MB-TaylorFormer: Multi-branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing

Yuwei Qiu[1] Kaihao Zhang[2] Chenxi Wang[1] Wenhan Luo[1] Hongdong Li[2] Zhi Jin[1,3*]

[1] Sun Yat-sen University [2] Australian National University
[3] Guangdong Provincial Key Laboratory of Robotics and Digital Intelligent Manufacturing Technology

Abstract

In recent years, Transformer networks are beginning to replace pure convolutional neural networks (CNNs) in the field of computer vision due to their global receptive field and adaptability to input. However, the quadratic computational complexity of softmax-attention limits the wide application in image dehazing task, especially for high-resolution images. To address this issue, we propose a new Transformer variant, which applies the Taylor expansion to approximate the softmax-attention and achieves linear computational complexity. A multi-scale attention refinement module is proposed as a complement to correct the error of the Taylor expansion. Furthermore, we introduce a multi-branch architecture with multi-scale patch embedding to the proposed Transformer, which embeds features by overlapping deformable convolution of different scales. The design of multi-scale patch embedding is based on three key ideas: 1) various sizes of the receptive field; 2) multi-level semantic information; 3) flexible shapes of the receptive field. Our model, named Multi-branch Transformer expanded by Taylor formula (MB-TaylorFormer), can embed coarse to fine features more flexibly at the patch embedding stage and capture long-distance pixel interactions with limited computational cost. Experimental results on several dehazing benchmarks show that MB-TaylorFormer achieves state-of-the-art (SOTA) performance with a light computational burden. The source code and pre-trained models are available at https://github.com/FVL2020/ICCV-2023-MB-TaylorFormer.

1. Introduction

Single image dehazing is an image restoration task which aims to estimate latent haze-free images from hazy images. Starting with early CNN-based approaches [49, 5]

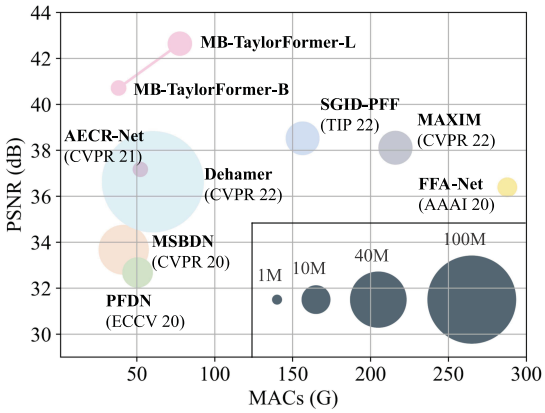

Figure 1: **Improvement of MB-TaylorFormer over the SOTA approaches.** The circle size is proportional to the number of model parameters. All models are trained on SOTS-Indoor [32].

and their revolutionary performance in dehazing, haze removal gradually shifts from prior based strategies [89, 20] to deep learning-based methods. In the past decade, deep dehazing networks achieve significant performance improvement due to advanced architectures like multi-scale information fusion [49, 47, 37], sophisticated variants of convolution [74, 36], and attention mechanisms [45, 86, 21].

Recently, Transformer has been popularly employed in various computer vision tasks and subsequently contributes greatly to the progress of high-level vision tasks [39, 71, 61, 6]. In the field of image dehazing, there are several challenges with the direct application of Transformer: 1) the computational complexity of Transformer is quadratic with the resolution of feature map, which makes it poorly suited to the pixel-to-pixel task of dehazing. Although some works apply self-attention in small spatial windows [35, 73] to relieve this problem, the receptive field of Transformer is restricted; 2) the basic elements of visual Transformer usually have more flexible scales [39]. However, existing

*Corresponding author: jinzh26@mail2.sysu.edu.cn

visual Transformer networks [78, 73] generally generate fixed-scale tokens by fixed convolution kernels. Thus, there is still room for improvement via introducing flexible patch embedding Transformer to the dehazing task.

To address the first challenge, we propose a Transformer variant expanded by Taylor formula (TaylorFormer), which applies self-attention on the entire feature map across spatial dimension and maintains linear computational complexity. Specifically, we calculate the weights of self-attention by performing a Taylor expansion on softmax, and then reduce the computational complexity of self-attention from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ by applying the associative law of matrix multiplication. This strategy brings three advantages: 1) it retains the ability of Transformer to model long-distance dependencies among data, and avoids reducing the receptive field caused by window splitting [58]; 2) it provides a stronger value approximation than methods using a kernel-based formulation of self-attention [28], and is similar to the vanilla Transformer [64]; 3) it makes the Transformer concerned with pixel-level interactions rather than channel-level interactions [78], which allows for more fine-grained processing of features.

Considering the error caused by ignoring Peano's form of remainder [43] in the Taylor formula, we introduce a multi-scale attention refinement (MSAR) module to refine TaylorFormer. We exploit the local correlation within image by convolving the local information of queries and keys to output a feature map with scaling factors. The number of feature map channels is equal to the number of heads in multi-head self-attention (MSA), so each head has a corresponding scaling factor. Our experiments show that the proposed MSAR module effectively improves model performance with tiny computational burden (see Table 3).

To tackle the second challenge, inspired by the success of inception modules [60, 59] and deformable convolutions [13] in CNN-based dehazing networks [18, 65, 44, 74], we propose a multi-branch encoder-decoder backbone for TaylorFormer, termed ad MB-TaylorForemr, based on multi-scale patch embedding. The multi-scale patch embedding has various sizes of receptive field, multi-level semantic information, and flexible shape of receptive field. Considering that the generation of each token should follow the local relevance prior, we truncate the offsets of the deformable convolutions. We reduce computational complexity and the number of parameters by the depthwise separable method. The tokens from different scales are then fed independently into TaylorFormer and finally fused. The multi-scale patch embedding module is capable of generating tokens with different scales and dimensions, and the multi-branch structure is capable of processing them simultaneously to capture more powerful features.

To summarize, our main contributions are as follows: (1) We propose a new variant of linearized Transformer based on Taylor expansion to model long-distance interactions between pixels without window splitting. An MSAR module is introduced to further correct errors in the self-attention of TaylorFormer; (2) We design a multi-branch architecture with multi-scale patch embedding. Among it, multiple field sizes, flexible shape of receptive field, and multi-level semantic information can help simultaneously generate tokens with multi-scales and capture more powerful features; (3) Experimental results on public synthetic and real dehazing datasets show that the proposed MB-TaylorForme achieves state-of-the-art (SOTA) performance with few parameters and MACs.

2. Related Works

CNN-based Image Dehazing. With the development of CNN, significant progress has been achieved in image restoration tasks [70, 85, 81, 84, 82, 67, 66], including image dehazing [5, 45, 19, 4, 88, 69]. One way is based on the atmospheric scattering model, such as DehazeNet [5], DCPDN [80], and Aod-Net [31]. However, the performance is hardly satisfactory when the atmospheric scattering model fails. The other way is based on the idea of image conversion, which is not dependent on the atmospheric scattering model. These kinds of methods predict haze-free images in an end-to-end manner, such as EPDN [47], FFA-Net [45] and AECRNet [74]. Nevertheless, it is difficult for CNN-based models to learn long-range pixel dependencies.
Efficient Self-attention. The computational complexity of the Transformer grows quadratically with the increasing spatial resolution of feature map, which makes it very demanding on computational resources. Some works reduce the computational burden by sliding window [48, 24] or shifted window [39, 35, 73, 58] based self attention. However, this design limits the ability of Transformer to model long-distance dependencies in the data. MaxViT [63] compensates for the decrease in receptive field with Grid attention. However, Grid attention is not strictly linear in computational complexity and is still quadratic on high-resolution images. Another approach is to modify the softmax-attention of the vanilla Transformer. Restormer [78] applies self-attention between channels, and ignores global relationships between pixels. Performer [12] achieves linear complexity by random projection. However, the queries, keys, and values require a large size, which results in increasing computation cost. Poly-nl [3] bridges the connection between attention and high-order polynomials. However, this has not been explored in a self-attention structure. [28, 52, 46, 75, 57] decompose softmax by kernel functions and use the associative law of matrix multiplication to achieve linear complexity. These methods are functional approximations, e.g., each element in the attention map is positive [46]. However, value approximations are not considered. Our method does not use the kernel func-

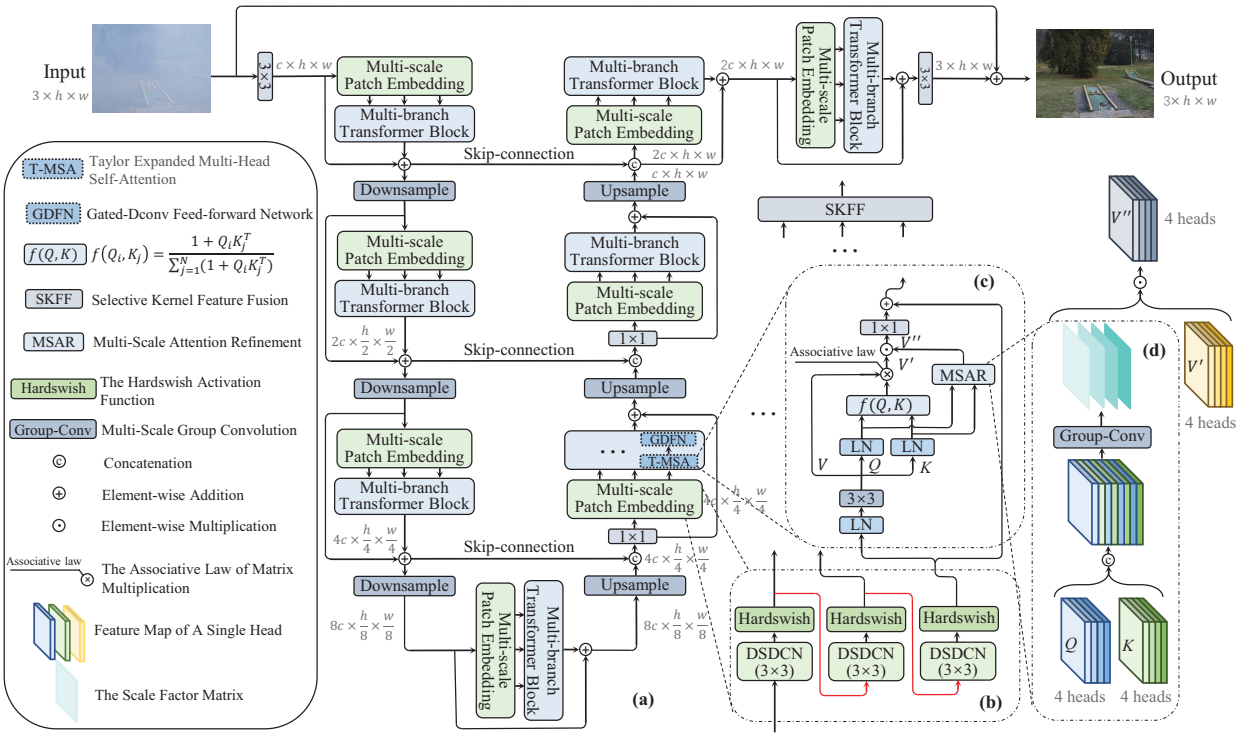

Figure 2: **Architecture of MB-TaylorFormer.** (a) MB-TaylorFormer consists of a multi-branch hierarchical design based on multi-scale patch embedding. (b) Multi-scale patch embedding embeds coarse-to-fine patches. (c) TaylorFormer with linear computational complexity. (d) MSAR module compensates for errors in Taylor expansion.

tion but performs the Taylor expansion on softmax directly, which guarantees a functional and numerical approximation to softmax.

Multi-scale Transformer Networks. In the field of high level vision, [41] is a simple pyramid structure. IFormer [55] applies inception structures to mix high and low frequency information. However, it does not utilize different patch sizes. CrossViT [7] and MPViT [30] process multi-scale patches via multiple branches to obtain multi-scale receptive fields. However, the receptive field shape is not flexible. In the field of low level vision, MSP-Former [76] uses multi-scale projections to help Transformer to represent complex degraded environment. Giqe [54] processes feature maps of different sizes via multi-branch. [87] represents different features related to the task utilizing multiple sub-network. The recent Transformer networks for recovery tasks [78, 73, 58, 26, 41] build a simple U-net network or with single-scale patches. However, these works hardly further explore multi-scale patches and multi-branch architectures. Although [29] uses deformable convolution in self-attention, the number of sampling points of the convolution kernel is fixed. While our multi-scale deformable convolution not only has flexible sampling points, but also provides multi-level semantic information.

3. MB-TaylorFormer

We aim to build an efficient and lightweight Transformer-based dehazing network. To reduce the computational complexity, we apply Taylor expansion of softmax-attention to satisfy the associative law and adopt a U-net structure similar to the Restormer [78]. To compensate for the effects of Taylor expansion errors, we propose an MSAR module. In the following parts, we first describe the overall architecture of MB-TaylorFormer (Fig. 2a). Then we introduce three core modules: multi-scale patch embedding (Fig. 2b), Taylor expanded self-attention (Fig. 2c) and MSAR module (Fig. 2d).

3.1. Multi-branch Backbone

Given a hazy image $I \in \mathbb{R}^{3\times h\times w}$, we apply convolution for shallow feature extraction to generate $F_0 \in \mathbb{R}^{c\times h\times w}$. Subsequently, we employ a four-stage encoder-decoder network for deep feature extraction. Each stage has a residual block containing a multi-scale patch embedding and a multi-branch Transformer block. We use multi-scale

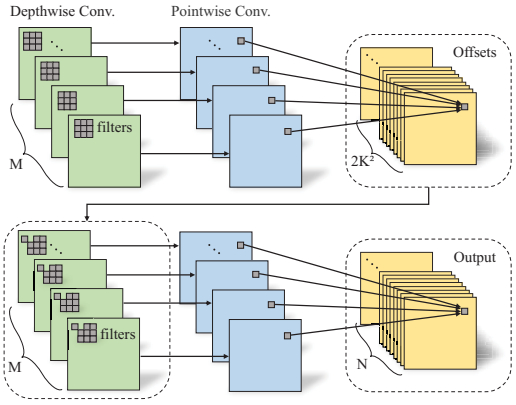

Figure 3: **An illustration of DSDCN.** The offsets are generated by $K \times K$ depthwise convolutions and pointwise convolutions, and the output is generated by $K \times K$ depthwise deformable convolutions and pointwise convolutions.

patch embedding to generate multi-scale tokens, and then feed them into multiple Transformer branches, respectively. Each Transformer branch contains multiple Transformer encoders. We apply the SKFF [79] module at the end of the multi-branch Transformer block to fuse features generated by different branches. Benefiting from the excellent performance of the Taylorformer and multi-branch design, we have the opportunity to compress the number of channels and Transformer blocks. We apply pixel-unshuffle and pixel-shuffle operations [53] at each stage to down-sample and up-sample features, respectively. Skip-connection [51] is used to aggregate information from encoder and decoder, and a 1×1 convolutional layer is employed for dimensionality reduction (except for the first stage). We also use a residual block after the encoder-decoder structure to restore fine structural and textural details. Finally, a 3×3 convolutional layer is applied to reduce channels and output a residual image $R \in \mathbb{R}^{3\times h\times w}$. We thus obtain the restored image by $I' = I + R$. To further compress the computation and parameters, we apply depthwise separable convolutions [23, 11] in the model.

3.2. Multi-scale Patch Embedding

Visual elements can vary greatly in scale. Existing works [73, 35, 78] employ convolution with fixed kernels into patch embedding, which may result in single scale of visual tokens. To address this issue, we design a new multi-scale patch embedding with three properties: 1) multiple sizes of receptive field, 2) multi-level semantic information, 3) flexible shapes of receptive field. Specifically, by designing multiple deformable convolutions (DCN) [13] with different scales of convolution kernels in parallel, we enable patch embedding to generate both coarse and fine visual tokens, as well as flexible transformation modeling

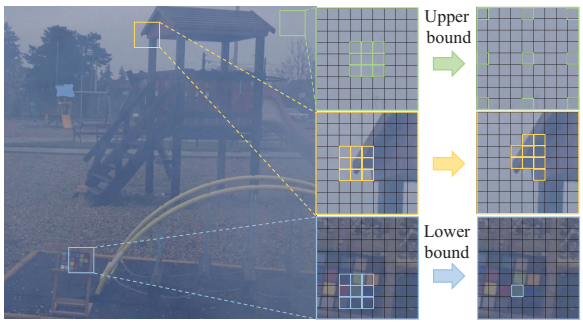

Figure 4: **An illustration of the receptive field of DSDCN (the offsets are truncated to [-3,3]).** The upper bound of the receptive field of the DSDCN is 9×9 and the lower bound is 1×1.

capabilities. Inspired by the operation of stacking conventional layers that can expand receptive fields [56], we stack several deformable convolutional layers with small kernels instead of a single deformable convolutional layer with large kernels. This not only increases the depth of network and consequently provides multi-level semantic information, but also helps to reduce parameters and computational burden. All deformable convolutional layers are followed by Hardswish [22] activation functions.

Similar to the strategy of depthwise separable convolutions [23, 11], we propose depthwise separable and deformable convolutions (DSDCN), which decomposes the parts of DCN with depthwise convolution and pointwise convolution, as shown in Fig. 3. The computational cost of standard DCN and DSDCN regarding a $h \times w$ image is as follows:

$$\Omega(\text{DCN}) = 2MK^4hw + MNK^2hw + 4MK^2hw, \quad (1)$$

$$\Omega(\text{DSDCN}) = 8MK^2hw + MNhw. \quad (2)$$

where M and N are the numbers of channels in the input and output, respectively, and K is the kernel size of convolution. The number of parameters of DCN and DSDCN are as follows:

$$\text{P}(\text{DCN}) = 2MK^4 + MNK^2, \quad (3)$$

$$\text{P}(\text{DSDCN}) = 4MK^2 + MN, \quad (4)$$

In summary, DSDCN greatly reduces computational complexity and number of parameters compared to DCN.

Considering that images are locally relevant and patch embedding captures the basic elements of feature maps, the visual elements (*i.e.*, tokens) should be more focused on local areas. We control the receptive field range of the patch embedding layer by truncating the offsets, which we choose in practice to be $[-3, 3]$. As it is shown in Fig. 4, depending on the shape of visual object, the model is able to select the

receptive field size by learning, which has an upper bound of 9×9, equivalent to a dilated convolution [77] of BF = 4, and a lower bound of 1×1. When we set up multi-scale patch embedding in parallel, the sizes of the receptive field of different branches are $x \in [1, 9]$, $y \in [x, x+8]$ and $z \in [y, y+8]$ in ascending order (for three branches). Experiments in the supplementary material show that setting constraints on the receptive field of each token in a reasonable way can improve the model's performance.

3.3. Taylor Expanded Multi-head Self-Attention

Let queries (Q), keys (K), and values (V) be a sequence of $h \times w$ feature vectors with dimensions D, where h and w is the height and width of the image, respectively. The formula of the vanilla Transformer [64] is as follows:

$$V' = Softmax\left(\frac{Q^T K}{\sqrt{D}}\right) V^T. \tag{5}$$

Since $Q \in \mathbb{R}^{hw \times D}$, $K \in \mathbb{R}^{hw \times D}$ and $V \in \mathbb{R}^{hw \times D}$, Softmax causes the computational complexity of self-attention to be $\mathcal{O}\left(h^2 w^2\right)$, resulting in expensive computational cost.

We aim to reduce the computational complexity of self-attention from $\mathcal{O}\left(h^2 w^2\right)$ to $\mathcal{O}\left(hw\right)$. To achieve it, we first write the generalized attention equation for Eq. (5), as follows:

$$V_i' = \frac{\sum_{j=1}^{N} f\left(Q_i, K_j\right) V_j}{\sum_{j=1}^{N} f\left(Q_i, K_j\right)}, \tag{6}$$

where the matrix with i as subscript is the vector of the i-th row of matrix, and $f(\cdot)$ denotes any similarity function. Eq. (6) degenerates to Eq. (5) when we let $f\left(Q_i, K_j\right) = \exp\left(\frac{Q_i^T K_j}{\sqrt{D}}\right)$. If we apply the Taylor formula to perform a first-order Taylor expansion on $\exp\left(\frac{Q_i^T K_j}{\sqrt{D}}\right)$ at 0, we can rewrite Eq. (6) as:

$$V_i' = \frac{\sum_{j=1}^{N} \left(1 + Q_i^T K_j + o\left(Q_i^T K_j\right)\right) V_j^T}{\sum_{j=1}^{N} \left(1 + Q_i^T K_j + o\left(Q_i^T K_j\right)\right)}. \tag{7}$$

Further, we generate $\tilde{Q}_i$, $\tilde{K}_i$ from the normalization of vectors Q_i and K_j to approximate $\exp\left(\frac{Q_i^T K_j}{\sqrt{D}}\right)$. When the norm of $\tilde{Q}_i$ and $\tilde{K}_i$ are smaller than 1, we can make the values of the attention map all positive, and in practice, we find the best results are achieved by normalizing the norm to 0.5. As shown in Fig. 5, we consider that there is an approximation to e^x and its first-order Taylor expansion is in the definition domain of $[-0.25, 0.25]$. So we eliminate Peano's form of remainder [43] and obtain the expression for the Taylor expansion of self-attention as follows:

$$V_i' = \frac{\sum_{j=1}^{N} \left(1 + \tilde{Q}_i^T \tilde{K}_j\right) V_j^T}{\sum_{j=1}^{N} \left(1 + \tilde{Q}_i^T \tilde{K}_j\right)}. \tag{8}$$

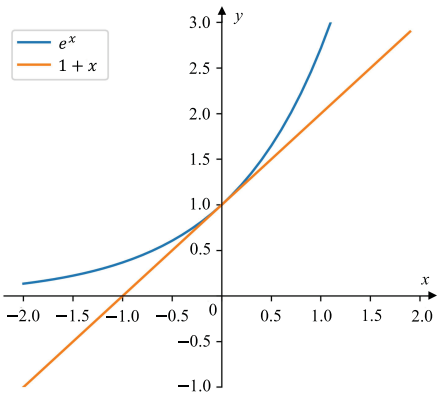

Figure 5: **e^x (blue) and its first-order Taylor expansion curve (orange).** The closer to 0, the tighter the approximation of the orange line to the blue line.

Finally, we apply the matrix multiplication associative law to Eq. (8), as follows:

$$\begin{aligned} V_i' &= \text{Taylor-Attention}\left(Q_i, K_i, V_i\right) \\ &= \frac{\sum_{j=1}^{N} V_j^T + \tilde{Q}_i{}^T \sum_{j=1}^{N} \tilde{K}_j V_j^T}{N + \tilde{Q}_i^T \sum_{j=1}^{N} \tilde{K}_j}. \end{aligned} \tag{9}$$

Supplementary materials provide the pseudo-code of Taylor expanded multi-head self-attention (T-MSA).

Similar to MDTA [78], we employ deep convolutional generation for Q, K, V to emphasize the local context. We also employ a multi-head structure, where the number of heads increases progressively from top to bottom of the level. The computational complexity of the standard multi-head self-attention (MSA) module and the T-MSA module regarding an image of $h \times w$ patches is as follows:

$$\Omega(\text{MSA}) = 4hwD^2 + 2h^2 w^2 D, \tag{10}$$

$$\Omega(\text{T-MSA}) = 18hwD + 7hwD^2. \tag{11}$$

Generally, $h \times w$ is much larger than D, so T-MSA provides more possibilities than MSA for testing high-resolution images, and ensures the values are close to the MSA.

3.4. Multi-scale Attention Refinement

Since we perform a first-order Taylor expansion of softmax in T-MSA and ignore Peano's form of reminder, there is an inevitable approximation error. For the n-th order remainder term $\frac{(Q_i K_j)^n}{n!} (n \geq 2)$ of Taylor expansion, the matrix multiplicative combination law cannot be used to make the computational complexity of T-MSA linear. However, the remainder term is related to the Q and K matrices. Considering that the images have local correlation, we learn the local information of the Q and K matrices to correct the inaccurate output V'. In addition, the conv-attention module

Table 1: **Quantitative comparisons of various methods on dehazing benchmarks.**"-" indicates that the result is not available. The best and second best results are highlighted in bold and underlined, respectively.

Methods	SOTS-Indoor		SOTS-Outdoor		O-HAZE		Dense-Haze		Overhead	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	#Param	MACs
DCP [20]	16.62	0.818	19.13	0.815	16.78	0.653	12.72	0.442	-	0.6G
DehazeNet [5]	19.82	0.821	24.75	0.927	17.57	0.770	13.84	0.430	0.01M	0.6G
GFN [50]	22.30	0.880	21.55	0.844	18.16	0.671	-	-	0.50M	14.9G
GDN [37]	32.16	0.984	30.86	0.982	18.92	0.672	14.96	0.536	0.96M	21.5G
PFDN [15]	32.68	0.976	-	-	-	-	-	-	11.27M	51.5G
MSBDN [14]	33.67	0.985	33.48	0.982	24.36	0.749	15.13	0.555	31.35M	41.54G
FFA-Net [45]	36.39	0.989	33.57	0.984	22.12	0.770	15.70	0.549	4.46M	287.8G
AECR-Net [74]	37.17	0.990	-	-	-	-	15.80	0.466	2.61M	52.2G
MAXIM-2S [62]	38.11	0.991	34.19	0.985	-	-	-	-	14.10M	216.0G
SGID-PFF [4]	38.52	0.991	30.20	0.975	20.96	0.741	12.49	0.517	13.87M	156.4G
Restormer [78]	38.88	0.991	-	-	23.58	0.768	15.78	0.548	26.10M	141.0G
Dehamer [19]	36.63	0.988	35.18	0.986	25.11	0.777	16.62	0.560	132.50M	60.3G
Ours (-B)	40.71	0.992	37.42	0.989	25.05	**0.788**	**16.66**	0.560	2.68M	38.5G
Ours (-L)	**42.64**	**0.994**	**38.09**	**0.991**	**25.31**	0.782	16.44	**0.566**	7.43M	88.1G

allows TaylorFormer to better handle high frequency information [42].

Specifically, for multi-heads $Q_m \in \mathbb{R}^{head\times\frac{D}{head}\times N}$ and $K_m \in \mathbb{R}^{head\times\frac{D}{head}\times N}$, we reshape them into $\hat{Q}_m \in \mathbb{R}^{head\times\frac{D}{head}\times H\times W}$ and $\hat{K}_m \in \mathbb{R}^{head\times\frac{D}{head}\times H\times W}$, where $head$ denotes the number of heads, and we concatenate $\hat{Q}_m$ and $\hat{K}_m$ along the channel dimension to generate a tensor $T \in \mathbb{R}^{head\times\frac{2D}{head}\times H\times W}$, which is subsequently passed through a multi-scale grouped convolutional layer to generate a gating tensor $G \in \mathbb{R}^{head\times 1\times H\times W}$ as:

$$G = \text{Sigmoid}\left(\text{Concat}\left(T_1 W_1^G, \ldots, T_{head} W_{head}^G\right)\right), \quad (12)$$

where $T_{head} \in \mathbb{R}^{\frac{D}{head}\times H\times W}$ is the $head$-th head of T, and $W_{(\cdot)}^G$ is the convolution with different kernels. Since different levels of the network have different numbers of heads, we choose the corresponding multi-scale grouped convolutions for different numbers of heads. Supplementary materials provide details of the structure of the MSAR module.

With the approach of T-MSA and the module of MSAR, the refined T-MSA module is computed as:

$$\begin{aligned} \hat{X} &= X + \text{Cat}\left(H_1 \odot G_1, \ldots, H_{head} \odot G_{head}\right) W^p, \\ H_i &= \text{Taylor-Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \end{aligned} \quad (13)$$

where X and $\hat{X}$ denote the input and output feature maps. The projections are parameter matrices $W^P \in \mathbb{R}^{D\times D}$, $W_i^Q \in \mathbb{R}^{D\times\frac{D}{head}}$, $W_i^K \in \mathbb{R}^{D\times\frac{D}{head}}$, and $W_i^V \in \mathbb{R}^{D\times\frac{D}{head}}$.

4. Experiments

In this section, we conduct experiments to demonstrate the effectiveness of the proposed MB-TaylorFormer. Further details including more qualitative results are provided in Supplementary Materials.

4.1. Experiment Setup

Implementation Details. We provide two architectures of MB-TaylorFormer including MB-TaylorFormer-B (the basic model) and MB-TaylorFormer-L (a larger variant). We employ random cropping and random flipping for data augmentation. We set the initial learning rate to 2e-4 and gradually reduce it to 1e-6 using cosine annealing [40]. we only use L1 loss as our loss function.

Datasets. We evaluate the proposed MB-TaylorFormer on synthetic dataset (RESIDE [32]) and real-world datasets (O-HAZE [2], Dense-Haze [1]). As subsets of RESIDE, ITS and OTS contain 13990 pairs of indoor and 313950 pairs of outdoor images, respectively. Models are evaluated on the SOTS subset. The real-world datasets, O-HAZE and Dense-Haze, contain 45 and 55 paired images, respectively. We use the last 5 images of each dataset as the testing set and the rest as the training set.

4.2. Experiments on Synthetic Hazy Images

Table 1 compares the performance of MB-TaylorFormer with the SOTA methods on synthetic datasets. Our basic model (MB-TaylorFormer-B) achieves 40.71dB PSNR and 0.994 SSIM on SOTS-Indoor. It improves the PSNR by 2.19dB over the previous SOTA method SGID-PFF [4], but with merely 10% number of parameters and 14% computational cost of SGID-PFF. Furthermore, our large model (MB-TaylorFormer-L) achieves 4.12dB gain over the SGID-PFF with approximately half of its complexity. Our method also surpasses the second best performing method Dehamer [19] by 2.91dB on the SOTS-Outdoor. We also compare the visual results of MB-TaylorFormer-L with other SOTA dehazing methods. Fig. 6 shows that the images generated by other methods are less natural in the shadows and high frequency regions. However, our

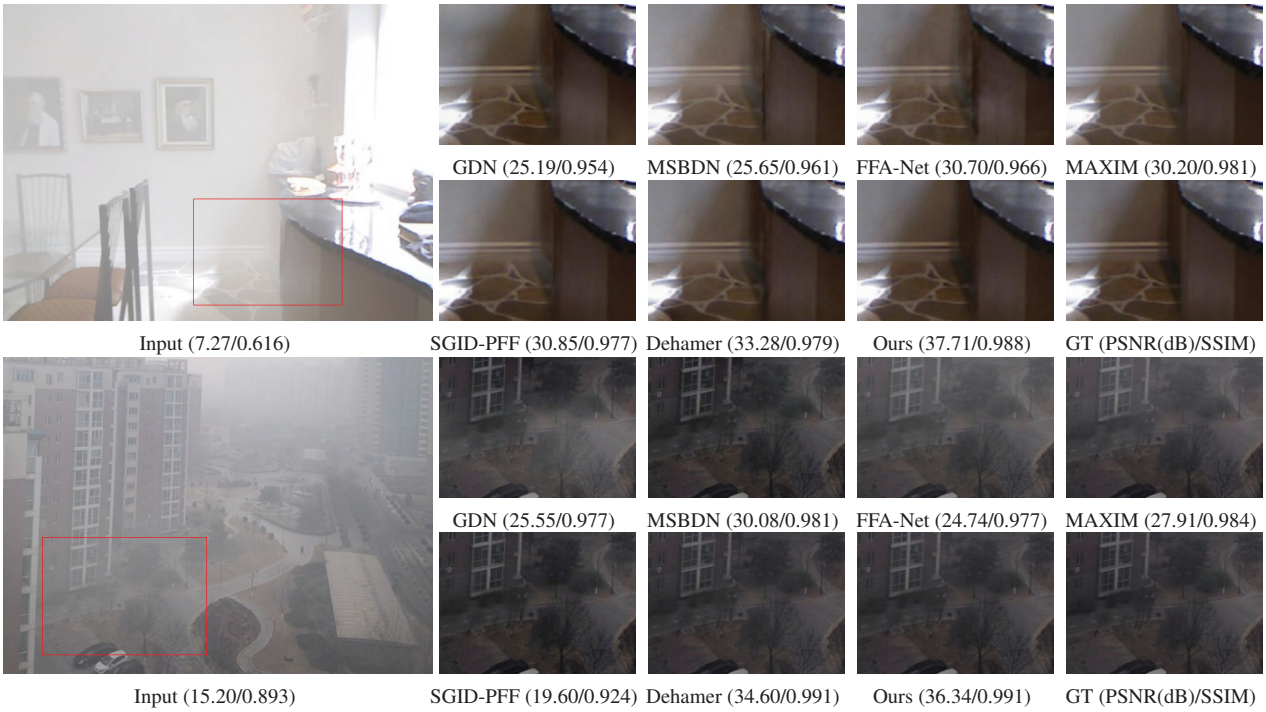

Figure 6: **Visual comparisons on synthetic hazy images.** The first two rows of images are from the SOTS-Indoor dataset, the last two rows are from SOTS-Outdoor. Our MB-TaylorFormer-L generates haze-free images with color fidelity and finer textures.

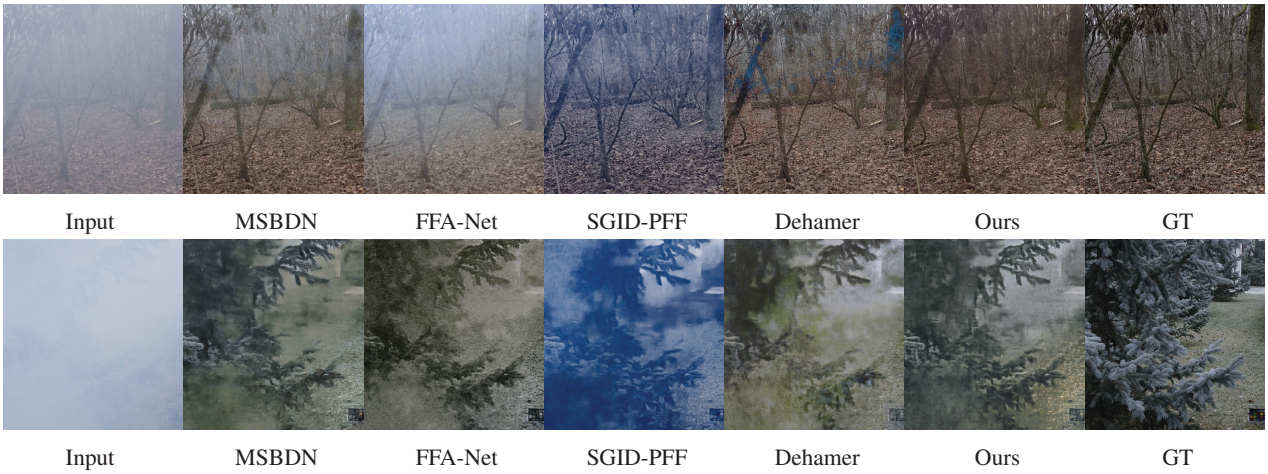

Figure 7: **Visual comparisons on real hazy images.** The first row of images is from the O-HAZE, The second row of images is from the Dense-Haze. Our MB-TaylorFormer generates haze-free images with color fidelity and without artifacts.

MB-TaylorFormer generates haze-free images with better-restored details in the shadows, and is most similar to ground truth, especially in the high-frequency region.

4.3. Experiments on Real Hazy Images

We further evaluate our MB-TaylorFormer against the SOTA methods on the real datasets O-HAZE [2] and Dense-Haze [1]. The quantitative comparison is shown in Table 1. We have the following findings: 1) MB-TaylorFormer outperforms previous methods by up to 0.20dB and 0.04dB PSNR magnitude on the O-HAZE and Dense-Haze, respectively. 2) Usually, when training on small-size datasets, the non-convex losses of visual Transformer are expected to lead to poor performance [42]. However, MB-TaylorFormer still achieves the best PSNR and SSIM scores than other CNN-based models on small-size datasets. We also com-

Table 2: **Ablation studies for the multi-scale patch embedding and multi-branch structure.** "-S" means two convolutional layers in series with the same kernel size of 3, and "-P" means two convolutional layers in parallel with the same kernel size of 3. Supplementary materials provide more details.

Branch	Type of Conv.	PSNR	SSIM	#Params	MACs
Single	Conv	38.27	0.991	2.655M	33.63G
Double	Conv-P	38.42	0.991	2.652M	37.89G
	Dilated Conv-P	38.77	0.991	2.652M	37.89G
	Conv-S	39.04	0.992	2.652M	37.89G
	DSDCN-S	**40.71**	**0.992**	2.677M	38.51G

Table 3: **Effectiveness of multi-scale attention refinement module.** It balances computational burden and performance.

Methods	PSNR	SSIM	#Params	MACs
MB-TaylorFormer-B	40.71	0.992	2.677M	38.51G
W/o MSAR	38.74	0.990	2.582M	37.60G
$G : \mathbb{R}^{h\times 1\times H\times W} \rightarrow \mathbb{R}^{h\times C\times H\times W}$	**40.88**	**0.993**	4.288M	68.43G

pare our MB-TaylorFormer with previous methods in terms of the visual quality of dehazed images, which are presented in Fig. 7. The dehazed images by MSBDN [14] hardly remove the haze, FFA-Net [45] generates images with graininess and loss of detail, SGID-PFF [4] generates images with color distortion, and the result images by Dehamer [19] suffer from artifacts and texture loss. The haze-free images generated by our method are much cleaner.

4.4. Ablation Studies

In this section, all MB-TaylorFormer models are trained on 256×256 pixel patches cropped from SOTS-Indoor. The epoch is set as 500. Based on MB-TaylorFormer-B, we analyze the effectiveness of different modules on our framework. MB-TaylorFormer-L is used to explore the effects of branch dimension, channel dimension, and depth.

Study of multi-scale patch embedding and multi-branch structure. In Table 2, we study the difference in the patch embedding and the different number of branches. Specifically, we set a single-branch model based on single-scale standard convolution as the baseline and modify it from the following aspects. 1) To examine the effect of multi-branch structure, we design the patch embedding on models with single-scale and multi-branching (Conv-P). 2) To study the effect of multiple receptive field sizes, we use parallel dilated convolutional layers (DF=1, 2) to embed patches (Dilated Conv-P). 3) To investigate the influence of multi-level semantic information, we further replace dilated convolution with standard convolution to embed patches, and employ the approach of connecting two convolutional layers in series (Conv-S). 4) To examine the effect of flexible receptive field shapes, we additionally replace standard convolution with DSDCN (DSDCN-S), as shown in Fig. 2(b). The experiment shows that the performance from best to worst is DSDCN-S, Conv-S, Dilated Conv-S, Conv-P and Conv. This indicates that our multi-scale patch embedding can embed patches flexibly. As shown in Fig. 8, we visualize the average of the feature maps in the first stage of our network, and find that the multi-scale tokens (DSDCN-S) help the network to obtain richer information than the single-scale tokens (Conv). Thanks to the depth-separable design, our DSDCN has only a tiny increase in the number of parameters and computational cost compared to the depth-separable convolution.

Table 4: **Comparison with other linear self-attention modules.** We replace the T-MSA of our proposed model with another linear self-attention module.

Methods	PSNR	SSIM	#Params	MACs
MB-TaylorFormer-B	**40.71**	**0.992**	2.677M	38.51G
W/o MSAR	38.74	0.990	2.582M	37.60G
T-MSA $\rightarrow$ MDTA [78]	38.57	0.991	2.582M	35.20G
T-MSA $\rightarrow$ Swin [35]	36.59	0.988	2.517M	36.38G
T-MSA $\rightarrow$ TAR [28]	36.74	0.987	2.517M	34.00G
T-MSA $\rightarrow$ PVTv2 [72]	38.10	0.990	10.89M	38.60G
T-MSA $\rightarrow$ Cswin [16]	38.19	0.987	3.30M	40.10G
T-MSA $\rightarrow$ LinFormer [68]	36.12	0.983	48.70M	371.68G

Table 5: **Analysis of approximation errors.** The smaller approximation error for softmax-attention, the better the performance.

Methods	PSNR	SSIM	#Params	MACs
Swin	**36.59**	**0.988**	2.517M	36.38G
Swin + T-MSA-2nd	36.50	**0.988**	2.517M	36.38G
Swin + T-MSA-1st	36.37	0.987	2.517M	36.38G

Effectiveness of multi-scale attention refinement module. Table 3 demonstrates our MSAR module provides a favorable gain of 1.97dB over the counterpart without MSAR module, with only a tiny increase in the number of parameters (0.1M) and MACs (0.905G). When we set the gating tensor $G \in \mathbb{R}^{h\times1\times H\times W} \rightarrow G \in \mathbb{R}^{h\times C\times H\times W}$, it provides only a tiny improvement in PSNR and SSIM, at the cost of a significant increase in the number of parameters and computation complexity. Taking into account both performance and computational burden, we set $G \in \mathbb{R}^{h\times1\times H\times W}$.

Comparison with other linear self-attention modules. Table 4 compares the proposed T-MSA with several common linear self-attentive modules. Results show that the TaylorFormer has significant advantages over existing linear self-attention modules, which is attributed to the ability of T-MSA to model long-distance pixels and the approximation to the softmax-attention.

Analysis of approximation errors. To explore the approximation error and its impact, we investigate the ef-

Input Conv (Single) DSDCN-S (Double)

Figure 8: **Comparison of feature map visualization for single-scale and multi-scale tokens.** The design of multi-scale and multi-branch can capture more powerful features.

Table 6: **Results on the CSD dataset [8] for snow removal.** The best and second best results are highlighted in bold and underlined, respectively.

Methods	CSD (2000)		Overhead	
	PSNR↑	SSIM↑	#Param↓	MACs↓
DesnowNet [38]	20.13	0.81	26.15M	1.7KG
CycleGAN [17]	20.98	0.80	7.84M	42.4G
All in One [33]	26.31	0.87	44.00M	12.3G
HDCW-Net [8]	29.06	0.91	6.99M	9.8G
TKL [9]	33.89	0.96	31.35M	41.6G
SMGARN [10]	31.93	0.95	6.86M	450.3G
Uformer [73]	33.80	0.96	9.03M	19.8G
Restormer [78]	<u>35.43</u>	<u>0.97</u>	26.10M	141.0G
Ours (-B)	**37.10**	**0.98**	2.68M	38.5G

Table 7: **Results on the RainCityscapes dataset [25] for rain removal.** "-" indicates that the result is not available. The best and second best results are highlighted in bold and underlined, respectively.

	RESCAN [34]	DAFNet [25]	PCNet [27]	EPRRNet [83]	Uformer [73]	Restormer [78]	Ours(-B)
PSNR↑	28.97	30.06	24.49	31.11	32.26	<u>34.37</u>	**36.55**
SSIM↑	0.885	0.953	0.946	0.974	0.981	<u>0.988</u>	**0.990**
#Params↓	0.15M	-	0.63M	137.10M	9.03M	26.10M	2.68M
MACs↓	9.7G	-	2.4G	182.2G	19.8G	141.0G	38.5G

fect of different orders of Taylor expansion for softmax attention. Considering that the associative law is not applicable to the second-order T-MSA (T-MSA-2nd), which leads to a significant computational burden, we perform first-order and second-order Taylor expansions for Swin. Table 5 shows that T-MSA can effectively approximate softmax-attention, and the T-MSA-2nd is already very close to softmax-attention in performance. However, the computational complexity of both T-MSA-2nd and softmax-attention increases quadratically with image resolution, so it is difficult to model long-range pixel relationships in practical applications. Thus, we finally choose T-MSA-1st for our proposed method.

4.5. The Generalization Capabilities

Though this paper focuses on image dehazing, we have also evaluated the generalization capabilities of our proposed model on other tasks, specifically snow and rain removal, using the CSD [8] and RainCityscapes [25] datasets. The setting of the training and testing dataset for snow removal follows the existing methods [8, 10], and the dataset setting for rain removal follows existing methods [25, 83]. The results, presented in Table 6 and Table 7, demonstrate that our model, MB-TaylorFormer, performs well in these tasks, indicating that its capabilities are not limited to dehazing. The supplementary material includes visualization results for these tasks.

5. Conclusion

In this paper, we propose a multi-branch linearized Transformer network, called MB-TaylorFormer, which consists of multi-scale patch embedding and Taylor expansion of self-attention. Multi-scale patch embedding with flexible shape of receptive field, multi-scale size receptive field, and multi-level semantic information, can enable flexible embedding of diverse visual tokens. The Taylor expansion of softmax-attention using the matrix multiplicative associate law reduces the computational complexity. Further, we correct the output of self-attention by gating attention, which allows MB-TaylorFormer to perform both long-range attention and local corrections. Experimental results on various datasets demonstrate the effectiveness, lightness and generalization of the proposed MB-TaylorFormer.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62071500. Supported by Sino-Germen Mobility Programme M-0421. Supported by Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515012839. Supported by Shenzhen Science and Technology Program No. JSGG20220831093004008.

References

[1] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *ICIP*, pages 1014–1018. IEEE, 2019.

[2] Codruta O Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *CVPRW*, pages 754–762, 2018.

[3] Francesca Babiloni, Ioannis Marras, Filippos Kokkinos, Jiankang Deng, Grigorios Chrysos, and Stefanos Zafeiriou. Poly-nl: Linear complexity non-local layers with 3rd order polynomials. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10518–10528, 2021.

[4] Haoran Bai, Jinshan Pan, Xinguang Xiang, and Jinhui Tang. Self-guided image dehazing using progressive feature fusion. *TIP*, 31:1217 – 1229, 2022.

[5] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *TIP*, 25(11):5187–5198, 2016.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[7] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, pages 357–366, 2021.

[8] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4196–4205, 2021.

[9] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17653–17662, 2022.

[10] Bodong Cheng, Juncheng Li, Ying Chen, Shuyi Zhang, and Tieyong Zeng. Snow mask guided adaptive residual network for image snow removal. *arXiv preprint arXiv:2207.04754*, 2022.

[11] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017.

[12] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017.

[14] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2157–2167, 2020.

[15] Jiangxin Dong and Jinshan Pan. Physics-based feature dehazing networks. In *ECCV*, pages 188–204. Springer, 2020.

[16] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022.

[17] Deniz Engin, Anil Genç, and Hazim Kemal Ekenel. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 825–833, 2018.

[18] Lucas Teixeira Goncalves, Joel De Oliveira Gaya, Paulo Drews, and Silvia Silva Da Costa Botelho. Deepdive: An end-to-end dehazing method using deep learning. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 436–441. IEEE, 2017.

[19] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, pages 5812–5820, 2022.

[20] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *PAMI*, 33(12):2341–2353, 2010.

[21] Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *CVPR*, pages 3462–3471, 2020.

[22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, pages 1314–1324, 2019.

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[24] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, pages 3464–3473, 2019.

[25] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 8022–8031, 2019.

[26] Haobo Ji, Xin Feng, Wenjie Pei, Jinxing Li, and Guangming Lu. U2-former: A nested u-shaped transformer for image restoration. *arXiv preprint arXiv:2112.02279*, 2021.

[27] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Zheng Wang, Xiao Wang, Junjun Jiang, and Chia-Wen Lin. Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining. *IEEE Transactions on Image Processing*, 30:7404–7418, 2021.

[28] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

[29] Ashutosh Kulkarni and Subrahmanyam Murala. Aerial image dehazing with attentive deformable transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6305–6314, 2023.

[30] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *CVPR*, pages 7287–7296, 2022.

[31] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, pages 4770–4778, 2017.

[32] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 28(1):492–505, 2019.

[33] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020.

[34] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018.

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021.

[36] Jing Liu, Haiyan Wu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Trident dehazing network. In *CVPRW*, pages 430–431, 2020.

[37] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, pages 7314–7323, 2019.

[38] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018.

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[40] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[41] Yiqun Mei, Yuchen Fan, Yulun Zhang, Jiahui Yu, Yuqian Zhou, Ding Liu, Yun Fu, Thomas S Huang, and Humphrey Shi. Pyramid attention networks for image restoration. *arXiv preprint arXiv:2004.13824*, 2020.

[42] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[43] G Peano. Une nouvelle forme du reste dans la formule de taylor. *Mathesis*, 9:182–183, 1889.

[44] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. Ntire 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *CVPR*, pages 691–700, 2021.

[45] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, volume 34, pages 11908–11915, 2020.

[46] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022.

[47] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *CVPR*, pages 8160–8168, 2019.

[48] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *NIPS*, 32, 2019.

[49] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, pages 154–169. Springer, 2016.

[50] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, pages 3253–3261, 2018.

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[52] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.

[53] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016.

[54] Pranjay Shyam, Kyung-Soo Kim, and Kuk-Jin Yoon. Giqe: Generic image quality enhancement via nth order iterative degradation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2077–2087, 2022.

[55] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *arXiv preprint arXiv:2205.12956*, 2022.

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

[57] Jeong-geun Song. Ufo-vit: High performance linear vision transformer without softmax. *arXiv preprint arXiv:2109.14382*, 2021.

[58] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *arXiv preprint arXiv:2204.03883*, 2022.

[59] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[60] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[62] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, pages 5769–5780, 2022.

[63] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017.

[65] Anna Wang, Wenhui Wang, Jinglu Liu, and Nanhui Gu. Aipnet: Image-to-image single image dehazing with atmospheric illumination prior. *TIP*, 28(1):381–393, 2018.

[66] Chenxi Wang and Zhi Jin. Brighten-and-colorize: A decoupled network for customized low-light image enhancement, 2023.

[67] Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information, 2023.

[68] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[69] Tao Wang, Guangpin Tao, Wanglong Lu, Kaihao Zhang, Wenhan Luo, Xiaoqin Zhang, and Tong Lu. Restoring vision in hazy weather with hierarchical contrastive learning. *arXiv preprint arXiv:2212.11473*, 2022.

[70] Tao Wang, Kaihao Zhang, Ziqian Shao, Wenhan Luo, Bjorn Stenger, Tong Lu, Tae-Kyun Kim, Wei Liu, and Hongdong Li. Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions. *arXiv preprint arXiv:2305.17863*, 2023.

[71] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021.

[72] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[73] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022.

[74] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2021.

[75] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *ICCV*, pages 9981–9990, 2021.

[76] Yan Yang, Haowen Zhang, Xudong Wu, and Xiaozhen Liang. Mstfdn: Multi-scale transformer fusion dehazing network. *Applied Intelligence*, pages 1–12, 2022.

[77] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[78] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022.

[79] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, pages 492–511. Springer, 2020.

[80] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *CVPR*, pages 3194–3203, 2018.

[81] Kaihao Zhang, Dongxu Li, Wenhan Luo, Wenqi Ren, and Wei Liu. Enhanced spatio-temporal interaction learning for video deraining: faster and better. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1287–1293, 2022.

[82] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing*, 30:7419–7431, 2021.

[83] Kaihao Zhang, Wenhan Luo, Yanjiang Yu, Wenqi Ren, Fang Zhao, Changsheng Li, Lin Ma, Wei Liu, and Hongdong Li. Beyond monocular deraining: Parallel stereo deraining network via semantic prior. *International Journal of Computer Vision*, 130(7):1754–1769, 2022.

[84] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020.

[85] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022.

[86] Xiaoqin Zhang, Tao Wang, Jinxin Wang, Guiying Tang, and Li Zhao. Pyramid channel-based feature attention network for image dehazing. *Computer Vision and Image Understanding*, 197:103003, 2020.

[87] Dong Zhao, Jia Li, Hongyu Li, and Long Xu. Complementary feature enhanced network with vision transformer for image dehazing. *arXiv preprint arXiv:2109.07100*, 2021.

[88] Lirong Zheng, Yanshan Li, Kaihao Zhang, and Wenhan Luo. T-net: Deep stacked scale-iteration network for image dehazing. *IEEE Transactions on Multimedia*, 2022.

[89] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *TIP*, 24(11):3522–3533, 2015.