel-wise data-preprocessing approach. We demonstrate its suitability for this competition setting, and argue that it is prospective to be the principle for augmenting similar dataset. Details of this data-centric AI inspired preprocessing approach are discussed in later sections. Regarding to the network architecture, we design our model under the two-branch framework [15,36,37]. In the first branch, we adopt the Swin Transformer V2 model [24] pre-trained on ImageNet dataset [12] as the encoder. The powerful Swin Transformer is accredited to be able to supersede the previous methods in many contexts of transfer learning, where the knowledge gained from large-scale benchmark is adapted to task-specific datasets [20,24]. Such pertinent features are of vital importance when dealing with small real-world non-homogeneous datasets [37]. Alongside a refined decoder and skip connections, the first branch extracts multi-level features of the hazy images. The second branch is introduced to complement the knowledge learned from the pre-trained model by exclusively working on the domain of target data. For simplicity, we follows [37]

to build the second branch with a RCAN [40]. Since there is no down-sampling and up-sampling operations in the second branch, we expect it to extract features distinct from the ones obtained by the first branch. Finally, a fusion tail aggregates the results from both branches and produces dehazed output images.

Overall, our contributions are summarized as follows. Firstly we put forward a simple but effective data-preprocessing approach inspired by data-centric AI, leveraging extra data to significantly enhance our model. Secondly, we incorporate the state-of-the-art backbone in the two-branch framework. By carefully balancing the two branches, our model demonstrates promising results using limit-sized datasets, and outperforms other current approaches adopting this pipeline. Finally, we conduct extensive experiments to demonstrate the competitive performance of our proposed method. With substantial ablation study on different combinations of models and data, we hope to convince the future competition participants to pay equal attention to model design and data engineering.

2. Related Works

In this section, we briefly review the literature of single image dehazing and learning with limited data.

Single Image Dehazing. Approaches proposed for single image dehazing are divided into two categories: prior-based methods and learning-based methods. To guarantee the performance, prior-based methods require reasonable assumptions and knowledge on hazy images to obtain accurate estimations about the transmission map and atmospheric light intensity in ASM modeling [26]. Representative works in this category include [9,14,18,34,44]. Specifically, [34] observed that clear images have higher contrast comparing to the hazy counterparts, and proposed a local contrast maximization method. Based on the assumption that image pixels in no-haze patches have intensity values close to zero in at least one color channel, [18] introduced Dark Channel Prior (DCP). [44] presented a linear model adapting color attenuation prior (CAP) to estimate the depth according to the knowledge about the difference between the brightness and the saturation of hazy images. Prior-based methods left a permanent mark in single image dehazing but their vulnerability when adapted in variable scenes pivoted the researchers to another direction, the learning-based methods. With the advances in neural networks, [10, 11, 21, 29, 39, 42] have proposed progressively more powerful models that are capable of directly recovering the clean image from hazy image without estimating the transmission map and depth. The superiority of these methods in removing homogeneous haze is attributed to the availability of large training datasets. When applied on non-homogeneous haze, they fail to yield comparable results. The limited quantity of existing non-homogeneous haze
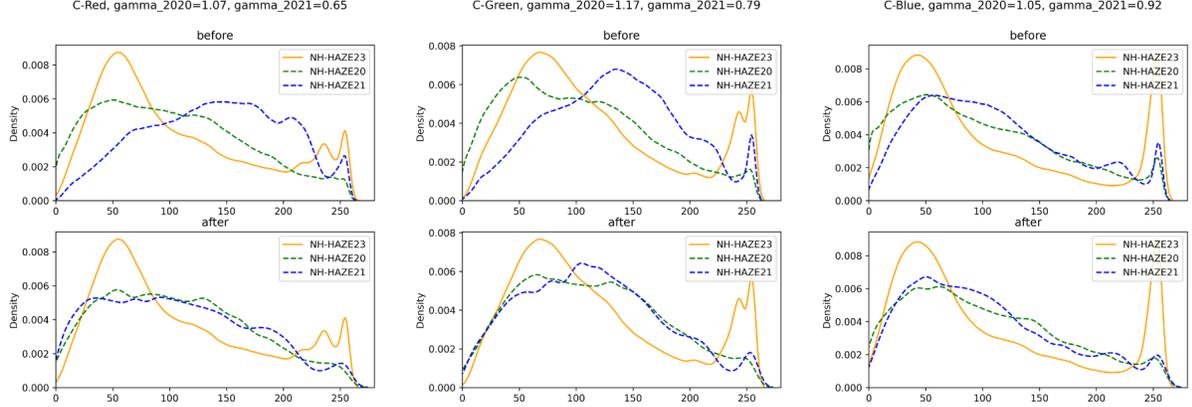
Figure 2. Comparison of RGB-wise distribution of datasets (GT) before and after being processed by our proposed method.

datasets prevents researchers from adopting simple end-to-end training methods.

Learning with Limited Data. Data is indispensable for all the AI models. Many of the models demand a huge dataset for training, but large dataset is not always available. Therefore, it urges the researchers to find solutions to accomplish training with limited data. In terms of dehazing, a seemingly straightforward solution to address the issues caused by small non-homogeneous training datasets is composing a relatively large dataset by combining several small datasets all together. In terms of NTIRE2023 challenge [8], it can be done by augmenting the NH-HAZE datasets (augmented dataset) [6, 7] with this year's data (target dataset). Surprisingly, against the common believe that larger dataset is always better in deep learning, [22] observed that the models perform better when training and testing are conducted on a single dataset (as opposed to the union of all datasets). This observation indicates that the augmented dataset locates in a different domain comparing to the target data. Direct aggregation introduces domain shift problem within the dataset. Thereby, [22] proposed a testing time training strategy to mitigate the problems, while [15, 31, 37] chose to adjust the domains of training data before sending them into the dehazing modules. Interestingly, the idea of focusing on improving the dataset rather than the model was introduced by the Data-Centric AI competition [1]. Data-Centric AI is anticipated to deliver a set of approaches for dataset optimization, thereby enabling deep neural networks to be effectively trained using smaller datasets [27]. The set of proposed techniques ranges widely from simple ones to complex combinations [38]. Through our experiments and qualitative analysis, we find that a too simple approach, such as the gamma correction adopted by [15, 37] fails to recover the color accurately. Nevertheless, a complicated method, like [31] applying domain adaptation to learn a separate neural network to translate the data, is infeasible due to the scarcity and lacking of depth information of the available data. In the next section, we introduce our innovative solution standing out in the NTIRE challenge settings.

3. Proposed Method

In this section, we introduce the details of our methodology following the order of the working pipeline. Firstly, we demonstrate the data-preprocessing method inspired by the idea of data-centric AI. Secondly, details of our model architecture are presented, as well as the functions of each component. Finally, we introduce the loss functions applied to train our proposed networks.

3.1. Data-Centric Engineering

Systematically engineering the data is a key requirement for training deep neural networks. The idea of data-centric AI moreover emphasizes on assessing the data quality before deployment [38]. By comparing the NH-HAZE20 and 21 dataset [6, 7] to the data provided this year both numerically and empirically, we notice obvious color discrepancy. When evaluating on this year's test data, training on a direct combination of all data does not boost the score comparing to training on this year's data only (see results in Section 4.3.1). Therefore, our goal is to propose an approach that reduces the color differences, and shifts the distribution of augmented data towards that of target data. Inspired by the success application of gamma correction [15, 37] as a simple yet effective data-preprocessing technique, we propose a more systematic solution for data engineering. Instead of the practice in [15, 37] by applying gray-scale gamma correction, we here introduce to correct on each R,G,B channel separately:

$$O_{R,G,B} = \left(\frac{I_{R,G,B}}{255}\right)^{\frac{1}{\gamma_{R,G,B}}} \tag{1}$$

where O and I are output and input pixel intensity ($\in [0, 255]$), respectively. γ is the gamma factor. The sub-
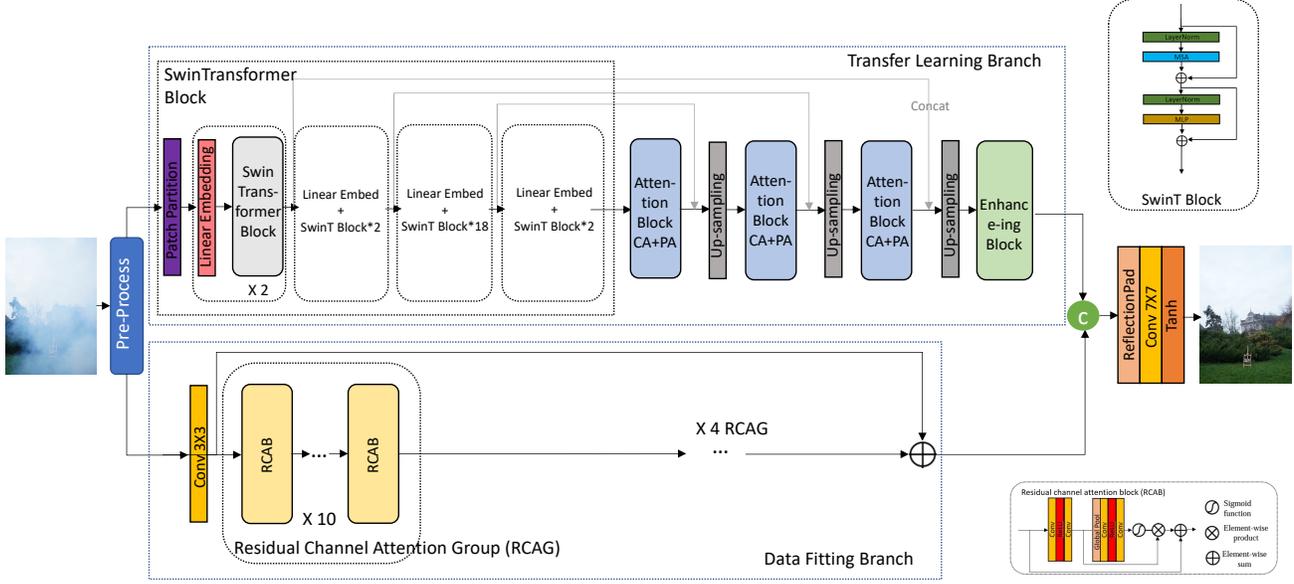
Figure 3. An overview of our network. The model consists of two branches. The transfer learning branch is composed by Swin Transformer based model. The data fitting branch consists of residual channel attention groups.

scripts R, G, B indicate that the values for different channels are unique.

As for implementation, we first calculate the average pixel intensity of each channel of the three datasets; then for each channel of the NH-HAZE20 or 21 dataset, we apply a transformation with a unique gamma value to all the pixels, resulting in similar mean and variance values comparing with the corresponding channel of NH-HAZE23 dataset. In Figure 2, we present the histogram change with corresponding γ values. From observation, our method adjusts the color of NH-HAZE20 and 21 data to become much similar to the NH-HAZE23 data. Numerically, the average pixel intensity of 2023 data is 107.46(R), 114.48(G), 101.92(B). After applying our method, the adjusted average pixel intensity of NH-HAZE20 data is 107.77(R), 114.33(G), 102.08(B); and the adjusted ones of NH-HAZE21 data is 107.43(R), 115.01(G), 102.13(B). Note that, we not only apply such preprocessing method on the clean ground truth images but also on the hazy images (as opposed to [15, 37] only manipulating the ground truth images).

With this novel data-preprocessing method, the distributions of all three color channels of NH-HAZE20 and 21 data are shifted closer to those of NH-HAZE23 dataset. Benefiting from more in-distribution data, the models gain substantial improvements. Being able to work with small but good dataset, rather than a larger but internally diverged one helps us stand out in the competition. This indeed aligns with the idea of data-centric AI [27, 38]. For future competition participants, we elect this approach to be a good starting point for data engineering.

3.2. Network Architecture

As shown in Figure 3, the pre-processed data is fed into a two-branch model architecture. This two-branch framework has been successfully employed in various computer vision tasks [19], and has facilitated several works [15, 36, 37] winning the awards in the past NTIRE challenges. In our implementation, the first Transfer Learning Branch aims to extract pertinent features of the inputs with pre-trained weights initialization. The second Data Fitting Branch is responsible to complement the knowledge learned from the first branch and work on the domain of target data. The fusion tail aggregates the outputs from both branches and produces dehazed images.

Swin Transformer based Transfer Learning. To leverage the power of transfer learning [33], we use the ImageNet [12] pre-trained Swin Transformer [24] as the backbone of our encoder. Swin Transformer achieves the state-of-the-art performance in many vision tasks. It is exceptionally efficient and more accurate as comparing to its predecessor, Vision Transformer (ViT) [13], which struggles with high resolution images because its complexity is quadratic to the input size. The working pipeline of the Swin Transformer is summarized as follows. First, Swin Transformer splits an input image into non-overlapping patches with a patch splitting module. Through a linear embedding layer, the patches and their features are set as a concatenation of the raw pixel RGB values, also referred to as "token", and then be projected to an arbitrary dimension. These tokens are processed by a cascade of stages. Each stage consists of a linear embedding layer and several Swin Transformer

Block (SwinT Block) modules. SwinT Block uses cyclic-shift with MSA modules to implement efficient batch computation for shifted window partitioning. From the previous stage to the next, the spatial dimension of the feature maps are effectively reduced, resulting in hierarchical feature maps. These modules compose our encoder part of the Transfer Learning Branch. As for the decoder part, we adopt the ideas from [15,37]. With skip connections, the attention blocks and up-sampling layers gradually restore the hierarchical feature maps and produce an output with the same spatial dimension as the input.

Rest of the Model. We adopt the Data Fitting Branch from [40], which is based on residual channel attention block [40]. Trained from scratch, this second branch complements the first one by exclusively working on the domain of target data. With no down-sampling and up-sampling operations, this branch operates in the full-resolution mode, thus extracts features distinct from the ones obtained by the first branch. A simple yet insightful fusion tail consisting of a reflection padding layer, a 7×7 convolutional layer and the Tanh activation [37] combines the features from two branches and produces dehazed images.

3.3. Loss Functions

Since our method mainly focuses on the data-centric engineering and implementing transformers, we follow [15, 37] to adopt a combination of several losses for training our model.

Smooth L1 Loss. For image fidelity reconstruction, the smooth L1 loss [17] has been proved to be more robust than the MSE loss in various image restoration tasks [41]. The formulation follows:

$$L_{l_1} = \frac{1}{N} \sum_i^N smooth_{L_1}(y_i - f_\theta(x_i)), \quad (2)$$

$$smooth_{L_1}(z) = \begin{cases} 0.5z^2 & \text{if } |z| < 1 \\ |z| - 0.5 & \text{otherwise,} \end{cases} \quad (3)$$

where x_i and y_i denote the i-th pixel of clean and hazy images, respectively. N is the total number of pixels. $f_\theta(\cdot)$ represents the network.

MS-SSIM Loss. Multi-scale Structure similarity (MS-SSIM) is based on the assumption that human eyes are adapted for extracting structural information, and therefore a metric of evaluating structural similarity can provide a good approximation to perceived image quality. Let O and G represent two windows centered at the i-th pixel in the dehazed image and the ground truth image, respectively. Gaussian filters are applied on both windows, and produce resulting values of corresponding means (μ_O, μ_G), standard deviations σ_O, σ_G, and covariance σ_{OG}. The SSIM formu-

lation for the i-th pixel follows:

$$\text{SSIM}(i) = \frac{2\mu_O\mu_G + C_1}{\mu_O^2 + \mu_G^2 + C_1} \cdot \frac{2\sigma_{OG} + C_2}{\sigma_O^2 + \sigma_G^2 + C_2}, \quad (4)$$

where C_1 and C_2 help stabilize the division.

Perceptual Loss. Besides pixel-scale supervision on perceptual quality, we adopt ImageNet [12] pre-trained VGG16 [32] to measure perceptual similarity, which helps reconstruct finer details [43]. Denoting x and y as hazy inputs and ground truth images respectively, the loss is defined as:

$$L_{perc} = \frac{1}{N} \sum_j \frac{1}{C_j H_j W_j} ||\phi_j(f_\theta(x)) - \phi_j(y)||_2^2, \quad (5)$$

where $f_\theta(x)$ is the dehazed image. $\phi_j(\cdot)$ denotes the feature map. We choose L_2 loss to measure the distances between them. N denotes the number of features.

Adversarial Loss. Compensating for the risks that pixel-wised loss functions fail to provide sufficient supervision when training on a small dataset, we employ the adversarial loss [43]:

$$L_{adv} = \sum_{n=1}^N -logD(f_\theta(x)), \quad (6)$$

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

Total Loss. The total loss is a weighted sum of aforementioned four components with pre-defined weights:

$$L = L_{l_1} + 0.5L_{\text{MS-SSIM}} + 0.01L_{perc} + 0.0005L_{adv}. \quad (7)$$

4. Experiments

In this section, we first introduce the datasets used to conduct experiments along with implementation details. Then, we conduct ablation studies to verify the effectiveness of our model design and data-preprocessing method. Finally, we evaluate the dehazing results of our proposed method qualitatively and quantitatively, and compare with several state-of-the-art methods.

4.1. Datasets

O-HAZE. With the help of the professional haze machine that generate real haze, O-HAZE [5] was published in 2018, containing 45 clean and hazy image pairs in total. Each pair has a unique spatial resolution. We conduct our evaluation based on the official train, and test split [2].

DENSE-HAZE. DENSE-HAZE [3, 4] was introduced with the NTIRE2019 challenge. It characterizes in dense and homogeneous haze. The dataset contains 45 training images, 5 validation images and 5 test images. All images
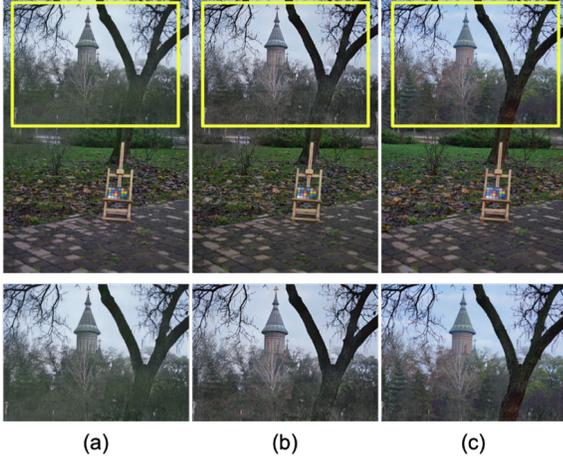
Figure 4. Qualitative ablation study on the data-centric design. (a) the results generated by the model trained on NH-HAZE20+21+23 GC, (b) the results generated by the model trained on NH-HAZE20+21+23 RGB, and (c) ground truth.

are of the same 1600×1200 dimension. In our experiments, we follow the official train, val and test split.

NH-HAZE20 & NH-HAZE21. In NTIRE2020 [6] and NTIRE2021 [7] challenges, NH-HAZE20 and NH-HAZE21 were released. The haze pattern in these two datasets is non-homogeneous. The images in these two datasets are of the size of 1600 × 1200. NH-HAZE20 contains 45 training data, 5 validation data and 5 testing data. We adopt the official train, and test split to conduct experiments on NH-HAZE20. For NH-HAZE21, we take the first 20 training images as our training set, and the rest 5 images are used for testing.

NH-HAZE23. Inheriting the non-homogeneous haze style from previous years, NTIRE2023 introduces 50 image pairs, each of a much higher resolution of 4000 × 6000. The increase in image size leads to larger volume of training data and greater demand in computation resources. As the ground truth images of the 5 validation data and 5 test data are not public so far, we can only make use of the 40 train data when not evaluating on the server. We adopt different train/test splitting strategies for performing quantitative comparisons of SOTA methods and ablation studies. For methods comparison, we choose the first 35 images of the official training set as our training data, and the rest 5 are used for testing. For ablation study, we use all 40 pairs to perform training, and get testing scores using the online validation server of the challenge.

4.2. Implementation Details

The input images are randomly cropped to a size of 256 × 256, and augmented by several data augmentation strategies, including 90, 180, 270 degrees of random ro-

Table 1. Ablation study for architectures and data-preprocessing techniques. The scores are evaluated using NTIRE2023 online validation server.

Data	Res2Net+RCAN		Ours	
	PSNR	SSIM	PSNR	SSIM
NH-HAZE23 only	20.68	0.678	21.54	0.682
NH-HAZE20+21+23	20.86	0.688	21.54	0.689
NH-HAZE20+21+23 GC	21.08	0.690	21.58	0.693
NH-HAZE20+21+23 RGB	**21.26**	**0.693**	**21.94**	**0.697**

tation, horizontal flip, and vertical flip. Note that, we do not apply any augmentation strategy related with brightness or color change as we have no intention to jeopardize the adjusted color distributions produced by our data-preprocessing method. We use the AdamW [25] ($\beta_1 = 0.9, \beta_2 = 0.999$) as our optimizer. The learning rate is initially set to $1e^{-4}$ and decreased to $1e^{-6}$ with a cosine annealing strategy. We implement with the PyTorch library [28] on two Nvidia Titan XP GPUs. Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) are used as two metrics for quantitative evaluation.

4.3. Ablation Study

We conduct comprehensive ablation studies to analyze and demonstrate the effectiveness of our data-preprocessing method and proposed network architecture.

4.3.1 Importance of Data-Centric Design.

In Section 3.1, we emphasize the importance of data for succeeding in non-homogeneous dehazing. To further demonstrate, we conduct experiments on several datasets with different data-preprocessing methods. There are four sets of training data, including: 1) **NH-HAZE23 only**: only the data from NTIRE2023 challenge is used; 2) **NH-HAZE20+21+23**: a direct combination of data from NH-HAZE20, NH-HAZE21 and NH-HAZE23 datasets; 3) **NH-HAZE20+21+23 GC**: a combination of data from NH-HAZE20, NH-HAZE21 and NH-HAZE23 with the GT data from NH-HAZE20 and 21 being processed by gray-scale gamma correction as [15, 37]; 4) **NH-HAZE20+21+23 RGB**: a combination of data from NH-HAZE20, NH-HAZE21 and NH-HAZE23 with both the hazy and GT data from NH-HAZE20 and 21 being processed by our proposed approach described in Section 3.1. We employ these sets of data on two different model architectures. The first model is from the NTIRE2021 challenge [37] (we refer to as Res2Net+RCAN), where Res2Net [16] is adopted as the backbone. The second model is our proposed one introduced in Section 3.2. In total, we conduct 8 individual experiments, and report their best results (in terms of the PSNR index) evaluated on the NTIRE2023 online validation sever. The results are shown in Table 1.
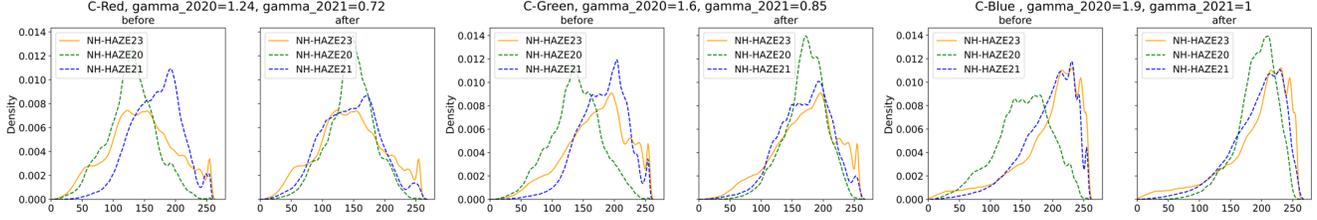
Figure 5. Comparison of RGB-wise distribution of datasets (hazy) before and after being processed by our proposed method.

By comparing the first and second row of Table 1, we find that the direct combination of all the available data yields limited improvements for both our model and [37]. By comparing the last two rows with the second row in Table 1, it can be observed that performing data-preprocessing is generally beneficial. Not surprisingly, the models trained on our dataset achieve the best performance. These results reinforce the importance of data-centric engineering.

To qualitative evaluate the importance of data engineering, we show in Figure 4 by the images generated from the models respectively trained on **NH-HAZE20+21+23 GC** and **NH-HAZE20+21+23 RGB**. By comparing the two results with ground truth, it is obvious that the model trained on our processed dataset can generate more faithful results in terms of color and brightness. Specifically, the colors of the building and trees of ours are much more in line with those in the ground truth, while the compared one tend to generate green objects.

4.3.2 Effectiveness of Transformer

As noted in Section 3.2, our network is built upon the recent work [37]. The main difference is that we replace the Res2Net branch in [37] with the proposed Transformer-based structure. By quickly check the performance of our method and that of Res2Net+RCAN on four datasets in Table 1, it can be easily observed that our method always outperforms Res2Net+RCAN by a significant margin. This illustrate the effectiveness of using Transformer in non-homogeneous dehazing.

4.4. Further Analysis on Data-Centric Engineering

In Figure 2, we show the distribution change of the ground truths after applying the proposed RGB gamma correction. In Figure 5, we provide the distribution change of NH-HAZE20 and 21 hazy images as a supplement. It can be observed that after our data-preprocessing, the distribution of the three image channels (RGB) of the NH-HAZE20 and 21 hazy images are more in line with those of NH-HAZE23 data. Figure 6 further qualitatively illustrates the images before and after data-preprocessing. The results shows that colors of the processed NH-HAZE20 and NH-HAZE21 data are much more similar to that of NH-
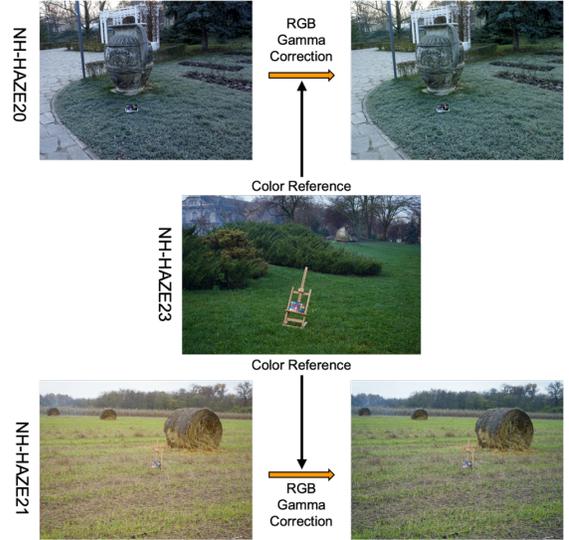

Figure 6. Examples from NH-HAZE20 and NH-HAZE21 datasets to visually showcase the color correction.

HAZE23. We emphasize that this data-centric engineering is the key that helps our method stand out in the competition. Based on both the analysis in this section and Section 4.3.1, we conclude that the data quality is one of the determining factors, possibly the most important one under the NTIRE dehazing challenges.

4.5. Comparisons with the State-of-the-art Methods

To conduct comparisons, we select five state-of-the-art methods, including DCP [18], AOD-Net [21], GCANet [11], FFA [29], and Res2Net+RCAN [37].

In Table 2, we illustrate the best PSNR and SSIM indexes of each method on five different datasets. The methods adopting the two-branch framework perform generally well on all the datasets, where in Figure 7, Res2Net+RCAN and our method can produce visual pleasing results on all datasets. They unveil significantly better performance when dealing with non-homogeneous haze patterns, as we could observe from the results on NH-HAZE20, 21 and 23. Therefore, the two branch framework remains dominating in the limited data scenarios.

(a) DENSE-HAZE

(b) NH-HAZE20

(c) NH-HAZE21

(d) NH-HAZE23

Figure 7. Qualitative evaluation on the four representative datasets, i.e., DENSE-HAZE, NH-HAZE20, NH-HAZE21 and NH-HAZE23. For DENSE-HAZE and NH-HAZE20, we follow the official train, val and test split. For NH-HAZE21 and NH-HAZE23, due to the unavailable of test data, we split the released official training data to our training set and test set.

Table 2. Quantitative evaluation on DENSE-HAZE, NH-HAZE20, NH-HAZE21 and NH-HAZE23 datasets. The best results are marked in **bold**, and the second bests are marked with underlines.

Methods	O-HAZE		DENSE-HAZE		NH-HAZE20		NH-HAZE21		NH-HAZE23	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
DCP	12.92	0.505	10.85	0.404	12.29	0.411	11.30	0.605	11.87	0.470
AOD	17.69	0.616	13.30	0.469	13.44	0.413	13.22	0.613	12.47	0.369
GCANet	19.50	0.660	12.42	0.478	17.58	0.594	18.76	0.768	16.36	0.512
FFA	22.12	0.768	16.26	0.545	18.51	0.637	20.40	0.806	18.09	0.585
Res2Net+RCAN	<u>25.54</u>	<u>0.783</u>	**16.36**	**0.582**	**21.44**	<u>0.704</u>	<u>21.66</u>	**0.843**	<u>20.11</u>	<u>0.627</u>
Ours	**25.98**	**0.789**	<u>16.31</u>	<u>0.561</u>	**21.44**	**0.710**	**21.67**	<u>0.838</u>	**20.53**	**0.636**

It is worth noticing that our model substantially outperforms the Res2Net+RCAN model only on O-HAZE and NTIRE2023. We argue the reason behind is that due to the huge increase of image resolution on O-HAZE and NH-HAZE23 datasets. For example, the number of pixels in NH-HAZE23 data is 6.25 times larger than that of the combination of NH-HAZE20 and NH-HAZE21 datasets. Since our transformer-based model contains more learnable parameters, a larger training dataset can essentially alleviate the overfitting problem. This phenomenon further indicates that when it comes to a limited data setting, it is more critical to investigate in a data-centric manner other than simply improving the model's capacity.

5. Conclusion

In this paper, we propose a method targeting on non-homogeneous dehazing. It consists of a data-preprocessing strategy inspired by data-centric AI and a Transformer based two-branch model structure. Combining them together, we construct a solution that outperforms the SOTA methods, which stimulates our advocation on treating the model and the data equally important. Additionally, extensive experimental results provide strong support to the effectiveness of our method.

References

[1] Data-centric ai competition submission guide, 2021. 2, 3

[2] Cosmin Ancuti, Codruta O Ancuti, and Radu Timofte. Ntire 2018 challenge on image dehazing: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 891–901, 2018. 5

[3] Codruta O. Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE International Conference on Image Processing (ICIP)*, IEEE ICIP 2019, 2019. 5

[4] C. O. Ancuti, C. Ancuti, R. Timofte, L. Van Gool, L. Zhang, M. Yang, T. Guo, X. Li, V. Cherukuri, V. Monga, H. Jiang, S. Yang, Y. Liu, X. Qu, P. Wan, D. Park, S. Y. Chun, M. Hong, J. Huang, Y. Chen, S. Chen, B. Wang, P. N. Michelini, H. Liu, D. Zhu, J. Liu, S. Santra, R. Mondal, B. Chanda, P. Morales, T. Klinghoffer, L. M. Quan, Y. Kim, X. Liang, R. Li, J. Pan, J. Tang, K. Purohit, M. Suin, A. N. Rajagopalan, R. Schettini, S. Bianco, F. Piccoli, C. Cusano, L. Celona, S. Hwang, Y. S. Ma, H. Byun, S. Murala, A. Dudhane, H. Aulakh, T. Zheng, T. Zhang, W. Qin, R. Zhou, S. Wang, J. Tarel, C. Wang, and J. Wu. Ntire 2019 image dehazing challenge report. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2241–2253, 2019. 5

[5] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *IEEE Conference on Computer Vision and Pattern Recognition, NTIRE Workshop*, NTIRE CVPR'18, 2018. 5

[6] 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. 2, 3, 6

[7] 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. 2, 3, 6

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

[9] 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. 2

[10] 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, 2

[11] 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 (WACV)*, pages 1375–1383, 2019. 1, 2, 7

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 4, 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[14] Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1):1–14, 2014. 2

[15] 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 (CVPR) Workshops*, pages 203–212, June 2021. 1, 2, 3, 4, 5, 6

[16] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 6

[17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 5

[18] 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, 2, 7

[19] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 4

[20] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[21] 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, 2, 7

[22] 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. 1, 2, 3

[23] Jing Liu, Haiyan Wu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Trident dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[24] 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. 2, 4

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] William Edgar Knowles Middleton. *Vision through the atmosphere*. University of Toronto Press, 1952. 2

[27] Mohammad Motamedi, Nikolay Sakharnykh, and Tim Kaldewey. A data-centric approach for training deep neural networks with less data, 2021. 2, 3, 4

[28] 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

[29] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11908–11915, Apr. 2020. 1, 2, 7

[30] 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

[31] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2020. 3

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

[33] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning, 2018. 4

[34] Robby T Tan. Visibility in bad weather from a single image. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[35] Haiyan Wu, Jing Liu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Knowledge transfer dehazing network for nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[36] Haiyan Wu, Jing Liu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Knowledge transfer dehazing network for nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 478–479, 2020. 1, 2, 4

[37] Yankun Yu, Huan Liu, Minghan Fu, Jun Chen, Xiyao Wang, and Keyan Wang. A two-branch neural network for nonhomogeneous dehazing via ensemble learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2021. 1, 2, 3, 4, 5, 6, 7

[38] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. Data-centric ai: Perspectives and challenges, 2023. 3, 4

[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. 1, 2

[40] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2, 5

[41] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 5

[42] Zhaorun Zhou, Zhenghao Shi, Mingtao Guo, Yaning Feng, and Minghua Zhao. Cggan: A context guided generative adversarial network for single image dehazing, 2020. 1, 2

[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. 5

[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. 2