EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

A hierarchically sampling global sparse transformer in data stream mining for lightweight image restoration

Mingzhu Shi[1,2*], Bin Zao[1,2], Chao Wang[3], Muxian Tan[1,2], Siqi Kong[1,2] and Shouju Liu[3]

*Correspondence:
shimz@tjnu.edu.cn

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
[3] School of Engineering and Digital Arts, University of Kent, Canterbury CT2 7NT, UK

Abstract

With the rapid development of information technology, mining valuable information from multi-source data stream is essential for redundant data, particularly in image processing; the image is degraded when the image sensor acquires information. Recently, transformer has been applied to the image restoration (IR) and shown significant performance. However, its computational complexity grows quadratically with increasing spatial resolution, especially in IR tasks to obtain long-range dependencies between global elements through attention computation. To resolve this problem, we present a novel hierarchical sparse transformer (HST) network with two key strategies. Firstly, a coordinating local and global information mapping mechanism is proposed to perceive and feedback image texture information effectively. Secondly, we propose a global sparse sampler that reduces the computational complexity of feature maps while effectively capturing the association information of global pixels. We have conducted numerous experiments to verify the single/double layer structure and sampling method by analyzing computational cost and parameters. Experimental results on image deraining and motion deblurring show that the proposed HST performs better in recovering details compared to the baseline methods, achieving an average improvement of 1.10 dB PSNR on five image deraining datasets and excellent detail reconstruction performance in visualization.

Keywords: Image restoration, Transformer, Global sparse attention, Stochastic sampler

1 Introduction

Data stream mining from multi-source, massive dynamic data to extract valuable features is of significant importance. Especially in image processing, since some details are lost during imaging and transmission, data stream mining plays a key role in recovering clean images. Image restoration (IR) aims to reconstruct high-quality images by mining useful information from degraded, disturbed, and detail-loss data (e.g., rainy image, blurry image). Images have a lot of redundant information in the process, such as similar lines and edges, symmetric local regions and consistent distribution of noise. Therefore, mining useful features from replicated information at different scales to improve the image quality and relieve the computational burden, data stream mining is urgently

needed. This paper aims to process multiple image restoration tasks by extracting data features to form generalizable priors through a sparse transformer approach.

Recent methods [1–4] based on convolutional neural networks (ConvNets) show excellent performance compared with traditional prior-based methods on the basis of evaluation metrics such as PSNR and SSIM, attributed to the combination of model design and the local connectivity of convolutions. However, the convolution operation lacks perception of the long-distance dependence of pixels of the limited receptive field. To address this problem, an efficient self-attention mechanism is generated to associate global weights, and interact the pixels in the feature map directly. Transformer was first proposed in natural language tasks [5] and designed to solve sequence-to-sequence tasks, and then, the methods [6–9] explored its performance in image restoration tasks, where it exhibited remarkable performance in global textures, but the computational complexity grows quadratically with spatial resolution attributed to its unique computational approach. Whether ConvNets or transformer, learning sufficient generalization priors often need enormous model structures [2, 6, 9]; fortunately, the lightweight networks with outstanding advantages in inference speed and memory storage provided capable backbones for related vision tasks. Mobilenets [10] proposed depth separable convolution reduces the Multi-Adds of the model by about nine times; the lightweight transform network [11] combined the advantages of ConvNets and vision transformer (ViT) [12] to improve the inference speed and reduce the parameters. Then, the strategy was rapidly upgraded and optimized to be applied in multiple tasks such as object detection and image classification. [13–16].

In this paper, we aim to exploit the global perception capability of the self-attention mechanism for image restoration while reducing the computational effort. To this end, we propose a transformer-based image restoration structure hierarchical sparse transformer (HST) network. Existing transformer-based methods mostly use double-layer computational module, of course, although the equipment is advanced enough and more parameters can support to get better results, specific tasks inevitably encounter computational bottlenecks and it is hard to achieve satisfactory experimental results with as few layers and parameters as possible. Therefore, we attempt to design a single-layer structure to reduce the amount of the parameters and get a good trade-off between performance and complexity. Moreover, we have conducted extensive experiments to compare the performance and found that our results are comparable to the current excellent methods, the overall number of parameters is reduced by about 48.1% and the training time is reduced by about 44.7%, compared to the double-layer model, it is simple and efficient.

Specifically, we present a three-stage scheme to recover high-quality images, which is achieved by optimally integrating self-attention and convolution to form a highly cost-effective information exchange bottleneck. The first stage is local, local means local context aggregation, using efficient depth-wise convolutions to extract the spatial proximity of local pixels. The next stage is global that uses sparse attention to global pixels; to alleviate the computational burden of the transformer through a sparse strategy while obtaining long-range dependencies, a suitable stochastic sampler is designed to form a sparse set of evenly distributed delegate tokens for long-range information exchange by self-attention. The last stage is local, meaning local propagation; the updated long-range

information in the previous stage is diffused from delegated tokens to non-delegated tokens in the local neighborhood via pixel-unshuffle operations. With the above three stages, the local interaction and global sparse strategies are functionally complementary and the model forms a refined hybrid of self-attention and convolution, while the experimental data show a well balanced between lightweight and performance.

Secondly, we propose a global sparse attention (GSA) module to capture long-range pixel interactions and reduce the computational cost. Meanwhile, we design a stochastic sampler as the core component in the GSA module to enhance the generalization ability. In this work, firstly, the stochastic sampler downsamples the feature maps with different ratios according to layers, which is considering the inherent property of spatial redundancy of images and that interpreting all elements is inefficient. Then, the module conducts self-attention on the activated elements. Specifically, at each sampling, the elements participating in the interaction are specified randomly, unlike some operations (e.g., maximum pooling, average pooling) that use fixed rules to activate elements, this strategy ensures that the global relationships between pixels are fully interacted while computing attention maps. The experimental results in several datasets prove that our approach is feasible. Furthermore, we compare the experimental results of our proposed method with other pooling operations, as an example shown in Fig. 1, and show that it outperforms other methods in terms of detailed preservation and has better generalization capability.

Based on above components, we conduct extensive experiments to demonstrate the generalization performance of HST on eight benchmark datasets for image deraining and image motion deblurring. The streamlined structural design shows satisfactory results while enabling a significant reduction in training time and computational cost. In addition, ablation experiments are carried out to illustrate architectural efficiency and component functionality.

Overall, the contributions of this work are summarized as follows:

- We propose HST, a pyramid transformer with a streamlined layer structure. It adopts the ethos of local–global information flow interaction for image restoration and exploits long-range dependencies by reducing global pixel density, followed by sparse attention.
- A stochastic sampler suitable for global sparse attention is presented. The sampler evenly stochastically samples spatial elements without using a single rule to cluster the elements, thereby obtaining generalized priors and relieving the global computational burden.

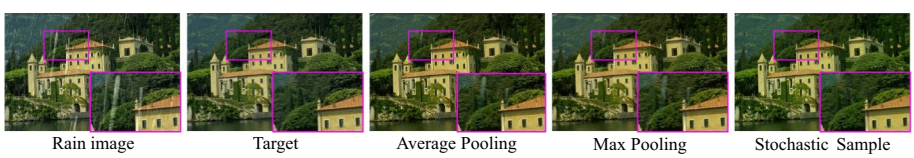

Fig. 1 An example of the Rain12 dataset. Selected regions are zoomed-in and displayed in the bottom right corner of each image. Details of raindrops in the image, which can be better reconstructed with HST compared to other methods (combination of pooling and self-attention), while the image color is closer to the target

- We design a global sparse attention module. The module forms a functional complementarity between stochastic sampler and self-attention mechanism by sparsely capturing global context. Experimental results show that our method outperforms other combinations in terms of generalization and detail recovery

In the following sections of this paper, we discuss in Sect. 2 the application of the classical UNet structure to image restoration and variants of transformer for specific tasks. In Sect. 3, we give a detailed explanation of our approach. In Sect. 4, we describe the experimental details and compare our results with other methods for evaluation metrics and visualization, followed by a concluding statement in Sect. 5.

2 Related work

2.1 UNet-based image restoration

Compared with traditional image restoration methods [17–20], deep learning based methods show outstanding performance and have become more and more popular in image restoration tasks [4, 21–23]. In terms of architecture design, UNet [24] exhibits an excellent perception of texture details in image segmentation tasks due to its elegant symmetric path. Subsequently, many efforts based on convolutional neural networks also use U-shaped architecture. DeblurGAN-v2 [25] applies this symmetrical structure as the core building block of the generator in the generative adversarial network for image deblurring, which greatly improves the deblurring efficiency and has high flexibility in the quality efficiency spectrum, but the spatial details are not preserved well enough. To improve the spatial accuracy, MPRNet [2] constructs an encoder-decoder subnetwork to learn contextual information in each progressive stage of restoring degraded images. Besides, considering the inability of CNN to model the dependencies between distant pixels, some recent works explore combining efficient self-attention mechanism with U-shaped architecture to improve performance. Uformer [8] keeps the same overall architecture as UNet, modifying the convolutional layers into transformer blocks and using non-overlapping window-based self-attention to handle the image restoration task, thus reducing computational overhead, however, its window operation limits the interaction of SA in the global domain. For reducing the computational burden and applying to image reconstruction tasks involving high resolution images, Restormer [9] also employs a similar structure and computes the cross-covariance between feature channels, without decomposing them into local windows, thereby exploiting the distant image context. Therefore, a symmetric pyramid structure with skip connections is an effective method to extract deep texture information from images. In this paper, we also carry out related work on this benchmark.

2.2 Vision transformer

The self-attention mechanism [5] is proposed in natural language processing tasks; due to its excellent contextual relevance, it is rapidly applied in image classification [12, 26], object detection [27, 28], and image segmentation [29, 30]. A pioneering VIT [12] model divides the processing window and flattens patches, significantly reducing the computational cost for all elements of the entire feature map. IPT [6] first adopted this approach in image restoration, but this strategy lost the boundary pixel information of each patch.

To this end, a swin transformer [31] that can slide the divided window is proposed, which can use window shifting to aggregate the information flow of neighboring image patches. This method is quickly verified on the image restoration task [7], achieving state-of-the-art performance. However, these schemes restrict the operational domain of self-attention. In order to perform self-attention globally, Restormer [9] calculates attention in the feature channel dimension after transposition in the image restoration task, reducing computational complexity while capturing long-range dependencies. Most of the above work focuses on the contradictory problem of global context information and computational complexity, and we notice that spatial redundancy in the feature map is unavoidable during calculation. Therefore, we design hierarchical sparse transformer to reduce the scale of continuous redundant regions, cooperating with a multilayer pyramid structure to gradually enrich long-range pixel relationships through multiple iterations.

2.3 Lightweight strategy

The research of lightweight image restoration based on deep learning has never stopped. Some methods use cascading and hierarchical architectures to optimize computational efficiency and memory consumption. Zhang [32] built a three-scale end-to-end network in which different convolution operations are associated with cascade and dense connections, respectively, to achieve fast image dehazing; Fu [33] constructed a hierarchical pyramid network for image deraining with few parameters by introducing a mature Gauss-Laplace decomposition technique. There are methods to perform related work on the basic calculation modules. Avisek [34] based on image restoration by designing streamlined modules (1x1 pointwise convolution with compressed number of channels and parallel branching structure strategy) instead of the current commonly used computational blocks (e.g., 3x3 convolution), allowing significant reduction of parameters and FLOPs in the network. There are also approaches based on neural architecture search algorithms [35], such as Shen [36] proposed a joint search operation to hunt for efficient lightweight image restoration networks. Therefore, using lightweight models for lower-level tasks is more elegant and efficient, and is a topic worth exploring.

3 Method

In this section, we firstly describe the overall architecture of the proposed hierarchically sampling global sparse transformer (HST). Then, we will introduce the details of the local–global interaction mechanism of the LGL block. Finally, we analyze the global sparse attention operation principle and the role of the stochastic sampler in the global long-range dependence tasks. The architecture of our proposed hierarchical sparse transformer (HST) network is shown in Fig. 2.

3.1 Overall pipeline

As shown in Fig. 2a, the overall structure of our model is a hierarchical network with a U-shaped pyramid structure that is connected across levels. Firstly, given a degraded image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$, where $H \times W$ denotes the spatial dimension and C is the number of channels. Preprocessing is performed before inputting to the HST network, where the images are randomly cropped and expanded into a 128×128 square

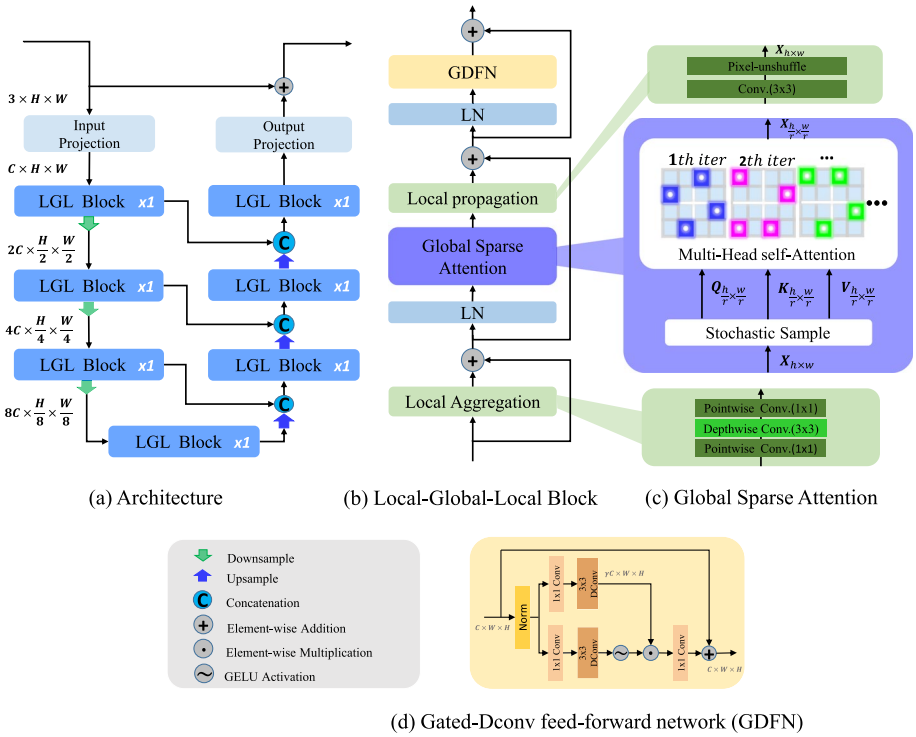

Fig. 2 The architecture of the hierarchical sparse transformer (HST) network. **a** Overview of the HST structure. **b** LGL transformer block. **c** Illustration of local–global information interaction operations, including sparse attention and stochastic sample. **d** Gated-Dconv feed-forward network (GDFN) [9]

image set for fast training and data enhancement. Subsequently, the network applies a 3×3 convolutional and LayerNorm layer to learn low-level features and generate $\mathbf{X_0} \in \mathbb{R}^{C \times 128 \times 128}$. Then, we will go through K encoder stages, each of which contains an LGL block and a down-sampling layer. The LGL block can correlate local–global information and effectively reduce the computational burden through global sparse attention module. The down-sampling layer doubles the channel of the feature map and reduces the height and width by half through a convolution operation with a convolution kernel of 4×4 and stride 2. The encoder produces the feature maps in l-th stage $\mathbf{X}_l \in \mathbb{R}^{2^l C \times \frac{128}{2^l} \times \frac{128}{2^l}}$. After the three-layer encode process, the deep features extracted hierarchically will be inputted to the bottom levels with the same LGL block.

The image features are reconstructed subsequently. It is a k-stage decoding path symmetrically distributed with the encoder, each of which contains an upsampling operations and an LGL block; a pixel-unshuffle operations [37] is used in the feature upsampling to halve the channels and double the dimension. Then, the skip connections concatenate the encoded features with the decoded features. To reduce the number of channels by half, the concatenation operation is followed by 1×1 convolution. The multilayer architecture by fusion of high and low layer feature information is efficient for preserving the fine structure and texture details of the image. Finally, we reshape the dimensions of the image using 3×3 convolution to obtain the residual image $\mathbf{M} \in \mathbb{R}^{C \times 128 \times 128}$. and add the residual image to the degraded image to acquire

the restored image by the formula $\mathbf{I}' = \mathbf{I} + \mathbf{M}$, where $\mathbf{I}'$ represents the restored image and $\mathbf{I}$ is the degraded image.

3.2 LGL block

Performing transformer on the global pixels would make the computational burden too heavy, and conducting the associated windowing operation on the feature map would limit the long-distance interaction between elements, the two have become an intractable conflict. Inspired by the excellent performance of lightweight networks [38] in long-range handling relationships between a set of delegation tokens, we propose a local–global information interaction transformer block, which includes a core module that performs self-attention computation after random sampling. In the block the local part captures useful local context, the global component models long-range pixel dependencies, and the local information and global information are exchanged with each other timely.

As shown in Fig. 2b, we build the block with three core designs: (1) local aggregation (LA); (2) global sparse attention (GSA); (3) local propagation (LP). The information flow of each computational module is element-wise addition. In the following, we elaborate three designs separately.

Local aggregation (LA): Neighboring pixel feature association is a vital information reference for image restoration tasks [8, 39]. Depth-wise convolution can decrease the number of parameters and improve the operation efficiency while ensuring the effect of feature extraction. Then, we introduce depth-wise and point-wise convolution in local aggregation to emphasize the local context. As shown in Fig. 2c, we use 1×1 point-wise convolution at the beginning and end to control the dimension and provide depth-wise convolutions in the middle to capture local interaction information, the convolution kernel size with 3×3, groups with dimension. An activation function LN is added after each convolutional layer. Local aggregation can be formulated as:

$$\mathbf{X}'_{LA} = Conv_D(Conv_P(LN(\mathbf{X}))) \tag{1}$$

$$\mathbf{X}''_{LA} = Conv_P(LN(\mathbf{X}'_{LA})) \tag{2}$$

where $Conv_P$,$Conv_D$ and LN denote pixel convolution, depth-wise convolution, and layer normalization.

Global Sparse Attention (GSA): Image features have so high redundancy that applying self-attention to global tokens is not cost-effective. On the other hand, the long-distance dependency between pixels is the crucial reference information for restoration work. Therefore, we have an idea for self-attention after reducing the scope of the tokens. We first define the feature information after local aggregation in the previous stage as a uniform area, then stochastically select elements in the area to form tokens instead of taking a single rule to activate the elements, and finally perform multi-head self-attention on the tokens. Given the feature maps $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, the computational complexity drops from $\mathcal{O}(H^2W^2C)$ to $\mathcal{O}((\frac{H}{r})^2(\frac{W}{r})^2C) = \mathcal{O}(\frac{H^2W^2}{r^4}C)$, where r represent the sampling ratio, which is set differently in structures of different levels, compared to global self-attention, global sparse attention can significantly reduce the computational cost.

Local Propagation (LP): Feeding back the feature information from the global interaction to neighboring pixels is also an indispensable step, which converts the sampled feature map of dimension $\mathbf{X} \in \mathbb{R}^{C \times \frac{H}{r} \times \frac{W}{r}}$ to $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$. Figure 2c shows that we use 3×3 convolution to expand the feature dimension and then use the pixel-unshuffle operation [37] to feed back the fused contextual information and return the original scale. It can be expressed as:

$$\mathbf{X}'_{LP} = pixelunshuffler(Conv_{3\times3}(\mathbf{X})) \tag{3}$$

To further capture the fine information of the image, we also employ the Gated-Dconv Feedforward Network (GDFN) [9]. It enables the network to focus on recovering high frequency details with contextual information; we put it at the end of the block to complement the fine details. as shown in Fig. 2d, given an input feature $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, GDFN can be expressed as:

$$\mathbf{X}_G^1 = GELU(Conv_D(Conv_P(LN(\mathbf{X})))) \tag{4}$$

$$\mathbf{X}_G^2 = Conv_D(Conv_P(LN(\mathbf{X}))) \tag{5}$$

$$\mathbf{Z}_G = \mathbf{X}_G^1 \odot \mathbf{X}_G^2 \tag{6}$$

$$\mathbf{Z} = Conv_P(\mathbf{Z}_G) \tag{7}$$

where $Conv_P$,$Conv_D$ and LN denote pixel convolution, depth-wise convolution, and layer normalization. The ability of GDFN to control the fine details of images has been demonstrated in several restoration tasks [9, 40].

Overall, our LGL block effectively helps the network to obtain global contextual information with less computational burden. The l-th stage of the LGL block can be formulated as:

$$\mathbf{X}' = LA(\mathbf{X}_{l-1}) + \mathbf{X}_{l-1} \tag{8}$$

$$\mathbf{X}'' = LP(GSA(LN(\mathbf{X}'))) + \mathbf{X}' \tag{9}$$

$$\mathbf{X}_l = GDFN(LN(\mathbf{X}'')) + \mathbf{X}'' \tag{10}$$

3.3 Global sparse attention

The computing overhead of the transformer is mainly concentrated in the self-attention operation. The memory complexity of query-key-value increases quadratically with the spatial resolution of the input. Therefore, applying self-attention to image restoration tasks requires high computing power platform support. To alleviate this issue, we propose a stochastic sampler to specify regions uniformly with a rate of r, take out a feature element per $r \times r$ region, and then reshape it into a token for multi-head self-attention, the number of factors after sampling:$N = \frac{HW}{r^2}$. The process can be formulated as follows:

$$\begin{aligned} \mathbf{X}' &= Sample_{stochastic}(\mathbf{X}^{C\times H\times W}) \\ \mathbf{X}' &\in \mathbb{R}^{C\times \frac{H}{r}\times \frac{W}{r}} \end{aligned} \tag{11}$$

Then, we generate the $\mathbf{Q}(query)$, $\mathbf{K}(key)$ and $\mathbf{V}(value)$ projections, $\mathbf{Q} = W_d^Q W_p^Q \mathbf{X}'$, $\mathbf{K} = W_d^K W_p^K \mathbf{X}'$, $\mathbf{V} = W_d^V W_p^V \mathbf{X}'$, where $W_d^{(*)}$, $W_p^{(*)}$ denote 3×3 depth-wise convolution, 1×1 point-wise convolution. In the Following calculation of attention, inspired by advanced experience of predecessors [9], we add a lightweight learnable scale parameter β to the self-attention matrix mapping to control the magnitude after $\mathbf{Q}$ and $\mathbf{K}$ dot product. Attention employs a multi-head processing mechanism and can be formulated as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \text{Softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}}/\beta) \tag{12}$$

Since the dimensionality of the feature map is reduced layer by layer in the pyramid structure, and different sampling rates are provided in the computational module at each level to decrease the density of global elements, this greatly reduces the overall computational burden. With stochastic element attention, a given pixel in the feature map has the opportunity to calculate the weighted sum with all other pixels at all other positions during long training time, which make the model obtain more generalized priors.

3.4 Stochastic sampler

The image restoration task requires the restoration of texture details to be closer to the truth. However, the maximum pooling only focuses on the maximum value of the region, and the average pooling has insufficient ability to analyze the edge and contour information of the image. These operations are unsuitable for restoration work with high requirements on texture features. For this reason, we design a stochastic sampler to enhance the generalization of model to various restoration tasks.

As shown in Fig. 3, after obtaining a complete feature map, different from the previous work [7, 8] to divide the window and then shift the window, our strategy is to set the sampling rate r, generate stochastic pointing numbers to sample elements in a square area with value r^2. This operation will generate $N = \frac{HW}{r^2}$ highlighting factors in the full-size feature maps and the objects of each iteration are different; as shown in Fig. 3c, blue represents the feature set of the first iteration, and pink represents the next iteration. This strategy enables elements from different regions to have the

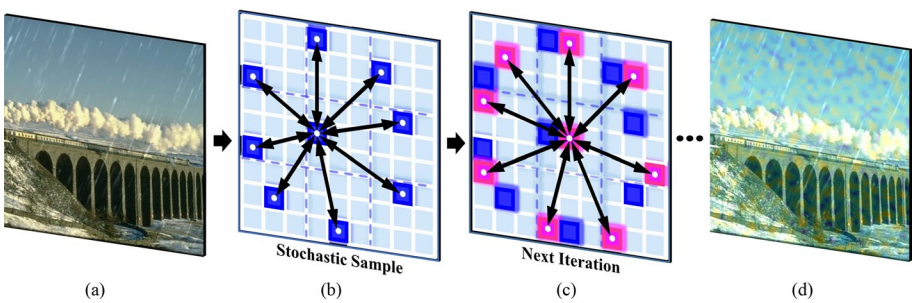

Fig. 3 Illustration of information flow for stochastic sampler

possibility to participate in the computation with constant iterations; the main steps of the scheme can be seen in Algorithm 1.

Algorithm 1: Stochastic Sampler

Input: original image X; sample ratio r
Output: sampled image Y
Compute the number of samples N according to X, r
Built an empty list $index = [\cdots]$
Partitioning the sampling area by dimensional transformation
Generate the overall index list
for $i = 0$ *to* N **do**
    Generate a random pointing number n according to N;
    Append n to the list index in order;
end
Sampling X according to $index$
Reverse the original position of the sampled elements by dimensional transformation
return Y

The stochastic sampler is able to associate global pixels with different combinations of sparse tokens at each indexing, so it plays a key role in maintaining long distance dependency and reducing the global computational burden.

4 Experiments

In this section, we discuss the experimental setup and our specific work on the two types of recovery tasks, showing the visualization results of HST and the performance comparison under the evaluation metrics. Finally, ablation studies analyze the effect of parameter settings on the overall performance.

4.1 Experimental setup

Preprocessing is performed before inputting to the HST network, where the images are randomly cropped and expanded into a 128 × 128 square image set for fast training and data enhancement, and set the batch size to 32. Following the common training strategy, we select Adamw optimizer [41] and the momentum terms of (0:9; 0:999), set the weight decay to 0.02, the initial learning rate is 0.0003, use the cosine decay strategy and reduce the learning rate. The network employs a 3-level encoder-decoder layer and a bottom layer, and the hierarchical structure corresponding transformer block number is [1,1,1,1], attention heads in GSA are [1; 2; 4; 8], the number of channels is [32, 64, 128, 256]. All experiments are trained on an NVIDIA GTX 2080Ti GPU.

4.1.1 Structural variants and parameters

We apply two structural variants, shown in Table 1, HST-tiny and HST-double. The specific parameter details, number of parameters, and the computational complexity in each variant are as follows:

Table 1 Parameter settings and corresponding size

HST	Channel	Depths	Sample rate	#params	GMACs
Tiny	32	{1,1,1,1}	{4,2,2,1}	5.53M	3.88G
Double	32	{2,2,2,2}	{4,2,2,1}	10.66M	6.32G

The computational cost and the number of parameters of HST shown in Table 1 are in a low order of magnitude, which can complete the network training in a short time. The experimental results in the following tasks are all based on HST-tiny training. More details of the experimental comparison of the two structures are presented in Section Ablation Studies.

4.1.2 Evaluation metrics and datasets

We calculate restoration performance using the commonly used PSNR/SSIM [60] metrics. For deraining, we evaluate the PSNR/SSIM on the Y channel in the YCbCr color space. And in Table 2, we list the datasets used for training and validation, including deraining and motion deblurring.

4.2 Image deraining results

With the same dataset as [2, 9], we train HST on 13712 clean-rainy image pairs shown in Table 2. The original dataset was randomly cropped to generate a small size dataset with patch size 128×128 for training. We train the deraining task for 250 epochs with the above settings and evaluate it on various testsets, including Rain100H [43], Rain100L [43], Test100 [42], Test2800 [44], and Test1200 [45].

We compare with 6 deraining methods: DerainNet [46], SEMI [47], DIDMDN [45], UMRL [48], RESCAN [49], PreNet [50]. As shown in Table 3, HST presents significantly better performance. The gain improved by 2.97dB on individual datasets, e.g., Test100. Evaluation performance on multiple testsets proves the generalizability of the HST. In Fig. 4, we show the visualization results for the three datasets, where HST can successfully remove rain streaks and capture more local details compared to other methods.

Table 2 Dataset descriptions for two types of image restoration tasks

Tasks	Datasets	Training samples	Testing samples	Testset rename
Deraining	Rain14000 [44]	11200	2800	Test2800
	Rain1800 [43]	1800	0	–
	Rain800 [42]	700	100	Test100
	Rain100H [43]	0	100	Rain100H
	Rain100L [43]	0	100	Rain100L
	Rain1200 [45]	0	1200	Test1200
	Rain12 [59]	12	0	–
Motion Deblurring	GoPro [51]	2103	1111	–
	HIDE [52]	0	2025	–
	RealBlur [53]	0	1960	–

Table 3 Image deraining results on five datasets

Methods	Test100 [42]		Rain100H [43]		Rain100L [43]		Test2800 [44]		Test1200 [45]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DerainNet [46]	22.77	0.810	14.92	0.592	27.03	0.884	24.31	0.861	23.38	0.835
SEMI [47]	22.35	0.788	16.56	0.486	25.03	0.842	24.43	0.782	26.05	0.822
DIDMDN [45]	22.56	0.818	17.35	0.524	25.23	0.741	28.13	0.867	29.65	0.901
UMRL [48]	24.41	0.829	26.01	0.832	29.18	0.923	29.97	0.905	30.55	0.910
RESCAN [49]	25.00	0.835	26.36	0.786	29.80	0.881	31.29	0.904	30.51	0.882
PreNet [50]	24.81	0.851	26.77	0.858	32.44	0.950	31.75	0.916	31.36	0.911
HST	27.97	0.869	27.73	0.843	32.70	0.938	32.72	0.931	31.54	0.914

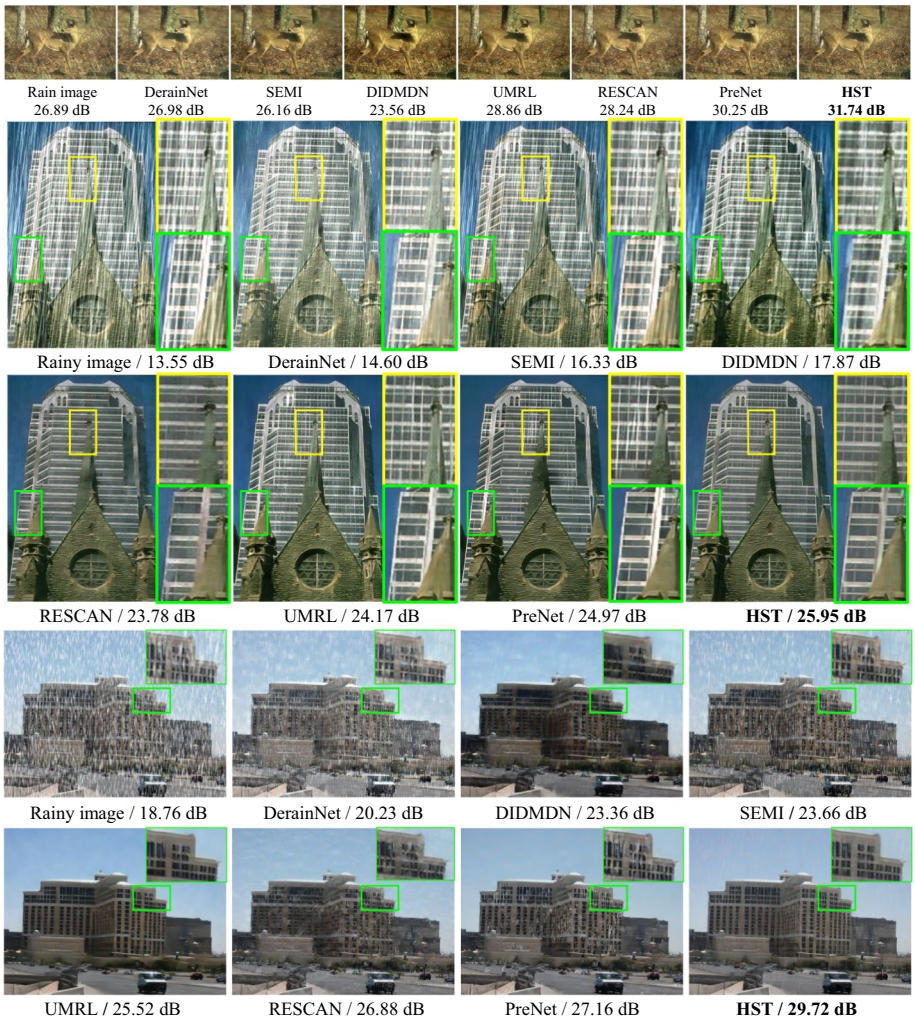

Fig. 4 Visualize results on different datasets, including Rain100L, Rain100H, and Test100

4.3 Image deblurring results

We follow the previous method [2, 8] to train HST on the GoPro [51] dataset and test it on the four datasets: synthesized test set of GoPro [51] and HIDE [52], two real-world datasets (RealBlur-R [53], RealBlur-J [53]).

Table 4 Performance comparison of image deblurring

Method	GoPro [51]		HIDE [52]		RealBlur-R [53]		RealBlur-J [53]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Xu et al. [54]	21.00	0.741	–		34.46	0.936	27.14	0.830
DeblurGAN [55]	28.70	0.858	24.51	0.871	33.79	0.903	27.97	0.834
Nah et al. [51]	29.08	0.914	25.73	0.874	32.51	0.841	27.87	0.827
Zhang et al. [56]	29.19	0.931	–		35.48	0.947	27.80	0.847
Deblur-GAN-V2 [25]	29.55	0.934	28.61	0.875	35.26	0.944	28.70	0.866
SRN [57]	30.26	0.934	28.36	0.915	35.66	0.947	28.56	0.867
Shen et al. [52]	–		28.89	**0.930**	–		–	
Gao et al. [58]	30.90	0.935	29.11	0.913	–		–	
HST	30.94	0.934	29.29	0.913	33.93	0.943	27.63	0.867

To validate the effectiveness of deblurring, we compare it with other excellent methods after training only on the GoPro dataset. The results are list in Table 4, which shows that our restoration method behaves well in real scenarios after training on synthetic datasets. Besides, we offer some visual results in Fig. 5, which shows that the restored image of HST is sharper and closer to the actual scene image compared with other methods (Table 4).

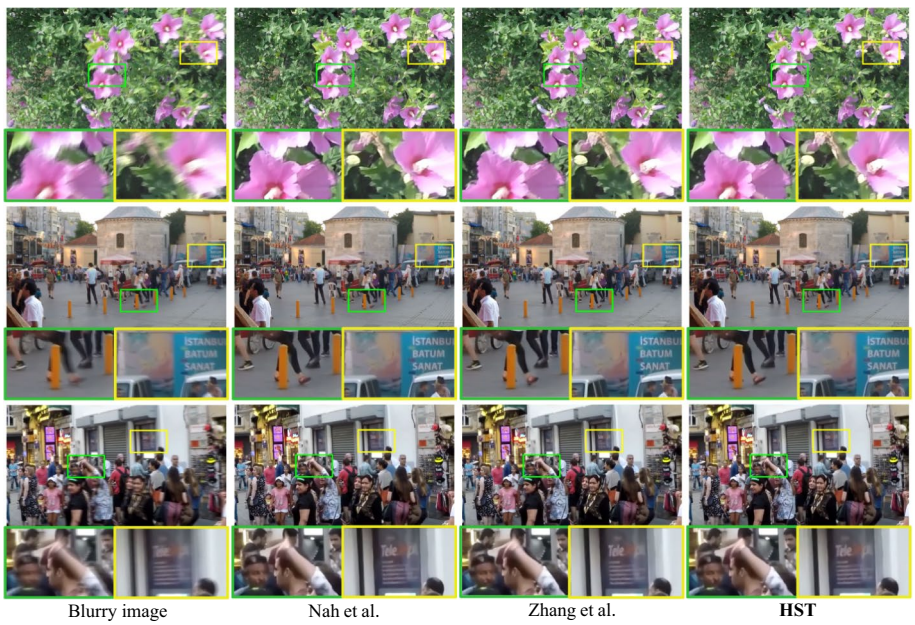

Fig. 5 Visualize results on image deblurring

Table 5 Results comparison of single/double layer structure

	Patch size	Channels	Epoch	GPUs	PSNR	Train time (h)
HST-tiny	128^2	32	200	1	29.17	12.6
HST-double	128^2	32	200	1	29.96	22.8

4.4 Ablation studies

In this section, we analyze the effect of GSA network variants on the restoration effect and the detailed analysis of the stochastic sampler in specific experiments. All the ablation studies are conducted on a heavy rainy dataset [43] with 1800 rainy images for training and 100 rainy images (Rain100H) for testing.

4.4.1 Single or double

In our task, whether the double-layer model structure can significantly improve the restoration effect has always been a key problem. Therefore, we conducted an experimental comparison between the double layer and single layer. The experiments only change the model single-double layer structure, input 36,000 patches of size 128×128 into the network for training, and other parameters are set the same as in the previous sections. The total training time and experimental results are shown in Table 5. It can be seen that the PSNR is not greatly improved, but the training time is significantly increased.

4.4.2 Stochastic sample, max pooling, and average pooling

Firstly, we conduct experiments to understand intuitively the feature maps involved in the computation under multiple iterations. We input the same image into the network, and under different sampling strategies, the position information of each activated pixel is recorded and accumulated with the number of iterations and finally plotted as a heat map. As shown in Fig. 6, the more frequently the relevant position is activated, the brighter the color is. It can be seen that the maximum pooling exercises a fixed rule and many elements in the feature map are not extracted to be involved in the computation, while our proposed stochastic sampling makes the overall activation of the pixels rise.

For the combination of stochastic sampler and self-attention mechanism, we replace this sampler with the maximum pooling and the average pooling for experimental

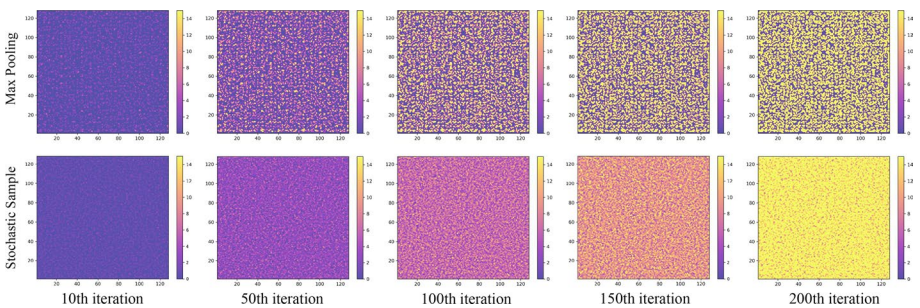

Fig. 6 Heat map of the degree of activation of the global pixels by the two approaches

Table 6 Results comparison between pooling and sampling

Method	Rain100H [43]		Rain100L [43]		Rain12 [59]		Train time (h)
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
Max pooling	29.40	0.847	33.86	0.966	33.62	0.954	12.4
Average pooling	29.29	0.838	33.71	0.963	32.93	0.950	12.5
Stochastic sample	29.27	0845	34.21	0.967	33.65	0.956	12.6

Fig. 7 The visualization results of the three combined operations for image textures

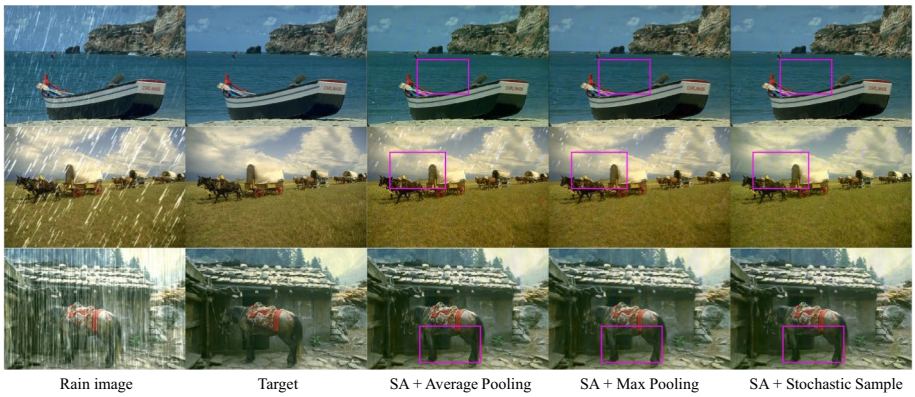

Fig. 8 The results of the three combined operations on different areas of the image

comparison. The experimental setup is based on a single-layer structure, replacing only the type of sampler. The experimental results are shown in Table 6. The visualization results of different strategies in image texture are shown in Fig. 7, and it can be seen that our proposed method is more generalized and restores images closer to the ground truth. We carefully observed the results of all raindrop removal experiments and found that stochastic sample had the best performance in areas with close pixel values to raindrops (e.g., sky, white walls, snow); in areas with significant color contrast to raindrops (black background, dark clothing, night sky), the maximum pooling operation was more effective in removing this part of raindrops, examples on the three datasets are shown in

Fig. 8; and for the subjective perception of the visualization results in terms of chromaticity, stochastic sample is closer to the target.

5 Conclusion

In this paper, we perform multi-source data stream mining in image restoration and propose a transformer network HST. The hierarchical pyramid structure can effectively keep the texture details; the sparse attention strategy enables to capture long-range pixel interactions; the randomly specified delegation token approach enhances the model generalization. The above works reduce the computational complexity and achieves fast training and image restoration.

We design the key approach for the core components of the transformer block. The stochastic sampler plays a critical role in the trade-off between generalizability and computational burden. Experimental results demonstrate that the stochastic sampler has excellent generalization performance, and HST is validated on eight datasets for image deraining and deblurring tasks, achieving excellent performance.

Abbreviations

IR	Image restoration
HST	Hierarchical sparse transformer
LGL	Local–global–local
GSA	Global sparse attention
GDFN	Gated-Dconv feedforward network

Acknowledgements
The authors would like to express their sincere thanks to the editors and anonymous reviewers.

Author contributions
All authors contributed equally to this work. All authors read and approved the final manuscript.

Funding
This work was supported by the National Science Foundation of China under Grant 61501328, Enterprise Joint Horizontal Science and Technology Project 53H21034, and the China Scholarship Council (File No.202008120045)

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 20 December 2022 Accepted: 13 April 2023
Published online: 01 May 2023

References

1. B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144 (2017)
2. S. Waqas Zamir, A. Arora, S. Khan, M. Hayat, F. Shahbaz Khan, M.-H. Yang, L. Shao, Multi-stage progressive image restoration. arXiv e-prints, 2102 (2021)
3. F. Wang, C. Wang, M. Chen, W. Gong, Y. Zhang, S. Han, G. Situ, Far-field super-resolution ghost imaging with a deep neural network constraint. Light Sci. Appl. **11**(1), 1–11 (2022)
4. S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, L. Shao, Learning enriched features for real image restoration and enhancement, in *European Conference on Computer Vision*, pp. 492–511 (2020). Springer
5. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł Kaiser, I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems 30* (2017)

6. H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao, Pre-trained image processing transformer, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12299–12310 (2021)
7. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844 (2021)
8. Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: A general u-shaped transformer for image restoration, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693 (2022)
9. S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739 (2022)
10. A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. S. Mehta, M. Rastegari, Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
12. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations* (2020)
13. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv 2: Inverted residuals and linear bottlenecks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520 (2018)
14. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324 (2019)
15. N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient CNN architecture design, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131 (2018)
16. R. Tahir, K. Cheng, B.A. Memon, Q. Liu, A diverse domain generative adversarial network for style transfer on face photographs (2022)
17. J. Yang, Z. Lin, S. Cohen, Fast image super-resolution based on in-place example regression. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1059–1066 (2013)
18. R. Timofte, V. De Smet, L. Van Gool, : A+: Adjusted anchored neighborhood regression for fast super-resolution, in *Asian Conference on Computer Vision*, pp. 111–126 (2014). Springer
19. W. Luo, Y. Zhang, A. Feizi, Z. Göröcs, A. Ozcan, Pixel super-resolution using wavelength scanning. Light Sci. Appl. **5**(4), 16060 (2016)
20. K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. **33**(12), 2341–2353 (2010)
21. S. Anwar, N. Barnes, Densely residual Laplacian super-resolution. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
22. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301 (2018)
23. M. Intriago-Pazmiño, J. Ibarra-Fiallo, A. Guzmán-Castillo, R. Alonso-Calvo, J. Crespo, Quantitative measures for medical fundus and mammography images enhancement (2022)
24. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer
25. O. Kupyn, T. Martyniuk, J. Wu, Z. Wang, Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8878–8887 (2019)
26. P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models. arxiv 2019. arXiv preprint arXiv:1906.05909
27. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European Conference on Computer Vision*, pp. 213–229 (2020) Springer
28. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in *International Conference on Machine Learning*, pp. 10347–10357 (2021) PMLR
29. B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
30. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
31. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022 (2021)
32. J. Zhang, D. Tao, Famed-net: a fast and accurate multi-scale end-to-end dehazing network. IEEE Trans. Image Process. **29**, 72–84 (2019)
33. X. Fu, B. Liang, Y. Huang, X. Ding, J. Paisley, Lightweight pyramid networks for image deraining. IEEE Trans. Neural Netw. Learn. Syst. **31**(6), 1794–1807 (2019)
34. A. Lahiri, S. Bairagya, S. Bera, S. Haldar, P.K. Biswas, Lightweight modules for efficient deep learning based image restoration. IEEE Trans. Circuits Syst. Video Technol. **31**(4), 1395–1410 (2020)
35. D. Song, C. Xu, X. Jia, Y. Chen, C. Xu, Y. Wang, Efficient residual dense block search for image super-resolution, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12007–12014 (2020)
36. H. Shen, Z.-Q. Zhao, W. Liao, W. Tian, D.-S. Huang, Joint operation and attention block search for lightweight image restoration. Pattern Recogn. **132**, 108909 (2022)
37. W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883 (2016)
38. J. Pan, A. Bulat, F. Tan, X. Zhu, L. Dudziak, H. Li, G. Tzimiropoulos, B. Martinez, Edgevits: Competing light-weight CNNs on mobile devices with vision transformers, in *European Conference on Computer Vision*, pp. 294–311 (2022) Springer
39. T. Huang, S. Li, X. Jia, H. Lu, J. Liu, Neighbor2neighbor: Self-supervised denoising from single noisy images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14781–14790 (2021)

40. W. Zou, T. Ye, W. Zheng, Y. Zhang, L. Chen, Y. Wu, Self-calibrated efficient transformer for lightweight super-resolution, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 930–939 (2022)
41. I. Loshchilov, F. Hutter, Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
42. H. Zhang, V. Sindagi, V.M. Patel, Image de-raining using a conditional generative adversarial network. IEEE Trans. Circuits Syst. Video Technol. **30**(11), 3943–3956 (2019)
43. W. Yang, R.T. Tan, J. Feng, J. Liu, Z. Guo, S. Yan, Deep joint rain detection and removal from a single image, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1357–1366 (2017)
44. X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, J. Paisley, Removing rain from single images via a deep detail network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3855–3863 (2017)
45. H. Zhang, V.M. Patel, Density-aware single image de-raining using a multi-stream dense network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 695–704 (2018)
46. X. Fu, J. Huang, X. Ding, Y. Liao, J. Paisley, Clearing the skies: a deep network architecture for single-image rain removal. IEEE Trans. Image Process. **26**(6), 2944–2956 (2017)
47. W. Wei, D. Meng, Q. Zhao, Z. Xu, Y. Wu, Semi-supervised transfer learning for image rain removal, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3877–3886 (2019)
48. R. Yasarla, V.M. Patel, Uncertainty guided multi-scale residual learning-using a cycle spinning CNN for single image de-raining, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8405–8414 (2019)
49. X. Li, J. Wu, Z. Lin, H. Liu, H. Zha, Recurrent squeeze-and-excitation context aggregation net for single image deraining, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 254–269 (2018)
50. D. Ren, W. Zuo, Q. Hu, P. Zhu, D. Meng, Progressive image deraining networks: a better and simpler baseline, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3937–3946 (2019)
51. S. Nah, T. Hyun Kim, K. Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3883–3891 (2017)
52. Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, L. Shao, Human-aware motion deblurring, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5572–5581 (2019)
53. J. Rim, H. Lee, J. Won, S. Cho, Real-world blur dataset for learning and benchmarking deblurring algorithms, in *European Conference on Computer Vision*, pp. 184–201. Springer (2020)
54. L. Xu, S. Zheng, J. Jia, Unnatural l0 sparse representation for natural image deblurring, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1107–1114 (2013)
55. O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, J. Matas, Deblurgan: Blind motion deblurring using conditional adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8183–8192 (2018)
56. J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R.W. Lau, M.-H. Yang, Dynamic scene deblurring using spatially variant recurrent neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2521–2529 (2018)
57. X. Tao, H. Gao, X. Shen, J. Wang, J. Jia, Scale-recurrent network for deep image deblurring, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8174–8182 (2018)
58. H. Gao, X. Tao, X. Shen, J. Jia, Dynamic scene deblurring with parameter selective sharing and nested skip connections, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3848–3856 (2019)
59. Y. Li, R.T. Tan, X. Guo, J. Lu, M.S. Brown, Rain streak removal using layer priors, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2736–2744 (2016)
60. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)

Publisher's Note