Vision Transformers for Single Image Dehazing

Yuda Song, Zhuqing He, Hui Qian, and Xin Du

Abstract—**Image dehazing is a representative low-level vision task that estimates latent haze-free images from hazy images. In recent years, convolutional neural network-based methods have dominated image dehazing. However, vision Transformers, which has recently made a breakthrough in high-level vision tasks, has not brought new dimensions to image dehazing. We start with the popular Swin Transformer and find that several of its key designs are unsuitable for image dehazing. To this end, we propose DehazeFormer, which consists of various improvements, such as the modified normalization layer, activation function, and spatial information aggregation scheme. We train multiple variants of DehazeFormer on various datasets to demonstrate its effectiveness. Specifically, on the most frequently used SOTS indoor set, our small model outperforms FFA-Net with only 25% #Param and 5% computational cost. To the best of our knowledge, our large model is the first method with the PSNR over 40 dB on the SOTS indoor set, dramatically outperforming the previous state-of-the-art methods. We also collect a large-scale realistic remote sensing dehazing dataset for evaluating the method's capability to remove highly non-homogeneous haze. We share our code and dataset at https://github.com/IDKiro/DehazeFormer.**

Index Terms—**Image processing, image dehazing, deep learning, vision transformer.**

I. Introduction

HAZE is a common atmospheric phenomenon that can impair daily life and machine vision systems. The presence of haze reduces the scene's visibility and affects people's judgment of the object, and thick haze can even affect traffic safety. For computer vision, haze degrades the quality of the captured image in most cases. It can impact the model's reliability in high-level vision tasks, further mislead machine systems, such as autonomous driving. All these make image dehazing a meaningful low-level vision task.

Image dehazing aims to estimate the latent haze-free image from the observed hazy image. For the single image dehazing problem, there is a popular model [1], [2], [3] to characterize the degradation process for hazy images:

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

where I is the captured hazy image, J is the latent haze-free image, A is the global atmospheric light, and t is the medium transmission map. And the transmission can be expressed as

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

where β is the scattering coefficient of the atmosphere, and d is the scene depth. As can be seen, image dehazing is a typically ill-posed problem, and early image dehazing methods tend to constrain the solution space with priors [4], [5], [6], [7]. They generally estimate A and $t(x)$ separately to lower the complexity of the problem and then use Eq.(1) to derive the results. These prior-based methods can produce images with good visibility. However, these images are often visibly different from haze-free images, and artifacts may be introduced in regions that do not satisfy the priors.

In recent years, deep learning has made a big hit in computer vision, and researchers have proposed a large number of image dehazing methods based on deep convolutional neural networks (CNNs) [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. With a sufficient number of synthetic image pairs, these methods can achieve superior performance over prior-based methods. Earlier CNN-based methods [8], [9], [10] also estimate A and $t(x)$ separately, where $t(x)$ is supervised using the transmission map used in synthesizing the dataset. And current methods [13], [14], [15], [16], [17], [18], [19], [20], [21] prefer to predict the latent haze-free image or the residuals of the haze-free image versus the hazy image since it tends to achieve better performance. Very recently, ViT [23] outperformed almost all CNN architectures in high-level vision tasks using plain Transformer architecture. Subsequently, many modified architectures [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41] have been proposed, and vision Transformer is challenging the dominance of CNNs in high-level vision tasks. So many works have demonstrated the effectiveness of vision Transformers, but there is still no Transformer-based image dehazing method that defeats the state-of-the-art image dehazing networks. In this work, we propose an image dehazing Transformer dubbed DehazeFormer, which is inspired by Swin Transformer [31]. It dramatically surpasses these CNN-based methods.

We find that the LayerNorm [42] and GELU [43] commonly used in vision Transformers harm the image dehazing performance. Specifically, the LayerNorm used in vision Transformer normalizes the tokens corresponding to the image patches separately, resulting in the loss of the relativity between the patches. Hence, we remove the normalization layer preceded by the multi-layer perceptron (MLP) and propose RescaleNorm to replace LayerNorm. RescaleNorm performs normalization on the entire feature map and reintroduces the mean and variance of the feature map lost after normalization. Besides, SiLU / Swish [44] and GELU work well in high-level vision tasks, but ReLU [45] works better

Manuscript received 11 April 2022; revised 16 January 2023 and 8 February 2023; accepted 25 February 2023. Date of publication 17 March 2023; date of current version 24 March 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jiaying Liu. *(Yuda Song and Zhuqing He contributed equally to this work.) (Corresponding author: Xin Du.)*

Yuda Song and Hui Qian are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China.

Zhuqing He and Xin Du are with the College of Information Science and Electronic, Zhejiang University, Hangzhou 310027, China (e-mail: duxin@zju.edu.cn).

Digital Object Identifier 10.1109/TIP.2023.3256763