Transformer based Image Dehazing

Patricia L. Suárez[1], Dario Carpio[1], Angel D. Sappa[1,2] and Henry O. Velesaca[1]

[1]*ESPOL Polytechnic University*
Guayaquil, Ecuador
{plsuarez, asappa, dncarpio, hvelesac}@espol.edu.ec

[2]*Computer Vision Center*
Barcelona, Spain
asappa@cvc.uab.es

***Abstract*—This paper presents a novel approach to remove non homogeneous haze from real images. The proposed method consists mainly of image feature extraction, haze removal, and image reconstruction. To accomplish this challenging task, we propose an architecture based on transformers, which have been recently introduced and have shown great potential in different computer vision tasks. Our model is based on the SwinIR an image restoration architecture based on a transformer, but by modifying the deep feature extraction module, the depth level of the model, and by applying a combined loss function that improves styling and adapts the model for the non-homogeneous haze removal present in images. The obtained results prove to be superior to those obtained by state-of-the-art models.**

***Index Terms*—atmospheric light, brightness component, computational cost, dehazing quality, haze-free image**

I. Introduction

Image dehazing is a process of removing the haze from an image, and can be done using a variety of methods. One common method has been to use classical approaches like a Haar wavelet transform to decompose the image into a series of wavelets and then use a thresholding technique to remove the haze. This method is however not very effective in removing the haze from the given image. The goal of dehaze is to obtain an image that appears clear and sharp. Clear images are necessary to apply other image processing and computer vision algorithms, like edge detection, segmentation, and object recognition, just to mention a few. The haze formulation is a simple model that simulates the effects of atmospheric haze. Also, it assumes that the atmospheric haze is an homogeneous medium with a uniform scattering coefficient. A simple image haze model [1] can be expressed as $I = J(x)t(x) + A(1-t(x))$; where I is the given hazy image, J is the latent haze-free image, A is the global atmospheric light, and t is the medium transmission map, which is based on the dispersion coefficient of the haze and the distance of the object from the camera [2].

Haze is caused by light scattering in the atmosphere, and can reduce the clarity and contrast of an image. It makes the image appear hazy and unclear. The atmospheric scattering is caused by the existence of small particles in the air, for instance fog, dust, water droplets, and smoke. These small particles in the air scatter the light in the atmosphere and reduce the amount of light that reaches the camera sensor [3]. Also, haze can cause a loss of contrast and color saturation in an image, as well as a reduction in overall clarity. Also in many cases, haze can completely obscure an image.

Identify applicable funding agency here. If none, delete this.

Recently, techniques for removing haze from images have been made using a convolutional neural network (CNN), which has proved to be very effective to remove haze from images [4]. CCN approaches are much more precise than the Haar wavelet transform-based methods and can remove the haze from the image without losing any of the details in the image. Although CCN networks are still booming, many authors have started transformer-based models, because these networks allow modeling data with high dependence on input using values large receptive fields [5]. Additionally, these models support parallel processing and multiple data modalities (e.g., images, video, text, and voice). Transformer architectures are based on a self-service mechanism, which allows the model to service different positions of the integrated linear input simultaneously.

The contribution of this paper can be pointed out in:

- The usage of Charbonnier Loss [6] to enhance image reconstruction during the haze removal process. This loss is most robust than $l1$ or $l2$ loss because it is less sensitive to outliers values and helps to incorporate more image details in the resulting haze-free image. In our approach, we combine this loss with structure similarity and adversarial loss.
- The modification of the attention mechanism of the encoder to extract not only the information based on pixels but also the spatial information, considering that the haze present in the images is non homogeneous.

The manuscript is organized as follows. Section II presents works related to the haze removal problem and some basic concepts of transformer networks. Section III presents the proposed haze removal architecture. Experimental results and comparisons with different implementations are given in Section IV. Finally, conclusions are presented in Section V.

II. Related Word

Several approaches for image dehazing have been proposed in the last decade, some of them are based on classical approaches, like those related to the estimation of map of transmission and atmospheric light [7], [8], [9]. Another classical technique related to fast visibility restoration has been presented by Fattal et al. [10], where the authors propose a color-line method based on the distribution often presented in real images showing regularity in image regions' pixels.

Another method proposed to improve the visibility of night haze is based on the use of synthetic night scenes guided by clear images, from which the spatial information, light, and reflectance are extracted to generate the haze effects [11]. Another well-known classical approach to remove haze is by using as a prior the dark channel [12]. In Berman et al. the authors propose an algorithm based on a new non-local prior since the degradation is non homogeneous on each pixel and changes according to the distance the camera sensor from the scene [8]. Continuing with the methods for estimating atmospheric light, in [13] a method for haze removal is presented. This method is based on the linear transformation expressing that a linear relation exists in the minimum channel between the haze and its corresponding clear image. Also, the average gray values of the pixels and the local gradients are used as evaluation criteria for the estimation of the transmission map. This focuses on the brightest areas of distortion to solve the haze removal problem.

There are some proposals based on the minimization of energy functions, for example in Galdran et al. [14] the authors propose the optimization of energy functions fused together to remove the haze present in an image. With this method, the authors introduce a method to enhance the contrast and saturation of the haze image while maintaining spatiality congruence in the image.In [15] The authors propose a method to remove haze based on ellipsoids generated based on statistical data such as the mean and variance of the pixels corresponding to the regions with haze. Another classical method to address the problem of image degradation in foggy weather is proposed by Wang et al. [16]. It is a technique based on a physical model and a multiscale retinex that works with the brightness components to restore the colors of the image. The method considers three components, the calculation of the atmospheric light, the radiation of the foggy scene, and the estimation of the transmission map taking into account the dynamic range of the image to include more accurately the details of the scenes.

Most recent approaches are based on the usage of Convolutional Neural Networks (CNN), for instance, in Engin et al. [17] the authors propose a cyclic GAN model that performs domain transfer between unmatched images. It uses the well-known contextual and consistency losses to extract the haze-free regions of the images and combine them with the texture information of the light images to obtain the haze-free images. In [18] the authors present an adaptive deep convolutional network that removes haze using a light contrastive regularizer.

In [19] the authors present a trained end-to-end network that is responsible for estimating the average transmission maps. With the estimated map applying the atmospheric dispersion model, the haze-free image is obtained. This network, called DehazeNet applies layer design to generate feature maps with the most relevant information on haze images. Additionally, an activation function called bilateral rectified linear unit is proposed in this work. With this new activation function, the quality of the generated clear image is improved. Another CNN-based approach has been proposed by Ren et al. [2] to perform the mapping of the information domain between images with haze and its corresponding transmission map. The prediction of the transmission map performed by the model is combined with a refinement process that works by local regions in the images.

Although the CCN networks still present good results in their recently proposed techniques. Also at the same time, techniques based on transformer networks have appeared. Researchers have shown interest since this type of network can be considered general purpose. These models support parallel data processing, thus creating efficient blocks to transform information. One of the approaches of this type of models are those related to computer vision that have obtained better results and specialize in the implementation of attention and data modules [20]. With this model scaling, the relationship between the error rate of the data and their corresponding calculations can be established. This allows the architecture to reduce memory consumption and increase the accuracy of the resulting models. There is also the class of transformer networks that are dedicated to spatial dimension conversion and its effectiveness. These techniques focus on the reduction of spatial dimension applied to a transformer architecture. The reduction is designed based on a grouping (PiT) Pooling-based Vision Transformer, on the original ViT (Vision Transformer) model. This grouping manages to improve generalization for image classification and/or object detection tasks [21]. Another transformer model adapted for computer vision is presented in Liu et al. [22], where the authors introduce a general purpose network called Swin transformer. This technique is challenging due to the great diversity of resolution present on the images, as well as the different scales of the images. For which, local compensation windows are used that avoid overlapping and represent a hierarchical transformer model. This model will vary in complexity depending on the size of the image. This architecture is well suited to solve computer vision problems, such as image classification, object detection, and semantic segmentation. In Liang et al. [23] a model to restore images based on the Swin Transformer network is proposed. Image restoration is performed in three steps: surface feature extraction, deep feature extraction, and actual reconstruction. Swin transformer blocks with a residual connection are used for the extraction of the most relevant characteristics. This model has been also evaluated for reconstruction tasks, image noise reduction, and JPEG compression artifact reduction.

III. Proposed Approach

Removing a non-homogeneous haze present in an image is a complex challenge since the distribution of the existing particles in the air is not always the same. Therefore, the model that is designed cannot assume an homogeneous global removal of the haze. The feature extractor should consider this pattern when obtaining the most relevant features.

Considering the aforementioned problem, in the present work a modified transformer network is proposed, based on the SwinIR [23] model see Fig. 1. The changes implemented starts with the addition of more levels of self-service headers. The limits of the activation function in the Multilayer Perceptron

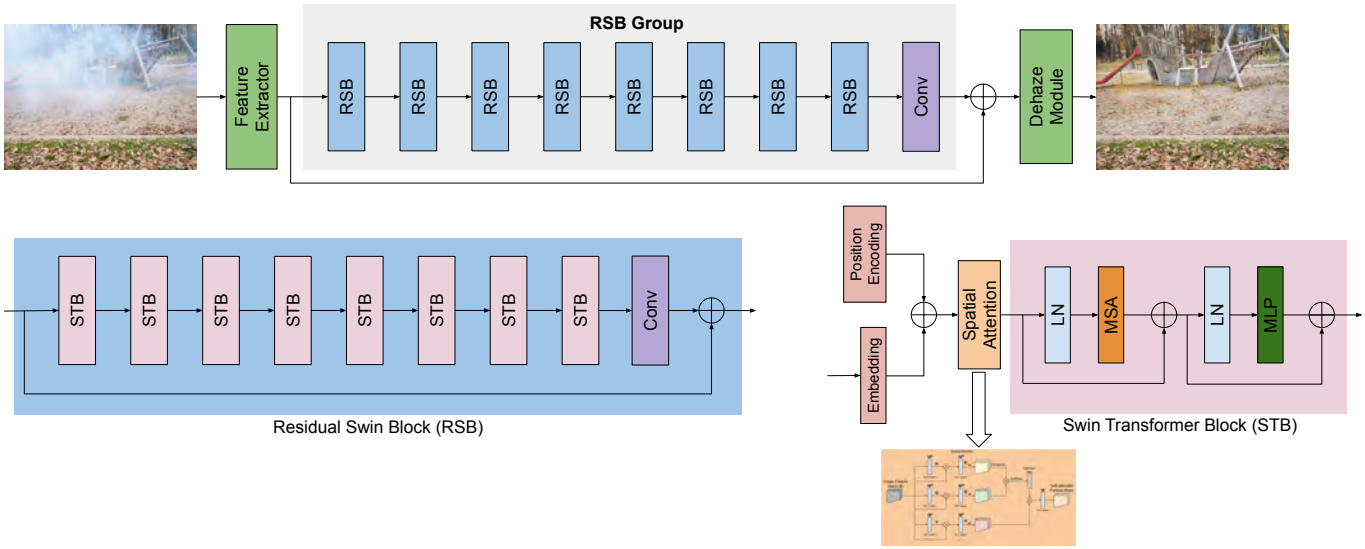

Fig. 1. Proposed image dehaze transformer based model.

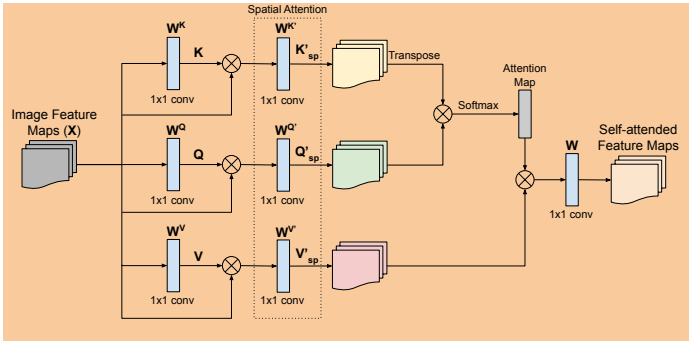

Fig. 2. Attention module.

(MLP) layer have also been modified. With the proposed changes, the construction of the feature maps will allow multiple resolution so that they can converge on sharper and more precise regions. In addition, the modifications implemented in the construction of the embedded vectors manage to capture the order relationships between patches (spatial information). This is crucial in a haze removal process as spatial smoothing is provided to the extracted descriptors in the model.

Additionally, our transformer encoder has been modified to enhance the low-level task of removing non-homogeneous haze, because our self-attention mechanism has a behavior restricted to learning the interplay across local regions on the images with haze and are more robust against data corruptions, image occlusions, and high-frequency noises.

According to Vaswani et al. [5], the scaled dot product attention is an attention mechanism where a set of embedded inputs named queries Q, keys K of dimension dk, and values V of dimension dv. simultaneously packed into arrays Q, K and V respectively are scaled down by $\sqrt{d_k}$. The function can be formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

The outputs of the vectors of keys, queries and values will be processed by a parallel attention method, generating multiple dimensional outputs according to the corresponding vectors inputs. All these outputs are then grouped and projected to obtain the definitive results.

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right) W^O \\ \text{where } \text{head}_\text{i} &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)\end{aligned} \tag{2}$$

where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$

In our model, we have modified the input embedding to enhance the attention map. This is one of the components of the self-attended feature map. The encoder will receive projections of queries Q, keys K and values V with their corresponding spatial information aggregated. First, we divide each image into patches, then each patch is flattened to generate lower-dimensional linear embeddings to reduce dimensionality. Our contribution is to add to each patch, its corresponding most relevant spatial information patch. The objective is to generate a map that contains the interspatial relationship extracted from the features of each image patch. This additional spatial information will be added to input vectors previous to generate linear embedding input of the transformer encoding. Figure 2 show the spatial attention implementation.

To obtain the spatial information, average and maximum grouping operations are applied to each image feature vector

for each channel of the patch. The outputs of these grouping operations are then concatenated. The main idea is to use them as feature descriptors. Next, a convolution is performed over the obtained descriptors, resulting in the sought spatial attention map [24].

The spatial attention map is obtained as follows:

$$\begin{aligned}\mathbf{SMap}(\mathbf{P}) &= \delta\left(fct^{s\times s}([\mathrm{AvgPool}(\mathbf{P}); \mathrm{MaxPool}(\mathbf{P})])\right) \\ &= \delta\left(fct^{sg\times sg}\left([\mathbf{P}\mathrm{avg}^{\mathbf{smap}}; \mathbf{P}\mathrm{max}^{\mathbf{smap}}]\right)\right),\end{aligned} \tag{3}$$

the sigmoid function represented by δ and the pooling operation denoted by $(fct^{sg\times sg})$, where $f^s \times f^s$ is the filter size.

Additionally, in our research, the use of multiple losses is proposed. In order to enhance the quality of the generated hazy free image. Futhermore, in order to optimize the learning process and accelerate the model generalization, this multiple loss can be defined as;

$$L_{\text{multiple}} = \alpha \times L_{\text{Charbonier}} + \beta \times (1 - L_{SSIM}) + \lambda \times L_{Identity} \tag{4}$$

where L_{Char} is the is the Charbonnier loss [6], $(1 - L_{SSIM})$ is the negative structure similarity loss and L_{Ide} is the identity loss.

The α, β and λ loss coefficients are set to 0.4, 6, and 0.65, respectively; these values have been empirically obtained.

The Charbonier loss has been used to compute the pixel loss between the predicted image and the ground truth:

$$L_{Charbonier} = \sqrt{\|X_S - Y\|^2 + \varepsilon^2} \tag{5}$$

where Y represents the ground-truth image, X_S is the hazy free image obtained from the last layer of the generator. ϵ is the region for which the loss function changes from being approximately quadratic to approximately linear, and for our model, this parameter was set to 0.00016.

Also, the loss known as structural similarity [25] L_{SSIM} is used to evaluate whether the content of two images have structural correspondence. It helps to evaluate changes in local structures, since the human visual perception system is sensitive to these changes. The idea behind this loss function is to help the learning model to obtain an enhanced image. This loss can be expressed as:

$$-L_{SSIM} = -\,\mathrm{SSIM}\left(X_s, Y\right). \tag{6}$$

where Y represents the ground-truth image, X_S is the hazy free image obtained from the last layer of the generator.

Finally, the identity loss evaluates whether the generated haze-free image matches the target clear image. That is, they share the same data domain and do not degrade the pixel intensity levels of the resulting haze-free image, and the mathematical expression can be expressed as:

$$L_{\text{Identity}} = L_{\text{Identity}}\left(Y_S, Y\right), \tag{7}$$

where Y_s is the prediction result obtained by inputting the Y hazy image into the network.

Also, in this paper we use $h = 8$ parallel attention layers, to enhance the global feature extractor. For each of these we use $d_k = d_v = d_{\text{model}}\ /h = 32$.

TABLE I
RESULTS FROM THE VALIDATION DATASETS (NH-HAZE 2020/2021). BEST RESULTS IN **BOLD**, AND SECOND BEST UNDERLINED.

Approaches	NH-Haze 2020		NH-Haze 2021	
	PSNR	SSIM	PSNR	SSIM
SwinIR+L_1	17.3552	0.5473	17.1163	0.7044
SwinIR+$L_{Charbonier}$	<u>17.9170</u>	<u>0.6065</u>	<u>19.4660</u>	<u>0.8000</u>
Ours+$L_{Charbonier}$	17.5055	0.5695	17.9842	0.7520
Ours+$L_{Multiple}$	**18.4038**	**0.6371**	**20.1259**	**0.8273**

IV. EXPERIMENTAL RESULTS

This section firstly describes the data set that has been used to validate the proposed approach. Then, training settings details are provides. Finally, results from the proposed approach and comparisons with state of the art technique are provided, both quantitative and qualitative.

A. Datasets

A non homogeneous realistic dataset has been used to evaluate the proposed approach. It contains pairs of real hazy and free haze images introduced by Ancuti et al. [26]. This dataset has been used in the NTIRE 2020 dehazing challenge and the extended version used in NTIRE 2021 dehazing challenge. The dataset contains 55 and 25 pairs of images respectively with their corresponding haze-free images. A set of 5 images has been selected from each dataset, NH-Haze 2020 and NH-Haze 2021. These images have been used to compute the metrics to evaluate results. To carry out the experiments, from the total of 80 image pairs, 60 pairs were selected for the training, 10 images have been used for testing, and the other 10 for validation.

B. Pre-Procesing and Training Settings

For transformer-based approaches, especially those related to computer vision, tokenization is used as a pre-process of the dataset. For this pre-processing we have divided the input images into regions with a size of 32 $\times$ 32. These regions are referred to as visual tokens, which are embedded in vectors of predetermined fixed dimension. The patch position is also embedded along with the image regions. These embedded vectors will be the input of the transformer model. For our paper, we have added an embedded vector that contains the most representative spatial information of each patch. With this information the model feature extractor can improve the quality of the generated haze-free images. Additionally, the model uses an initial learning rate of 0.00028 and ends with a value of 0.00017 when reaching local loss minimum values. For the training, we use Adam optimizer, β_1 and β_2 are initialized with 0.92 and 0.98, respectively. Quantitative evaluation is performed with PSNR and SSIM.

Fig. 3. Experimental results: 1st. row depicts input images; 2nd. to 3th. rows show results of state-of-the-art approaches; 4th. row shows results of the proposed approach with Charbonier Loss; 5th. row shows results of the proposed approach with Multiple Loss; 6th. row shows ground truth images from NH-Haze 2020.

C. Comparisons

The proposed approach has been evaluated and compared with the original model based on Swin et al. The experiments carried out include the validation of the original model with the loss L1 and Charbonier. On the other hand, our approach for comparison purposes has been validated with the Charbonier loss and the proposed multiple loss. The table I presents the results obtained from all approaches. Figures 3 and 4 show qualitative results with all the approaches evaluated with the field images. The proposed approach presents better results with respect to the others. Results of the 5 images in the NTIRE 2021 haze removal challenge validation set are provided in Figure 5.

V. Conclusions

Techniques based on transformer networks are lately achieving results that improve the state of the art techniques based solely on convolutional neural networks. In our paper, we have proposed to improve the spatial self-attention mechanism to make the haze removal process more efficient and enhance the quality of the generated hazy free images. The model used is the SwinIR, and with the contribution made, the input data will not only be the patches and their relative position, but also the most representative spatial information. With this additional information, the feature extractor will have the high-level pixels necessary to remove the haze more efficiently. The best results in haze-free image quality with the proposed approach validate the effectiveness of the model. As a future work, it is thought to improve the Swin transformer self-attention mechanism, to incorporate residual information by

Fig. 4. Experimental results: 1st. row depicts input images; 2nd. to 3th. rows show results of state-of-the-art approaches; 4th. row shows results of the proposed approach with Charbonier Loss; 5th. row shows results of the proposed approach with Multiple Loss; 6th. row shows ground truth images from NH-Haze 2021.

channel and space that improves the score obtained by the encoder and decoder of the transformer-based model.

ACKNOWLEDGEMENTS

This work has been partially supported by the ESPOL Polytechnic University; the Spanish Government under Project PID2021-128945NB-I00; and the "CERCA Programme / Generalitat de Catalunya". The authors gratefully acknowledge the support of the CYTED Network: "Ibero-American Thematic Network on ICT Applications for Smart Cities" (REF-518RT0559) and the NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

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

[2] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 154–169.

[3] W. Wang and X. Yuan, "Recent advances in image dehazing," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 410–436, 2017.

[4] P. L. Suarez, A. D. Sappa, B. X. Vintimilla, and R. I. Hammoud, "Deep learning based single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need 31st conference on neural information processing systems," *(NIPS), Long Beach, CA, USA*, pp. 1–11, 2017.

[6] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proceedings of 1st International Conference on Image Processing*, vol. 2. IEEE, 1994, pp. 168–172.

[7] W. Sun, H. Wang, C. Sun, B. Guo, W. Jia, and M. Sun, "Fast single image haze removal via local atmospheric light veil estimation," *Computers & electrical engineering*, vol. 46, pp. 371–383, 2015.

[8] D. Berman, T. Treibitz, and S. Avidan, "Air-light estimation using haze-lines," in *International Conference on Computational Photography (ICCP)*. IEEE, 2017, pp. 1–9.

Fig. 5. Results from our approach on the validation images of NH-Haze 2021—ground truth of these images are not provided since they were used as benchmark at the challenge of NTIRE 2021, hence only qualitative results are provided.

[9] D. Ngo, S. Lee, and B. Kang, "Robust single-image haze removal using optimal transmission map and adaptive atmospheric light," *Remote Sensing*, vol. 12, no. 14, p. 2233, 2020.

[10] R. Fattal, "Dehazing using color-lines," *ACM transactions on graphics (TOG)*, vol. 34, no. 1, pp. 1–14, 2014.

[11] Zhang J. et al., "Nighttime dehazing with a synthetic benchmark," in *28th ACM International Conference on Multimedia*, 2020, pp. 2355–2363.

[12] S. Lee, S. Yun, J.-H. Nam, C. S. Won, and S.-W. Jung, "A review on dark channel prior based image dehazing algorithms," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 4, pp. 1–23, 2016.

[13] W. Wang, X. Yuan, X. Wu, and Y. Liu, "Fast image dehazing method based on linear transformation," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1142–1155, 2017.

[14] Galdran A. et al., "Fusion-based variational image dehazing," *IEEE Signal Processing Letters*, vol. 24, no. 2, 2016.

[15] T. M. Bui and W. Kim, "Single image dehazing using color ellipsoid prior," *IEEE Trans on Image Processing*, vol. 27, no. 2, 2017.

[16] J. Wang, K. Lu, J. Xue, N. He, and L. Shao, "Single image dehazing based on the physical model and msrcr algorithm," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2190–2199, 2017.

[17] Engin D. et al., "Cycle-dehaze: Enhanced cyclegan for single image dehazing," in *Proceedings of the IEEE CVPRW*, 2018.

[18] Wu H. et al., "Contrastive learning for compact single image dehazing," in *Proceedings of the IEEE/(CVPR)*, June 2021, pp. 10 551–10 560.

[19] 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.

[20] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE CVPR*, 2022, pp. 12 104–12 113.

[21] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 936–11 945.

[22] Liu Z., et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVPR*, 2021.

[23] 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/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 1833–1844.

[24] P. L. Suárez, D. Carpio, and A. D. Sappa, "Non-homogeneous haze removal through a multiple attention module architecture," in *International Symposium on Visual Computing*. Springer, 2021, pp. 178–190.

[25] 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.

[26] Ancuti C.O., ET AL., "NH HAZE: an image dehazing benchmark with non-homogeneous hazy and haze-free images," in *Proceedings of the IEEE CVPRW*, ser. IEEE, 2020.