ISSN: 2710-0588

Research on Feature Fusion Image Defogging Algorithms based on Transformers

Jingjing Liu

School of Computer science and Artificial Intelligence, Lanzhou University of Technology, Lanzhou 730050, People's Republic of China

Abstract

Under hazy conditions, airborne particulate matter and chemical substances reduce image contrast and blur details by absorbing and scattering light, severely compromising the accuracy of computer vision tasks such as autonomous driving and remote sensing monitoring. Addressing the limitations of traditional physical defogging models prone to estimation errors, the insufficient global feature modeling capabilities of mainstream convolutional neural networks (CNNs), and the high computational complexity and inadequate local detail capture of existing Transformer-based methods, this paper proposes a Transformer-based feature fusion image defogging algorithm. This algorithm employs a U-shaped encoder-decoder as its backbone network. It introduces a Hub-and-Spoke Multi-Head Attention (HSMHA) mechanism to replace traditional self-attention, significantly reducing computational overhead while preserving global context modeling capabilities. A Feature Refinement Block (FRB) is embedded within the feedforward neural network to enhance the recovery of image texture and detail information. A Multiscale Residual Enhancer (MRE) is constructed to effectively eliminate redundant high-frequency features and deepen learning of subtle feature variations. A contrastive regularization (CR) learning strategy is introduced, using blurred images as negative samples and clear images as positive samples to guide the model toward learning more discriminative feature representations, thereby enhancing the consistency between defogged images and their original clear counterparts. Experimental results on the SOTS-indoor and SOTS-outdoor synthetic datasets demonstrate that the proposed algorithm achieves optimal or suboptimal levels in both Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM). Specifically, on the SOTS-indoor dataset, PSNR reaches 35.79dB and SSIM attains 0.984, and on the SOTS-outdoor dataset, PSNR was 34.31 dB and SSIM is 0.987. Qualitative results demonstrate the algorithm's superior performance in restoring image color fidelity, contrast, and detail integrity, effectively addressing issues such as incomplete defogging, color bias, or edge blurring present in existing methods.

Keywords

Computer Vision, Image Defogging, Transformer, Feature Fusion.

1. Introduction

Computer vision technology analyzes and interprets image and video information by simulating the human visual system, finding widespread application in critical fields such as autonomous driving [1], intelligent surveillance, remote sensing image processing, and medical imaging analysis. However, under adverse weather conditions such as smog, sandstorms, and fog, atmospheric particulate matter and chemical pollutants cause absorption, refraction [2], and scattering [3] of light propagation. This leads to degraded images characterized by reduced contrast, color distortion, and blurred details. In autonomous driving scenarios, for instance, haze-induced image degradation can prevent vehicle perception systems from accurately

ISSN: 2710-0588

identifying traffic lights, pedestrians, and obstacles, directly threatening driving safety. In remote sensing monitoring, degraded foggy-day images compromise the precision of crop growth assessments and disaster area estimations. Therefore, developing efficient and robust image defogging algorithms to restore degraded images' true information holds significant theoretical value and practical importance for enhancing the reliability of computer vision systems in complex environments.

Research on image defogging technology spans decades and can be broadly categorized into two main types based on technical paths: traditional physical modeling methods and deep learning-based methods. Traditional methods center on physical models, reconstructing fog-free images by establishing atmospheric scattering models [4]. He et al [5] proposed the Dark Channel Prior (DCP) algorithm, a landmark achievement in this field. It leverages the prior knowledge that at least one color channel pixel value approaches zero in local regions of fog-free images to estimate atmospheric light and transmittance. However, this method is prone to estimation errors in sky and highly illuminated areas, leading to color casts in defogged images. Addressing limitations of DCP, Wu et al. [6] introduced an adaptive atmospheric light estimation module and transmittance optimization strategy to construct an enhanced defogging framework. This method segments foggy input images into foreground and background regions, applying segmented mapping based on distinct grayscale values to effectively mitigate color distortion. Li et al. [7] integrated depth priors into an iterative scheme to adaptively optimize transmission maps. They designed residual convolutions to extract depth priors from a small training dataset, guiding the iterative process to derive adaptive transmission optimization formulas. Shi et al. [8] proposed an improved multi-scale dark channel prior algorithm, reconstructing defogged images through multi-scale feature map reconstruction. Wang et al. [9] introduced a dynamic scattering coefficient as an exponential function of image depth, replacing the constant scattering coefficient.

With the advancement of deep learning technology, CNN-based defogging algorithms have emerged as the mainstream with its strong feature learning capabilities. Wang et al. [10] proposed a parallel heterogeneous twin end-to-end model-free defogging algorithm. By leveraging dual-path feature collaboration between the backbone network and a lightweight detail enhancement module, and employing multi-scale feature adaptive fusion, it demonstrates superior restoration accuracy and detail preservation in complex, non-uniform haze scenarios. Ren et al. [11] divided the network into coarse-scale and fine-scale components to optimize the recovery of transmission maps and detail regions, respectively. However, the local receptive field nature of CNNs makes it challenging to capture long-range pixel correlations in images, leading to insufficient modeling of global scene information. This often results in incomplete defogging and blurred edges when handling large-scale dense fog regions.

In recent years, Transformer models have been progressively applied to image defogging tasks due to their powerful global feature modeling capabilities enabled by self-attention mechanisms. Guo et al. [12] combined CNNs with Transformers for image defogging by integrating positional embedding modules with physical prior information. Yuan et al. [13] proposed a hierarchical Tokens-to-Token transformation that fuses local and global features by merging adjacent tokens to structure image information. Wang et al. [14] introduced a Transformer-based defogging backbone network, which significantly outperformed traditional CNNs across various experimental metrics. This network achieved superior results in computer vision and demonstrated broad applicability across diverse downstream tasks.

To address the limitations of existing algorithms, such as inadequate balance between global modeling and local details and high computational complexity, this paper proposes a Transformer-based feature fusion image defogging algorithm. The main research contributions include:

ISSN: 2710-0588

(1) Proposed a Transformer-based feature fusion image defogging algorithm, enhancing the traditional Transformer architecture.

(2) Designed HSMHA to replace conventional attention mechanisms, significantly improving computational efficiency while simultaneously refining the feedforward neural network structure to strengthen capture and representation of fog-specific degraded features.

(3) Designed MRE to deepen learning of subtle feature variations and introduced CR learning strategies to simultaneously enhance feature learning quality and model robustness.

2. Model Design

2.1. Overall Model Architecture

The proposed algorithm adopts a three-tiered collaborative framework centered on efficient feature extraction, precise feature fusion, and high-quality feature optimization. This framework comprises a Transformer-based U-shaped encoder-decoder, a feature fusion block (FFB), and a CR module, as illustrated in Figure 1. The Transformer-based U-shaped encoder-decoder serves as the backbone network. On the one hand, it retains the U-shaped structure's advantage in multi-scale feature extraction, progressively capturing global degraded features of foggy images, such as overall fog concentration distribution and large-scale contrast decay patterns, through the down-sampling process. On the other hand, it leverages the up-sampling process to restore fine-grained features layer by layer. Simultaneously, it incorporates an enhanced Transformer Block (with HSMHA and FRB) to address the shortcomings of traditional CNNs in modeling global features. This effectively correlates the dependency relationships between distant pixels in the image, laying a global semantic foundation for subsequent defog feature processing. FFB plays a pivotal role in cross-scale feature integration. Through a gated selection mechanism, it dynamically adjusts the fusion weights of feature maps at different resolutions, prioritizing the retention of detail features beneficial for defogging while suppressing redundant fog interference features. This achieves precise complementarity between global degradation features and local detail features, avoiding the issues of incomplete defogging or detail loss caused by single-scale features.

At the network's terminal, a specially designed MRE module refines the feature maps output by the encoder-decoder through a multi-branch residual structure. On one hand, it employs convolutional kernels of varying sizes to capture multi-scale high-frequency information, precisely identifying and eliminating redundant high-frequency noise caused by fog interference. On the other hand, residual connections fuse raw features with processed features, enhancing the network's learning capacity for subtle feature variations. This ensures that defogged images retain overall clarity without sacrificing critical local details.

Finally, to further enhance feature learning quality and the realism of defogging results, the CR strategy is introduced. This strategy uses paired data of foggy and clear images as training basis, defining the input blurry foggy image as a negative sample and its corresponding true clear image as a positive sample. Simultaneously, the model's output defogged image serves as an anchor sample. During training, CR optimizes the contrastive loss function to force anchor samples to minimize their distance from positive samples while maximizing their distance from negative samples within the feature representation space. This constraint mechanism not only guides the model to learn more discriminative defogging features—effectively distinguishing fog-induced artifacts from true image characteristics, but also prevents issues like color bias, overexposure, or detail distortion in model outputs. Ultimately, it ensures defogged results exhibit high consistency with true fog-free images in color coherence and structural similarity, significantly enhancing the algorithm's defogging accuracy and robustness.

ISSN: 2710-0588

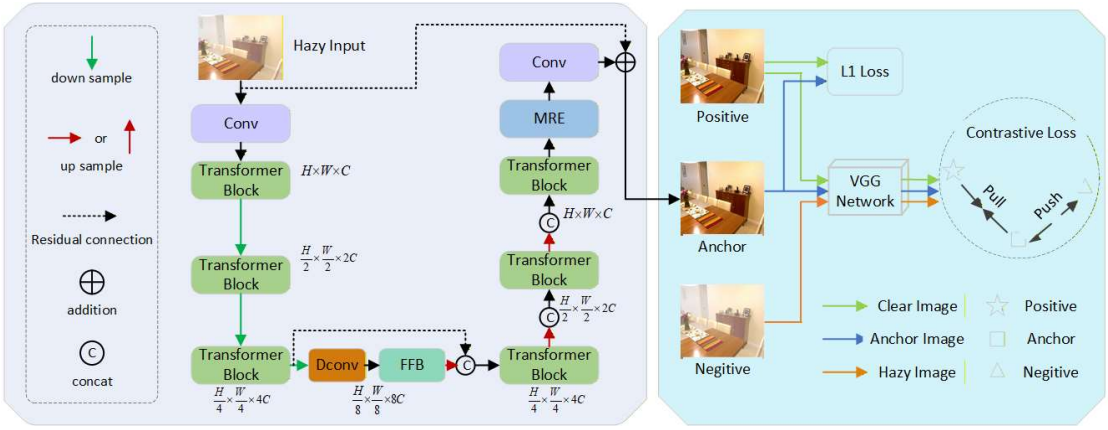

Figure 1. Overall structure of the model

2.2. Transformer

The improved Transformer architecture, as shown in Figure 2(b), the HSMHA is designed to enhance the traditional self-attention mechanism, effectively reducing computational overhead. Simultaneously, the FRB is incorporated into the feedforward neural network to improve the recovery performance of image texture and detail information.

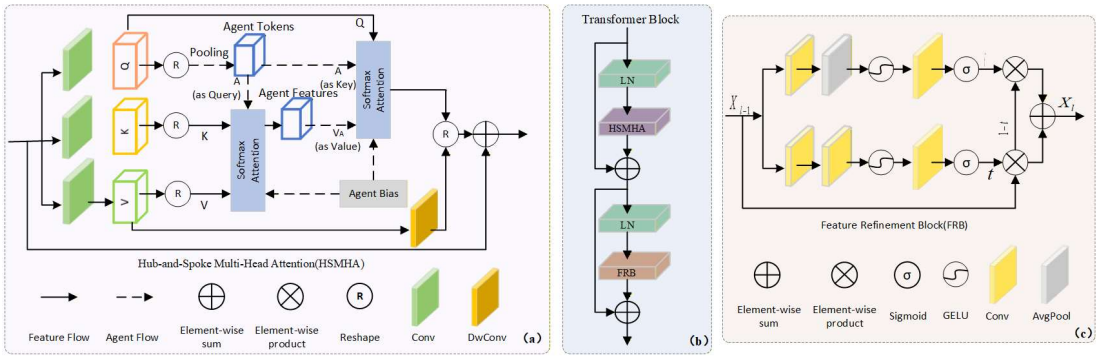

Figure 2. Transformer structure

2.2.1. HSMHA

The structure of HSMHA is shown in Figure 2(a). Its core innovation lies in introducing an additional token A into the attention triplet (Q, K, V), thereby constructing a quadruple attention paradigm (Q, A, K, V). The first Softmax attention is applied to the triplet (A, K, V), where A is derived from the Query matrix via simple pooling operations. This serves to aggregate information from V, and an attention matrix is computed between A and K to yield the feature VA. The second Softmax attention is executed on the triplet (Q, A, VA), forming the final attention output.

First, we enhance the traditional self-attention mechanism by extracting an additional proxy token A from the input features through a simple pooling strategy. Acting as a proxy for Q, A aggregates information from both K and V to obtain a proxy feature V_A that synthesizes information from all values, as illustrated in Equation (1):

$$V_A = Soft\max(AK^T) \cdot V \tag{1}$$

Then, using A as the Key and V_A as the Value, attention calculations are performed with the Query to broadcast the global information of the proxy features to each Query token, yielding the final output O_A, as shown in Equation (4.2). By employing proxy tokens as an intermediary,

ISSN: 2710-0588

this method preserves the global context modeling capability of Softmax attention while significantly reducing computational complexity by decreasing the number of Query tokens.

$$O_A = Soft\max(QA^T) \cdot Soft\max(AK^T) \cdot V \tag{2}$$

To better utilize location information, an Agent Bias O_A is designed, as shown in Equation (3):

$$O_A = Soft\max(QA^T + B_2) \cdot Soft\max(AK^T + B_1) \cdot V \tag{3}$$

Among these, $B_1 \in \mathbb{R}^{n \times N}$, $B_2 \in \mathbb{R}^{N \times n}$. To enhance parameter efficiency, each proxy bias is constructed using three bias components, namely $B_1 = (B_{1c} + B_{1r} + B_{1b})$, where $B_{1c} \in \mathbb{R}^{n \times 1 \times w}$ represents column bias, $B_{1r} \in \mathbb{R}^{n \times h \times 1}$ represents row bias, and $B_{1b} \in \mathbb{R}^{n \times h_0 \times w_0}$ represents block bias.

2.2.2. FRB

To minimize the loss of image detail during feature extraction, the FRB is designed with the structure shown in Figure 2(c). This module effectively mitigates the loss of structural details during sampling, adaptively aggregates sampled features, and thereby better restores fine image structures and texture details.

Specifically, the input features are first subjected to global average pooling, followed by further processing through convolutional layers and activation functions to obtain feature $\hat{F}$. Subsequently, the input feature X undergoes convolution and activation function operations to yield another feature representation F . Features $\hat{F}$ and F are then fused to produce the final output $\hat{X}$. This process is illustrated by Equations (4), (5), and (6):

$$\hat{F} = \sigma(Conv(\phi(Conv(GAP(X))))) \tag{4}$$

$$F = \sigma(Conv(\phi(Conv(Conv(X))))) \tag{5}$$

$$\hat{X} = X \odot F + (1 - F) \odot \hat{F} \tag{6}$$

Where, σ represents the activation function, ϕ represents the GELU activation function, and $\odot$ represents the Hadamard product.

3. Experimental Design and Results Analysis

On the synthetic foggy image datasets SOTS-indoor and SOTS-outdoor, the proposed method is compared with existing image defogging techniques in terms of PSNR and SSIM. Quantitative results are shown in Table 1 (where data represent the average of all test results), while qualitative results are illustrated in Figure 3. The optimal result is indicated in bold, and the second-best result is underlined.

As shown in Table 1, the proposed method achieves optimal or second-best results on the core evaluation metrics, PSNR reflecting image fidelity and SSIM reflecting image structural consistency, across the synthetic foggy image datasets SOTS-indoor and SOTS-outdoor. This fully validates the effectiveness of the method for fog removal in diverse foggy scenarios. Specifically, on the SOTS-indoor dataset, our method achieves a PSNR of 35.79 dB. This represents a significant improvement over traditional CNN-based methods such as

ISSN: 2710-0588

GridDehazeNet (32.03 dB) and MSBDN (33.64 dB). Even when compared to other Transformer-based methods like DehazeFormer-T (34.98 dB) and PCSformer (35.23 dB), our method still outperforms by 0.81–3.76 dB, indicating superiority in restoring clarity and suppressing noise interference in indoor foggy images. The SSIM value of 0.984, while slightly lower than AECR-Net's 0.986, significantly outperforms mainstream methods like FFA-Net (0.981) and IGTB-Dehazing (0.981), demonstrating the proposed method's superior performance in preserving the structural integrity of indoor scene images.

On the SOTS-outdoor dataset, the performance advantage of the proposed method is even more pronounced: achieving a PSNR of 34.31 dB, it not only surpasses traditional CNN-based methods such as MSBDN (33.46 dB) and FFA-Net (33.38 dB), but also outperforms most Transformer-based approaches like DehazeFormer-T (33.97 dB) and IGTB-Dehazing (33.95 dB). It demonstrates superior capability in restoring image dynamic range, particularly when handling complex fog concentration distributions in outdoor scenes. The SSIM value of 0.987 ranks first among all comparison methods, significantly outperforming GridDehazeNet (0.963), CFEN-ViT (0.969), and even outperforms AECR-Net, which excels in indoor scenes. This demonstrates that our method achieves superior restoration of object texture details in outdoor scenes, effectively addressing issues like edge blurring and insufficient contrast that commonly arise in traditional methods during large-scale foggy outdoor conditions.

It is worth noting that while AECR-Net achieves the highest SSIM value (0.986) on the SOTS-indoor dataset, it does not provide experimental results for outdoor scenes. Furthermore, its indoor PSNR (36.52 dB) shows only a small gap compared to our method (35.79 dB). In contrast, our method demonstrates superior overall performance in outdoor scenes, exhibiting stronger scene adaptability. Combining the above quantitative results, it is evident that our method, through the synergistic effects of HSMHA, FRB, MRE, and CR, not only addresses the limitations of traditional CNNs in global modeling but also overcomes the computational complexity and poor detail recovery issues of some Transformer-based approaches. It achieves high-fidelity, structurally consistent defogging effects across various foggy scenes, fully demonstrating the effectiveness and superiority of the proposed algorithm.

Table 1. Quantitative results of different methods on the SOTS dataset

Methods	SOTS-indoor		SOTS-outdoor	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
GridDehazeNet	32.03	0.971	30.71	0.963
MSBDN	33.64	0.984	33.46	<u>0.982</u>
FFA-Net	36.38	0.981	33.38	0.980
AECR-Net	<u>36.52</u>	**0.986**	--	--
MSTN	34.63	0.978	32.52	0.981
CFEN-ViT	32.15	0.966	31.04	0.969
DehazeFormer-T	34.98	0.976	33.97	0.963
PCSformer	35.23	0.981	33.03	0.978
IGTB-Dehazing	35.36	0.981	33.95	0.976
Ours	**35.79**	<u>0.984</u>	**34.31**	**0.987**

Figure 3 presents the qualitative results of different methods on the synthetic foggy dataset SOTS. The top three rows show qualitative results on SOTS-indoor, while the bottom three rows display results on SOTS-outdoor. As seen in the figure, images de-hazed by GridDehazeNet and MSBDN exhibit overall darker colors; images processed by FFA-Net and AECR-Net show color distortion; compared to the ground truth (GT) images, the proposed method achieves the

ISSN: 2710-0588

closest representation to standard clear-weather images in terms of color, contrast, and detail preservation.

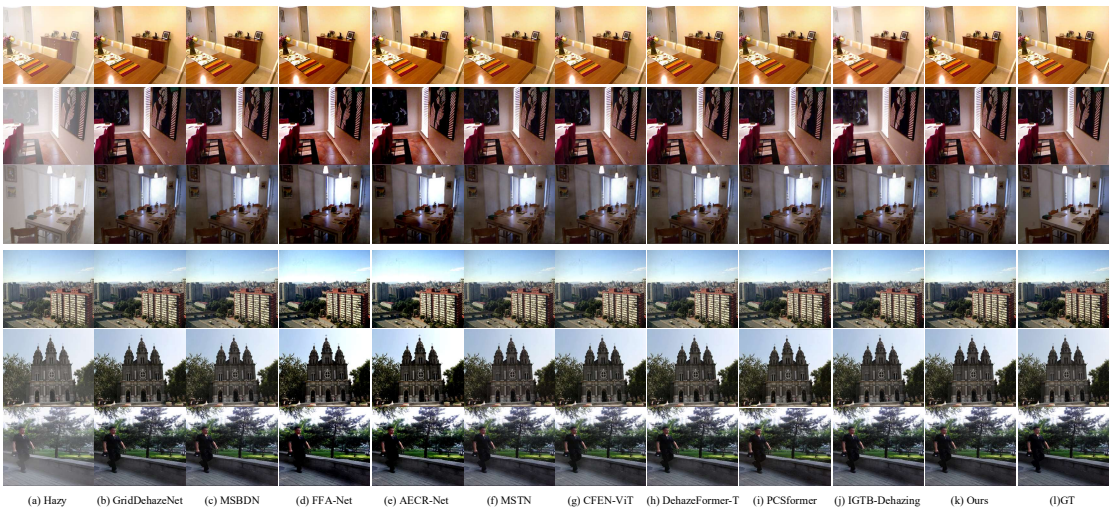

Figure 3. Qualitative results of different methods on the SOTS dataset

4. Conclusion

This paper addresses two core challenges in image defogging: balancing global feature modeling with local detail recovery, and optimizing computational efficiency. We propose a defogging algorithm that integrates an improved Transformer with multi-module collaborative optimization. By designing the HSMHA architecture, we successfully overcome the high computational complexity bottleneck of traditional Transformers, significantly enhancing computational efficiency while preserving the model's ability to model global feature correlations. The design of FRB enables precise capture of fine image structures and texture information, mitigating detail loss during feature extraction. The synergistic effect of MRE and CR learning strategies further enhances the model's ability to learn subtle features, ensuring de-fogged images exhibit high fidelity in color and contrast compared to original fog-free images. Experimental results validate the effectiveness of each innovative module and the overall algorithm's superiority. The proposed algorithm outperforms most existing mainstream methods on quantitative metrics (PSNR, SSIM) and qualitative effects across SOTS datasets, effectively addressing contrast reduction, detail blurring, and color distortion caused by fog-induced image degradation. Future research may extend to real foggy-day datasets to optimize the model's adaptability to practical scenarios such as non-uniform haze and complex lighting conditions. Concurrently, exploring lightweight deployment solutions for the model could provide more efficient and robust image defogging technology support for practical applications like autonomous driving and remote sensing monitoring.

References

[1] Wang Y, Bi X. The Application of Computer Vision Target Recognition Technology in Autonomous Driving [C]. 2024 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS).IEEE,2024:519-524.

[2] Baker D J, Halvorson H. of Modern Physics[J]. Studies in History and Philosophy of Modern Physics, 2013, 44: 464-469.

[3] Burchard W. Light scattering techniques[M]//Physical techniques for the study of food biopolymers. Boston, MA: Springer US, 1994: 151-213.

ISSN: 2710-0588

[4] Cox L. Optics of the atmosphere-scattering by molecules and particles [J]. Optica Acta: International Journal of Optics, 1977, 24(7): 779-779.

[5] He K, Sun J, Tang X. Single image haze removal using dark channel prior[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(12): 2341-2353.

[6] Wu Zifan, Luo Weiping Improved algorithm for image dehazing based on dark channels [J]. Journal of Wuhan Textile University, 2023, 36 (05): 47-52

[7] Li S, Liu R, Fan X, et al. Single Image Dehazing via Adaptive Transmission Optimization with Deep Prior[C]//2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM). IEEE, 2018: 1-5.

[8] Shi S, Zhang Y, Zhou X, et al. Cloud removal for single visible image based on modified dark channel prior with multiple scale[C]//2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, 2021: 4127-4130.

[9] Wang Q, Zhao L, Tang G, et al. Single-image dehazing using color attenuation prior based on haze-lines[C]//2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019: 5080-5087.

[10] Wang K, Yang Y, Li B, et al. Uneven Image Dehazing by Heterogeneous Twin Network[J]. IEEE access, 2020, 8: 118485-118496.

[11] Ren W, Liu S, Zhang H, et al. Single image dehazing via multi-scale convolutional neural networks[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 154-169.

[12] Guo C L, Yan Q, Anwar S, et al. Image dehazing transformer with transmission-aware 3d position embedding[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5812-5820.

[13] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 558-567.

[14] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 568-578.