Image Dehazing Transformer with Transmission-Aware 3D Position Embedding

Chunle Guo[1] Qixin Yan[2] Saeed Anwar[3] Runmin Cong[4] Wenqi Ren[5] Chongyi Li[6*]

[1] TMCC, CS, Nankai University [2] Tianjin University [3] Australian National University

[4] Beijing Jiaotong University [5] Sun Yat-sen University [6] S-Lab, Nanyang Technological University

guochunle@nankai.edu.cn qxyan@tju.edu.cn saeedanwarcse@gmail.com

rmcong@bjtu.edu.cn rwq.renwenqi@gmail.com lichongyi25@gmail.com

https://li-chongyi.github.io/Proj_DeHamer.html

Abstract

Despite single image dehazing has been made promising progress with Convolutional Neural Networks (CNNs), the inherent equivariance and locality of convolution still bottleneck dehazing performance. Though Transformer has occupied various computer vision tasks, directly leveraging Transformer for image dehazing is challenging: ***1)*** *it tends to result in ambiguous and coarse details that are undesired for image reconstruction;* ***2)*** *previous position embedding of Transformer is provided in logic or spatial position order that neglects the variational haze densities, which results in the sub-optimal dehazing performance.*

The key insight of this study is to investigate how to combine CNN and Transformer for image dehazing. *To solve the feature inconsistency issue between Transformer and CNN, we propose to modulate CNN features via learning modulation matrices (i.e., coefficient matrix and bias matrix) conditioned on Transformer features instead of simple feature addition or concatenation. The feature modulation naturally inherits the global context modeling capability of Transformer and the local representation capability of CNN. We bring a haze density-related prior into Transformer via a novel transmission-aware 3D position embedding module, which not only provides the relative position but also suggests the haze density of different spatial regions. Extensive experiments demonstrate that our method, DeHamer, attains state-of-the-art performance on several image dehazing benchmarks.*

1. Introduction

Single image dehazing aims to restore the haze-free image from the hazy counterpart that suffers from the reduced contrast and dull colors caused by spatial variant haze densities. This task has been a longstanding and challenging problem with a wide range of applications, such as surveillance systems and autonomous driving. To solve this ill-posed problem, prior-based methods like Dark Channel Prior (DCP) [14] and Color Attenuation Prior (CAP) [37] adopt priors as external information to estimate the parameters of the hazy image formation model. The robustness of these methods is limited, especially facing challenging scenes. With the learning capability of CNNs, CNN-based dehazing networks have achieved impressive performance by either estimating the imaging model's parameters [25] or directly learning the haze-free counterpart [20]. However, these networks are still bottlenecked by the local nature of the convolution for modeling the long-range dependencies and the translation equivariance [12]. Global context and spatially variant operations are particularly important for haze removal [8].

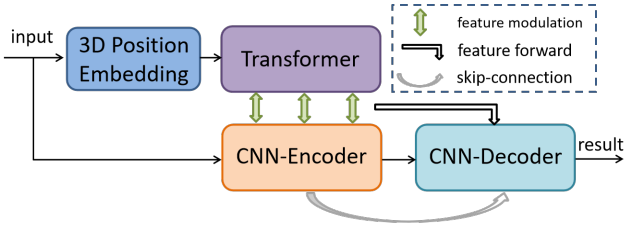


Figure 1. **Thumbnail of main idea**. Transformer is introduced into image dehazing via **1)** transmission-aware 3D position embedding and **2)** feature modulation. The proposed method combines the global modeling capability of Transformer and the local representation capability of CNN.

Although Transformer has swept across many computer vision tasks [6, 21, 31], directly using it in image dehazing exits some inherent issues: **1)** despite Transformer is able to provide long-distance feature dependencies via the cascaded self-attention, even in the early stage, it lacks the capability of retaining local feature details, thus leading to ambiguous and coarse details for image reconstruction; **2)** previous position embedding methods neglect the differences among the regions with variational haze densities, which affects image dehazing performance.

*Chongyi Li (lichongyi25@gmail.com) is the corresponding author.

To overcome these barriers, we propose several novel designs to bring the power of Transformer to image dehazing. The main idea is illustrated in Figure 1. Specifically, we attempt to combine the best world of the global modeling capability of Transformer and the local representation capability of CNN for image dehazing. To achieve that, given a hazy image, we separately extract the hierarchical global features via a Transformer module while the corresponding hierarchical local features are obtained by a CNN module. We propose a transmission (suggesting the haze density by prior information)-aware 3D position embedding module, which provides the relative position information and haze density information for the Transformer, thus improving image dehazing performance. Instead of simply concatenating or adding Transformer features and CNN features, we propose to integrate these features by a feature modulation module that learns the modulation matrices, which solves the feature inconsistency issue. With the modulated features, a CNN decoder module is utilized to enlarge image resolution and render local details of the haze-free image.

The inspired designs in this study can provide guidance for Transformer-based image reconstruction, especially about how to **1)** inherit the advantages of both Transformer features and CNN features via feature modulation and **2)** introduce prior information into Transformer via position embedding. Experiments and comparisons demonstrate the superiority of our method (called DeHamer) over state-of-the-art image dehazing methods.

In a nutshell, our **contributions** are as follows:

- In comparison to pure CNN-based image dehazing networks, our work is the first to introduce the power of Transformer into image dehazing via novel designs.
- We propose a novel transmission-aware 3D position embedding to involve haze density-related prior information into Transformer.
- Extensive experiments on image dehazing benchmark datasets demonstrate the outstanding performance of our method against state-of-the-art methods.

2. Related Work

Image Dehazing. For single image dehazing, existing solutions can be mainly divided into physical model-based methods and deep learning-based methods. Early methods employ haze or image degradation-related priors to estimate the transmission map and global atmospheric light that are key parameters in hazy image formation models such as the atmospheric scattering model [22]. Along this line, DCP [14] assumes that the pixels in non-haze regions have low intensity in at least one color channel. Subsequently, a variety of priors are proposed, such as color-line prior (CLP) [13] and haze-line prior (HLP) [3].

With the success of CNNs, data-driven-based networks have achieved promising results in image dehazing [7, 18, 19, 34]. These methods adopt CNNs to estimate the key parameters of the atmospheric scattering model or directly learn the haze-free image. For instance, Zhang et al. [33] proposed a densely connected pyramid network to estimate the transmission map and the atmospheric light. These estimated parameters are used to obtain a haze-free image. To avoid the accumulated errors in the process of estimating multiple parameters, end-to-end networks have been investigated to directly estimate a haze-free image. For example, Li et al. [17] proposed an all-in-one network for end-to-end image dehazing by reformulating the atmospheric scattering model. Liu et al. [20] proposed a GridDehazeNet, which consists of pre-processing, backbone, and post-processing. In the GridDehazeNet, an attention-based multi-scale estimation on a grid network is used to achieve robust dehazing results. Singh et al. [26] proposed a back-projected pyramid network for image dehazing, which contains iterative U-Net blocks and pyramid convolution blocks. Physics model-based feature learning was proposed for image dehazing [10]. In addition to $\mathcal{L}_1$ and $\mathcal{L}_2$ losses, various losses such as contrastive loss [30] and adversarial loss [8, 11] have been used into image dehazing networks.

Different from previous image dehazing methods, we bring the long-range modeling capability of Transformer to image dehazing and effectively combine such a capability with the local representation capability of CNN via a serial of novel designs. Unlike previous position embedding methods, we involve both the haze density-related prior and spatial position information into Transformer by 3D position embedding. In comparison to the adaptive instance normalization [15, 16, 29] that impose constraints on reference image or semantic information to align the content features, we leverage the feature modulation to inherit the advantages of both CNN and Transformer. These designs produce state-of-the-art dehazing performance and provide insights into Transformer-based image reconstruction.

Visual Transformer. Transformer [28] has been successfully applied in natural language processing tasks. Based on its strong capability of modeling long-range dependencies by stacked self-attention and feed-forward, it has inspired the computer vision community to investigate how to apply Transformer in related tasks such as object detection [5], image segmentation [31], and autonomous driving [23]. For instance, Strudel et al. [31] extended the Vision Transformer (ViT) to semantic segmentation while Xie et al. [12] built the self-supervised learning on Swin Transformer [21]. Chen et al. [6] proposed a Transformer backbone for multi-task image restoration; however, this Transformer relies on large-scale training data for optimal performance. Large-scale paired trained data is scarce for image enhancement and restoration tasks in the real world.

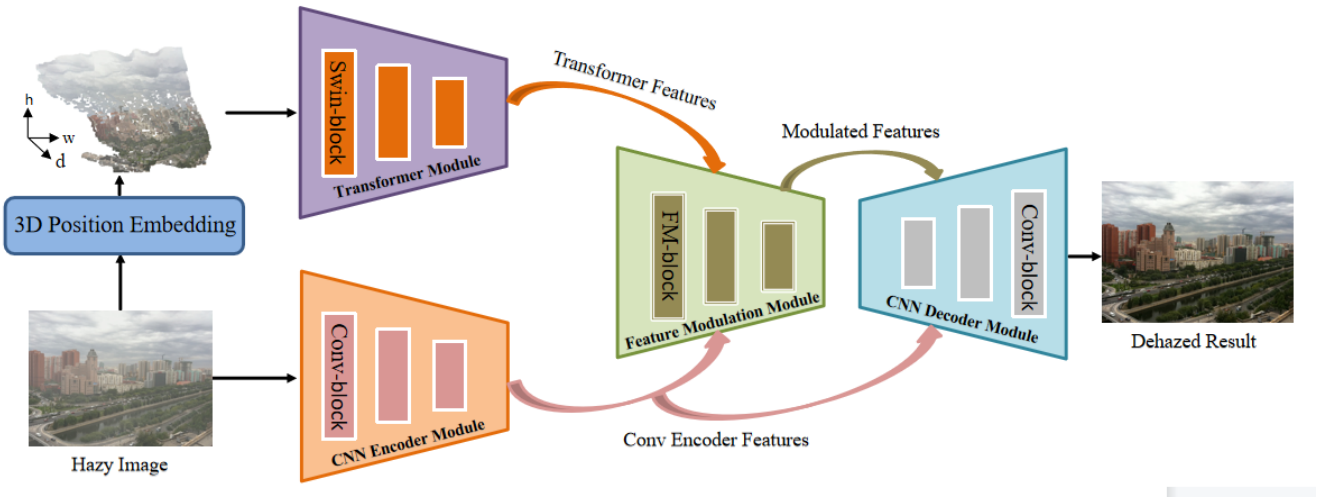


Figure 2. **Overview structure of our method.** Our method consists of five key modules: a transmission-aware 3D position embedding module, a Transformer module, a CNN encoder module, a feature modulation module, and a CNN decoder module.

Although visual Transformers have made great efforts to improve the performance of visual tasks, it is still difficult to directly follow the existing schemes for effectively introducing Transformer into the image dehazing task. This is because Transformer lacks the local representation capability and performs unsuitable position embedding for image dehazing. Therefore, we are motivated to explore exquisite designs to combine the best world of Transformer and CNN in image dehazing.

3. Methodology

The overview structure of our method is presented in Figure 2. Given a hazy image, we first introduce a haze density-related prior into a Transformer module via a transmission-aware 3D position embedding module. Then, our network separately extracts global features and local features via the Transformer module and a CNN encoder module. Afterward, we treat the Transformer features as the condition information and feed them to a feature modulation module to predict modulation matrices (*i.e.*, coefficient matrix and bias matrix) that are employed to scale and shift the corresponding CNN encoder features. In this way, the modulated encoder features enhance the global modeling capability of local features. Following this scheme, the hierarchical Transformer features and CNN encoder features are adaptively integrated. At last, the haze-free image is obtained through a CNN decoder module that gradually enlarges resolutions and renders local details.

In what follows, we will detail these modules. More detailed network structure and parameters can be found in the supplementary material.

3.1. 3D Position Embedding

In vision Transformers, position embedding is crucial to retain spatial position information. However, previous position embedding is provided in logic or spatial position order, which neglects the variational haze densities of different spatial regions in a hazy image. Moreover, variational haze densities challenge existing image dehazing methods.

To solve this issue, we propose a new position embedding method for image dehazing, transmission-aware 3D position embedding, that embeds the haze density-related prior information (e.g., transmission map) into the position encoder. Such a manner suggests the haze densities of different spatial regions. We expect the regions with similar haze density could share similar non-linear mapping relationships in the dehazing process.

To achieve that, we first compute the Dark Channel Prior [14] of the input hazy image I:

$$DCP(I) = \min_{y \in \Omega(x)} (\min_{c \in \{r,g,b\}} I^c(y))) \tag{1}$$

where $\Omega(x)$ is a local patch centered at x. Assuming the value of the atmospheric background light is 1, $DCP(I)$ would be $1 - t$, where t is the transmission map [14]. Note that we choose Dark Channel Prior to generate the haze density information based on its robust performance for image dehazing. Other priors can also be used in our method.

The pipeline of our 3D position embedding module is illustrated in Figure 3. Following the previous work of vision computer [12, 21], we first adopt patch partition and linear embedding to reduce the spatial dimension and increase the channel dimension of the image for efficiently and accurately obtaining long-range dependencies. After patch partition and linear embedding, the dimension of the token vec-

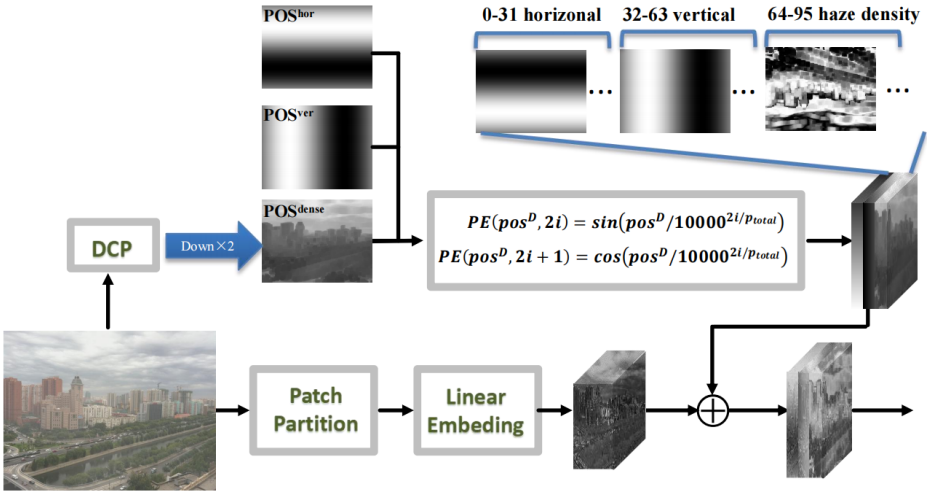


Figure 3. **Illustration of our 3D position embedding.** The dimensions from 0 to 31 represent the horizontal positions, from 32 to 63 represent the vertical positions, and from 64 to 95 represent the haze density positions.

tors becomes $\frac{H}{2} \times \frac{W}{2} \times C$, where H, W, and C represent the height and width of the input image and the number of channels of the token vectors, respectively. Specifically, C is set to 96 in our implementation. Then, we encode the position and haze density information via sinusoidal position encoding as:

$$\begin{aligned} PE(pos^D, 2i) &= sin(pos^D / 10000^{2i/p_{total}}), \\ PE(pos^D, 2i+1) &= cos(pos^D / 10000^{2i/p_{total}}), \end{aligned} \tag{2}$$

where PE represents the 3D encoder information, and pos^D is the position of image patch in dim D (i.e., horizontal dimension, vertical dimension, and haze density dimension). By adding the haze density information $DCP(I)$, we expand the two-dimensional spatial coordinates (x, y) of each pixel to three-dimensional coordinates (x, y, d). The variable i is the position in the token vectors. We set p_{total} to 32 for each dimension, thus forming 96 positions that consist of spatial positions and haze density information. We set the number to 96 for matching the dimension number of the token vectors. Finally, the token vectors and position coding information are combined through an addition operation, as shown in Figure 3.

As the illustration of 3D position embedding presented in Figure 3, for the horizontal position embedding, each column of patches shares the same embedding information, while different embedding values in the horizontal direction represent their relative positional relationships. Similarly, for the vertical position embedding, each row of patches shares the same embedding information, while different embedding values in the vertical direction indicate their relative positional relationships. For the haze density positions, a patch-level embedding, the embedding values represent the haze densities of different spatial regions.

3.2. Network Structure

Transformer Module. To achieve global context to deal with the spatially variant haze, we adopt a Transformer that has a strong capability of modeling the long-range dependencies. Concretely, we adopt Swin Transformer [21] as the backbone to extract the hierarchical Transformer features based on its good trade-off between effectiveness and efficiency. Other Transformer backbones can also be used in our framework. Although a larger image patch could improve the computational efficiency of the Swin Transformer [21], it leads to obvious border artifacts around each patch. Thus, instead of using the default image patch size *i.e.*, 4, we set the patch size to 2. We only adopt the three-stage Swin Transformer, where the lightweight transformer parameters are adopted, i.e., the depth and numbers of attention heads are set to [2,2,2] and [3,6,9], respectively. We did not find obvious gains by using more parameters.

CNN Encoder Module. To obtain local features, we adopt three convolution blocks to extract hierarchical convolution features that correspond to the three-stage Transformer features. In each convolution block, two convolution layers are followed by the ReLU activation function. After the last convolution layer, a max-pooling layer is employed to reduce the image size. The purpose is to ensure the sizes of CNN features are consistent with the corresponding features' sizes of the Swin Transformer. To achieve larger receptive fields, we employ a pyramid pooling module (PPM) [35] at the end of each convolution block, which fuses features under four different scales.

Feature Modulation Module. We found that the features extracted by Transformer have unique characteristics such as long-range attention but coarse textures in comparison to CNN features that have local attention and clear details,

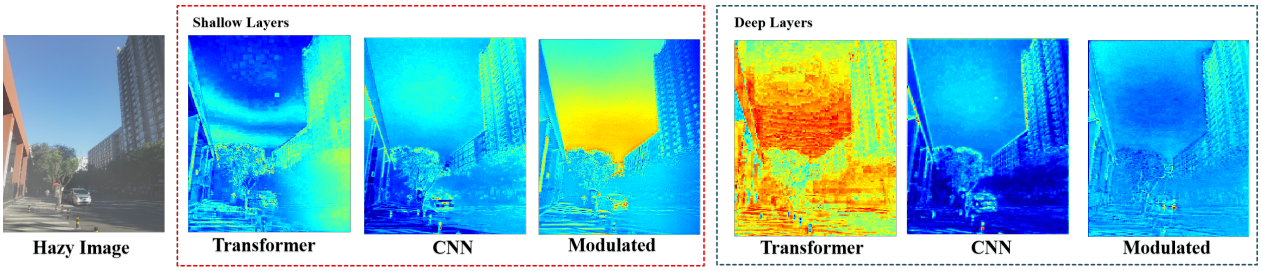


Figure 4. **Visual results of the intermediate features in the CNN encoder module and the Transformer module.** The corresponding modulated features are also presented. The feature maps are illustrated in heatmaps. The features in the Transformer have long-range attention but coarse textures, while the features in the CNN are with clear details. The modulated features inherit the characteristics of both Transformer features and CNN features, i.e., long-range dependencies and clear textures.

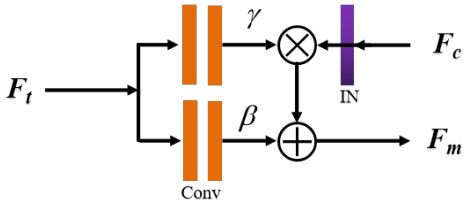


Figure 5. **Illustration of feature modulation block.** F_t: Transformer features; F_c: CNN features; F_m: modulated features.

see Figure 4. We argue the differences are stemmed from the nature of self-attention-based Transformer features and convolution-based CNN features. Thus, directly leveraging the commonly used feature fusion methods such as concatenation and addition may produce suboptimal performance.

Inspired by the style transfer and conditioned image enhancement [16, 29], we treat the Transformer features as the condition information to predict modulation matrices and then modulate the CNN features. In this way, we expect to migrate the long-range attention of Transformer to CNN features without damaging the details of CNN features, which can be expressed as:

$$F_m^s = G_\gamma^s(F_t^s) \otimes IN(F_c^s) \oplus G_\beta^s(F_t^s), \tag{3}$$

where F_m represents the modulated features, $s \in \{1, 2, 3\}$ denotes the stage level. We adopt three stages, in which we half the resolution of features. IN represents the instance normalization operation [27]. $\gamma^s = G_\gamma^s(F_t^s)$ and $\beta^s = G_\beta^s(F_t^s)$. γ and β are the scaling and shifting parameters which both have the same spatial dimensions with the corresponding CNN features F_c. $G_\gamma(\cdot)$ and $G_\beta(\cdot)$ are the modulation matrices estimation blocks that contain two convolution layers conditioned on Transformer features F_t. $\oplus$ and $\otimes$ denote the element-wise addition and element-wise multiplication, respectively. A feature modulation block is shown in Figure 5.

CNN Decoder Module. At last, we use sufficient feature representations to reconstruct the haze-free counterpart with the same size as the input hazy image. More specifically, we first concatenate the modulated features, the corresponding CNN encoder features, and the upsampled decoder features. Here, we discard the corresponding Transformer features due to the coarse texture. Then, these concatenated features are fed to a convolution block consisting of three convolution layers. After that, we adopt the multiscale residual block [32] which includes multiple fully convolutional streams connected in parallel to produce spatially precise features to select the effective features for image dehazing adaptively. After each convolution block, a $2\times$ up-sampling operation is followed to enlarge resolutions. After three convolution blocks, the features are sent to a convolution layer to generate a high-quality haze-free image.

4. Experiments

4.1. Experimental Settings

Implementation Details. Our method is implemented with the PyTorch on an NVIDIA Tesla V100 GPU. We use an ADAM optimizer with default parameters to optimize our method. We set the initial learning rate to 0.0001 and utilized the cosine annealing strategy to adjust the learning rate until convergence. Instead of using complex loss functions, we use only $\mathcal{L}_1$ loss to optimize our network. We randomly crop image patches for training and gradually enlarge the size of the image patch from 128×128 to the full size in the training process.

Training and Testing Datasets. Following previous works [10,20,30], we use **ITS** and **OTS** subsets of RESIDE dataset [18] as the training datasets and conduct the evaluations on **SOTS** subset that contains 500 indoor and 500 outdoor hazy images. In addition, we also include the real-world **Dense-Haze** [1] and **NH-HAZE** [2] datasets in the experiments. Dense-Haze consists of 45 training images, 5 validation images, and 5 test images. The hazy images of Dense-Haze are captured in the dense and homogeneous hazy scenes. NH-Haze also consists of 45 training images, 5 validation images, and 5 test images that are captured in dense and

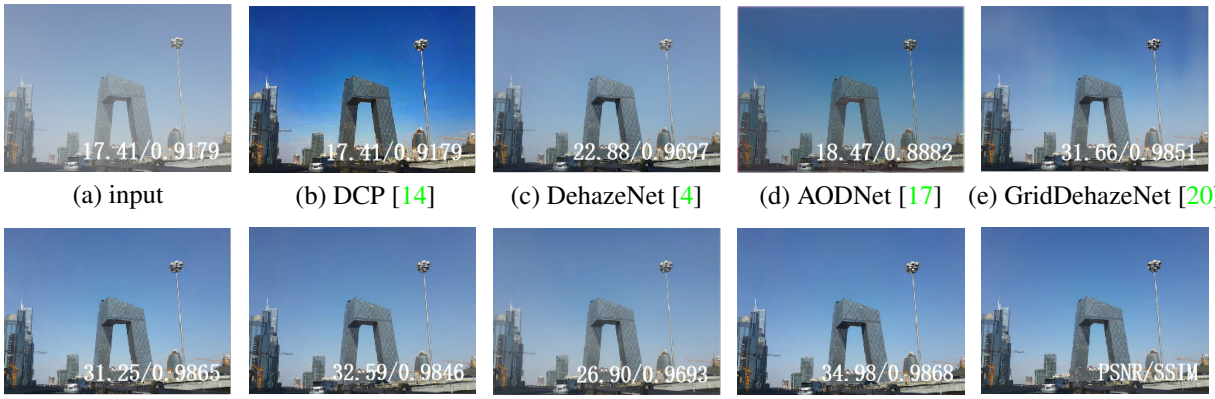


Figure 6. **Visual comparisons on a synthetic hazy image sampled from SOTS-Outdoor testing set.** Zoom in for best view.

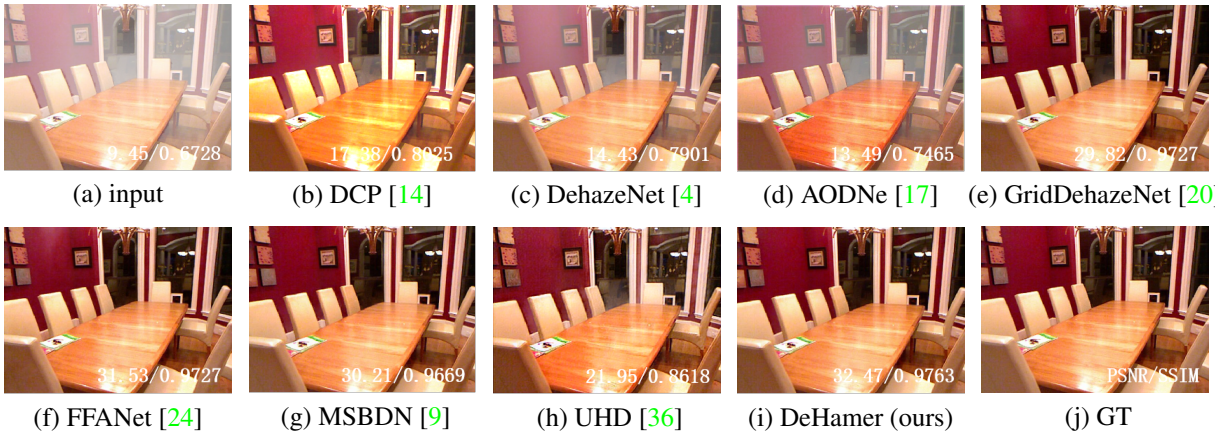


Figure 7. **Visual comparisons on a synthetic hazy image sampled from SOTS-indoor testing set.** Zoom in for best view.

nonhomogeneous hazy scenes.

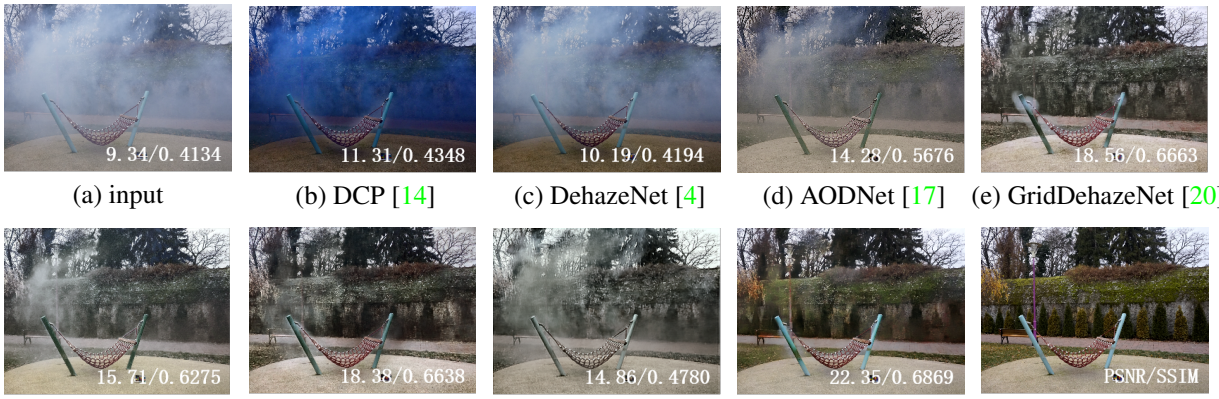
Comparison Methods and Evaluation Metrics. We compare our method with one prior-based method (**DCP** [14]) and six state-of-the-art deep learning-based methods (**DehazeNet** [4], **AODNet** [17], **GridDehazeNet** [20], **FFANet** [24], **MSBDN** [9], **UHD** [36]). We use the released code of these methods for fair comparisons if they are publicly available, otherwise we retrain them using the same training data with our method. We employ commonly-used PSNR (dB) and SSIM to quantify the dehazing performance of different methods.

4.2. Experiments on Synthetic Hazy Images

We first compare different methods on synthetic hazy image datasets SOTS-Indoor and SOTS-Outdoor. The visual comparisons on the hazy images sampled from the SOTS-Outdoor and SOTS-Indoor testing sets are presented in Figure 6 and Figure 7, respectively. As shown, the compared methods either remain haze on the results or produce color deviations while the results of our method are most close to the ground truth images. The better performance of our method is also reflected by the PSNR and SSIM scores on the results.

Table 1. **Quantitative comparisons on synthetic dehazing datasets: SOTS-Indoor and SOTS-Outdoor** The bold numbers denote the best performer under each case.

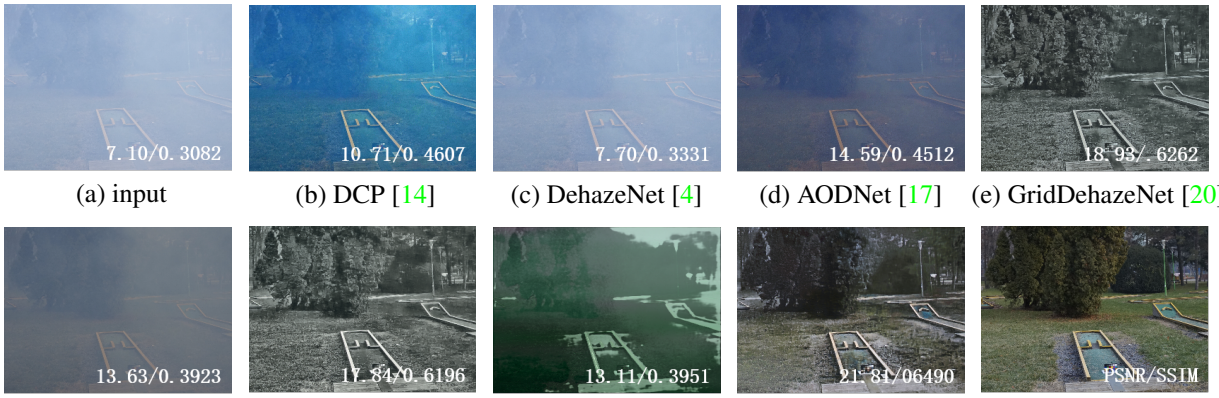
Methods	SOTS-Indoor		SOTS-Outdoor	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DCP	16.61	0.8546	19.14	0.8605
DehazeNet	19.82	0.8209	27.75	0.9269
AODNet	20.51	0.8162	24.14	0.9198
GridDehazeNet	32.16	0.9836	30.86	0.9819
FFANet	36.39	**0.9886**	33.57	0.9840
MSBDN	32.77	0.9812	34.81	0.9857
UHD	21.75	0.8786	26.48	0.9420
DeHamer (ours)	**36.63**	0.9881	**35.18**	**0.9860**



(a) input (b) DCP [14] (c) DehazeNet [4] (d) AODNet [17] (e) GridDehazeNet [20]

(f) FFANet [24] (g) MSBDN [9] (h) UHD [36] (i) DeHamer (ours) (j) GT

Figure 8. **Visual comparisons on a real hazy image sampled from NH-HAZE testing set.** Zoom in for best view.



(a) input (b) DCP [14] (c) DehazeNet [4] (d) AODNet [17] (e) GridDehazeNet [20]

(f) FFANet [24] (g) MSBDN [9] (h) UHD [36] (i) DeHamer (ours) (j) GT

Figure 9. **Visual comparisons on a real hazy image sampled from Dense-Haze testing set.** Zoom in for best view.

Table 2. **Quantitative comparisons on real dehazing datasets: Dense-Haze and NH-HAZE.** The bold numbers denote the best performer under each case.

Methods	Dense-Haze		NH-HAZE	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DCP	11.01	0.4165	12.72	0.4419
DehazeNet	9.48	0.4383	11.76	0.3988
AODNet	12.82	0.4683	15.69	0.5728
GridDehazeNet	14.96	0.5326	18.33	0.6667
FFANet	12.22	0.4440	18.13	0.6473
MSBDN	15.13	0.5551	17.97	0.6591
UHD	12.16	0.4594	16.05	0.4612
DeHamer (ours)	**16.62**	**0.5602**	**20.66**	**0.6844**

Additionally, the quantitative results on all testing sets are compared in Table 1. As presented, our method achieves the highest PSNR and SSIM scores on the SOTS-Outdoor. Moreover, the PSNR score of our method is the highest among the compared methods on the SOTS-Indoor while our SSIM score (0.9881 versus 0.9886) is just 0.0005 lower than the state-of-the-art performer FFANet [24]. The results suggest the good performance of our method, which benefits from the combination of Transformer and CNN with the novel designs.

4.3. Experiments on Real Hazy Images

To further validate the performance of our method, we compare different methods on real hazy images sampled from Dense-Haze and NH-HAZE testing sets. The visual results are presented in Figure 8 and Figure 9, respectively. As presented in Figure 8(a) and Figure 9(a), the real hazy images are extremely challenging, especially the hazy image captured in the nonhomogeneous hazy scene. In comparison to the results of different methods in Figure 8, only our method can remove the haze and recover a similar color with the ground truth image. Besides, our results look more visually pleasing than the compared results. For the results

(a) input (b) *full model* (c) *w/ 2DPE* (d) *w/o Transformer* (e) *w/o ConvE* (f) *w/ add*

Figure 10. **Visual comparisons for ablated models.** Red arrows suggest the obvious differences between the ablated model and the full model. Zoom in for best view.

in Figure 9, only GrideDehazeNet [20], MSBDN [9], and our method can unveil the structure of the input image captured in the dense, hazy scene. In contrast, our method achieves more realistic results and is more similar to the ground truth image than the compared results in terms of color and details.

Quantitative results on the real hazy images are compared in Table 2. For PSNR and SSIM scores, our method achieves the best performance across all testing sets. The PSNR scores of our method on these two testing sets exceed current methods 1-4 dB. The results on such challenging datasets further demonstrate the effectiveness and advantages of our approach.

4.4. Ablation Study

We perform ablation studies to investigate the impacts of our designs on image dehazing performance. The studies include the following ablated models:
w/ 2DPE: 2D position embedding, i.e., removing the haze density position in our Transformer module;
w/o Transformer: removing the Transformer module, i.e., U-Net-like CNN for image dehazing;
w/o ConvE: removing the CNN encoder module, i.e., a Transformer module followed by a CNN decoder;
w/o PPM: removing the pyramid pooing module in the CNN encoder module;
w/ add and *w/ cat*: replacing the feature modulation block with the features addition or features concatenation;
w/o MRB: removing the multiscale residual block in the CNN decoder module.

These models are trained using the same training data as our method (i.e., the *full model*). The quantitative results of ablated models on the NH-HAZE testing set are shown in Table 3. Observing Table 3, we can see all modules could improve the dehazing performance of our method, which suggests the effectiveness of our designs. The result of *w/ 2DPE* demonstrates that the haze density information embedded in the Transformer module is essential for image dehazing, improving the PSNR/SSIM from 18.90/0.6373 to 20.66/0.6844. Besides, removing the Transformer module or CNN encoder module significantly degrades the performance, suggesting that the combination of Transformer and CNN is effective. Compared to the commonly used feature addition and concatenation, modulating the CNN features conditioned on the Transformer features is more suitable for combining Transformer features and CNN features.

Table 3. **Quantitative comparisons of ablated models.**

Modules	Baselines	NH-HAZE	
		PSNR↑	SSIM↑
	full model	**20.66**	**0.6844**
3D Position	*w/ 2DPE*	18.90	0.6373
Transformer	*w/o Transformer*	18.25	0.6058
CNN Encoder	*w/o ConvE*	16.31	0.5731
	w/o PPM	19.04	0.6545
Feature Mod	*w/ add*	17.07	0.5381
	w/ cat	18.69	0.6457
CNN Decoder	*w/o MRB*	18.68	0.6425

Some visual comparisons of ablated models are presented in Figure 10. As shown, *w/ 2DPE* remains haze on the result, as indicated by the red arrow. *w/o Transformer* cannot handle dense haze well while *w/o ConvE* produces coarse details in the result. *w/ add* cannot recover the color of the ColorChecker well, and the remaining haze can be found in its result. In contrast, our *full model* achieves a more visually pleasing result, which removes the dense haze and recovers relatively good details. The visual comparisons demonstrate the effectiveness of our modules again.

5. Conclusion

In this work, we propose a novel method for single image dehazing. The key insights of this work are to effectively integrate Transformer features and CNN features and bring the domain knowledge such as task-specific prior into Transformer for improving the performance. Leveraging feature modulation enables our method to enjoy the best world of Transformer and CNN. Besides, we found that prior information can be effectively introduced to Transformer via 3D position embedding, which further improves the dehazing performance. Extensive comparisons demonstrate that our method achieves state-of-the-art performance in synthetic and real benchmark datasets.

Acknowledgements. This work is funded by National Natural Science Foundation of China under Grants 61922046, S&T innovation project from Chinese Ministry of Education, and China Postdoctoral Science Foundation (NO.2021M701780). We also gratefully acknowledge the support of MindSpore, CANN, and Ascend AI Processor used for this research.

References

[1] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte. Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. In *ICIP*, pages 1014–1018, 2019. 5

[2] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte. Nh-haze: An image dehazing benchmark with nonhomogeneous hazy and haze-free images. In *CVPRW*, pages 444–445, 2020. 5

[3] D. Berman, T. Treibitz, and S. Avidan. Non-local image dehazing. In *CVPR*, pages 1674–1682, 2016. 2

[4] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. An end-to-end system for single image haze removal. *TIP*, 25(11):5187–5198, 2016. 6, 7

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2

[6] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 1, 2

[7] Z. Chen, Y. Wang, Y. Yang, and D. Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *CVPR*, pages 7180–7189, 2021. 2

[8] Q. Deng, Z. Huang, C. C. Tsai, and C. W. Lin. Hardgan: A haze-aware reresentation distillation gan for single image dehazing. In *ECCV*, pages 722–738, 2020. 1, 2

[9] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M. H. Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2157–2167, 2020. 6, 7, 8

[10] J. Dong and J. Pan. Physics-based feature dehazing networks. In *ECCV*, pages 188–204, 2020. 2, 5

[11] Y. Dong, Y. Liu, H. Zhang, S. Chen, and Y. Qiao. Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In *AAAI*, pages 10729–10736, 2020. 2

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16×16 words: Transformer for image recognition at scale. *arXiv:2010.11929*, 2020. 1, 2, 3

[13] R. Fattal. Dehazing using color-lines. *ACM TOG*, 34(1):1–14, 2014. 2

[14] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *TPAMI*, 33(12):2341–2353, 2011. 1, 2, 3, 6, 7

[15] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 2

[16] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy. Tsit: A simple and versatile framework for image-to-image translation. In *ECCV*, pages 206–222, 2020. 2, 5

[17] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. Aod-net: All-in-one dehazing network. In *ICCV*, pages 4770–4778, 2017. 2, 6, 7

[18] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. Benchmarking single-image dehazing and beyond. *TIP*, 28(1):492–505, 2018. 2, 5

[19] C. Li, C. Guo, J. Guo, P. Han, H. Fu, and R. Cong. Pdr-net: Perception-inspired single image dehazing network with refinement. *TMM*, 22(3):704–716, 2019. 2

[20] X. Liu, Y. Ma, Z. Shi, and J. Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *CVPR*, pages 7314–7323, 2019. 1, 2, 5, 6, 7, 8

[21] 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. *arXiv:2103: 14030*, 2021. 1, 2, 3, 4

[22] W. Middleton. Vision through the atmosphere. *Tornoto: University of Tornonto Press*, 1952. 2

[23] A. Prakash, K. Chitta, and A. Geiger. Multi-modal fusion transformer for end-to-end autonomouse driving. In *CVPR*, pages 7077–7087, 2021. 2

[24] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia. FFA-Net: Feature fusion attention network for single image dehazing. In *AAAI*, pages 11908–11915, 2020. 6, 7

[25] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M. H. Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, pages 154–169, 2016. 1

[26] A. Singh, A. Bhave, and D. K. Prasad. Single image dehazing for a variety of haze scenarios using back projected pyramid network. In *ECCV*, pages 166–181, 2020. 2

[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016. 5

[28] A. Vaswani, N. Shazeeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L .Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[29] X. Wang, K. Yu, C. Dong, and C. C. Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018. 2, 5

[30] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L . Ma. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2021. 2, 5

[31] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv:2105: 15203*, 2021. 1, 2

[32] S. Zamir, A. Arora, S. Khan, M. Hayat, F. Khan, M. H. Yang, and L. Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, pages 492–511, 2020. 5

[33] H. Zhang and V. Patel. Densely connected pyramid dehazing network. In *CVPR*, pages 3194–3203, 2018. 2

[34] X. Zhang, H. Dong, J. Pan, C. Zhu, Y. Tai, C. Wang, J. Li, F. Huang, and F. Wang. Learning to restore hazy video: A new real-world dataset and a new method. In *CVPR*, pages 9239–9248, 2021. 2

[35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *ICCV*, pages 2881–2890, 2017. 4

[36] Z. Zheng, W. Ren, X. Cao, X. Hu, T. Wang, F. Song, and X. Jia. Ultra-high-definition image dehazing via multi-guided bilateral learning. In *CVPR*, pages 16185–16194, 2021. 6, 7

[37] Q. Zhu, J. Mai, and L. Shao. A fast single image haze removal algorithm using color attenuation prior. *TIP*, 24(11):3522–3533, 2015. 1