Efficient and Explicit Modelling of Image Hierarchies for Image Restoration

Yawei Li[1] Yuchen Fan[2] Xiaoyu Xiang[2] Denis Demandolx[2]
Rakesh Ranjan[2] Radu Timofte[1,3] Luc Van Gool[1,4]
[1]Computer Vision Lab, ETH Zürich [2]Meta Reality Labs [3]University of Würzburg [4]KU Leuven

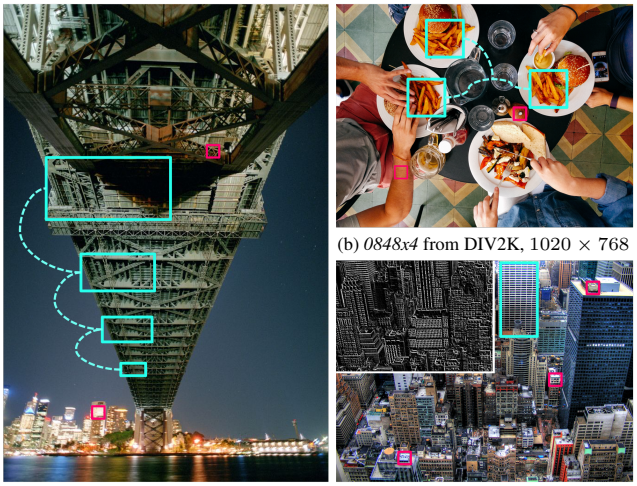
Abstract

The aim of this paper is to propose a mechanism to efficiently and explicitly model image hierarchies in the global, regional, and local range for image restoration. To achieve that, we start by analyzing two important properties of natural images including cross-scale similarity and anisotropic image features. Inspired by that, we propose the anchored stripe self-attention which achieves a good balance between the space and time complexity of self-attention and the modelling capacity beyond the regional range. Then we propose a new network architecture dubbed GRL to explicitly model image hierarchies in the Global, Regional, and Local range via anchored stripe self-attention, window self-attention, and channel attention enhanced convolution. Finally, the proposed network is applied to 7 image restoration types, covering both real and synthetic settings. The proposed method sets the new state-of-the-art for several of those. Code will be available at https://github.com/ofsoundof/GRL-Image-Restoration.git.

1. Introduction

Image restoration aims at recovering high-quality images from low-quality ones, resulting from an image degradation processes such as blurring, sub-sampling, noise corruption, and JPEG compression. Image restoration is an ill-posed inverse problem since important content information about the image is missing during the image degradation processes. Thus, in order to recover a high-quality image, the rich information exhibited in the degraded image should be fully exploited.

Natural images contain a hierarchy of features at global, regional, and local ranges which could be used by deep neural networks for image restoration. *First*, the local range covers a span of several pixels and typical features are edges and local colors. To model such local features, convolutional neural networks (CNNs) with small kernels (*e.g.* 3×3) are utilized. *Second*, the regional range is characterized by a window with tens of pixels. This range of pixels can cover small objects and components of large objects (pink squares in Fig. 1). Due to the larger range, modelling the regional features (consistency, similarity) explicitly with large-kernel CNNs would be inefficient in both parameters and computation. Instead, transformers with a window attention mechanism are well suited for this task. *Third*, beyond local and regional, some features have a global span (cyan rectangles in Fig. 1), incl. but not limited to symmetry, multi-scale pattern repetition (Fig. 1a), same scale texture similarity (Fig. 1b), and structural similarity and consistency in large objects and content (Fig. 1c). To model features at this range, global image understanding is needed.



(a) *bridge* from ICB, 2749×4049 (b) *0848x4* from DIV2K, 1020×768 (c) *073* from Urban100, 1024×765

Figure 1. Natural images show a hierarchy of features in a global, regional, and local range. The local (edges, colors) and regional features (the pink squares) could be well modelled by CNNs and window self-attention. By contrast, it is difficult to efficiently and explicitly model the rich global features (cyan rectangles).

Different from the local and regional range features, there are two major challenges to model the global range features. Firstly, existing image restoration networks based on convolutions and window attention could not capture long-range dependencies explicitly by using a single computational module. Although non-local operations are used in some works, they are either used sparsely in the network or applied to small image crops. Thus, global image under-

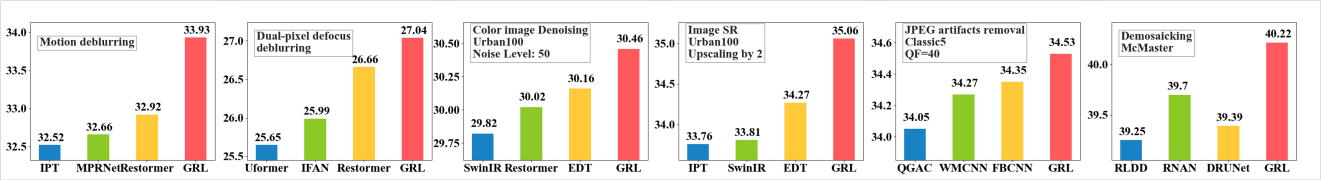


Figure 2. The proposed GRL achieves state-of-the-art performances on various image restoration tasks. Details provided in Sec. 5.

standing still mainly happens via progressive propagation of features through repeated computational modules. Secondly, the increasing resolution of today's images poses a challenge for long-range dependency modelling. High image resolution leads to a computational burden associated with pairwise pixel comparisons and similarity searches.

The aforementioned discussion leads to a series of research questions: 1) how to efficiently model global range features in high-dimensional images for image restoration; 2) how to model image hierarchies (local, regional, global) explicitly by a single computational module for high-dimensional image restoration; 3) and how can this joint modelling lead to a uniform performance improvement for different image restoration tasks. The paper tries to answer these questions in Sec. 3, Sec. 4, and Sec. 5, resp.

First, we propose anchored stripe self-attention for efficient dependency modelling beyond the regional range. The proposed self-attention is inspired by two properties of natural images including cross-scale similarity and anisotropic image features. Cross-scale similarity means that structures in a natural image are replicated at different scales. Inspired by that, we propose to use anchors as an intermediate to approximate the exact attention map between queries and keys in self-attention. Since the anchors summarize image information into a lower-dimensional space, the space and time complexity of self-attention can be significantly reduced. In addition, based on the observation of anisotropic image features, we propose to conduct anchored self-attention within vertical and horizontal stripes. Due to the anisotropic shrinkage of the attention range, a further reduction of complexity is achieved. And the combination of axial stripes also ensures a global view of the image content. When equipped with the stripe shift operation, the four stripe self-attention modes (horizontal, vertical, shifted horizontal, shifted vertical) achieves a good balance between computational complexity and the capacity of global range dependency modelling. Furthermore, the proposed anchored stripe self-attention is analyzed from the perspective of low-rankness and similarity propagation.

Secondly, a new transformer network is proposed to explicitly model global, regional, and local range dependencies in a single computational module. The hierarchical modelling of images is achieved by the parallel computation of the proposed anchored stripe self-attention, window self-attention, and channel-attention enhanced convolution. And the transformer architecture is dubbed **GRL**.

Thirdly, the proposed GRL transformer is applied to various image restoration tasks. Those tasks could be classified into three settings based on the availability of data including real image restoration, synthetic image restoration, and data synthesis based real image restoration. In total, seven tasks are explored for the proposed network including image super-resolution, image denoising, JPEG compression artifacts removal, demosaicking, real image super-resolution, single image motion deblurring, and defocus deblurring. As shown in Fig. 2, the proposed network shows promising results on the investigated tasks.

2. Related Works

Convolution for local range modelling. One of the basic assumptions for example and learning-based image restoration is that repetitive patterns could exist in either the same or different images [17] and that the redundant information they carry could help to restore the local patches. Thus, it helps if repetitive patterns could be detected and modelled [13, 33, 44, 61]. This intuition matches the computational procedure of convolution well, which slides the kernel across the image and detects local patterns similar to the learnable kernels. By stacking multiple convolutional layers, the receptive field of a CNN gets progressively enlarged and rich image features are captured. Since the advent of deep learning, great efforts have been made to design CNNs for image restoration [26, 39, 71, 72, 86].

Non-local and global priors. Besides the local features, it is also important to model the non-local and global image priors. The early work of non-local means serves this purpose, which computes an output pixel as the weighted sum of all the pixels within the image [4]. Inspired by that, later works have been developed to utilize the repetitive patterns in a non-local range for image denoising [10] and super-resolution [22]. Apart from the traditional methods, non-local operations are also introduced into deep neural networks for video classification [70] and image SR [45, 85].

Besides the non-local operations, self-attention has been developed to model the global range dependencies [12, 68]. However, the computational complexity of global self-attention grows quadratically with the number of tokens. Thus, the increase in efficiency of global self-attention is investigated by several works [7, 9, 31, 35, 69].

Regional self-attention. Among the methods for accelerating transformers, regional self-attention appears to

be promising. The idea is proposed in the pioneering works [54, 56] and improved as shifted window attention [47, 48]. Inspired by the success of shifted window attention for visual recognition and perception, this method is also used for image restoration [6, 42, 43]. Despite the good performance of the window attention mechanism, it is pointed out in recent works that a wider range of pixel involvement could lead to better image restoration [6, 23]. Thus, in this paper, we try to propose a method that efficiently brings the modelling capacity of self-attention beyond the regional range.

3. Motivation

3.1. Self-attention for dependency modelling

Self-attention is good at modelling long-range dependencies explicitly and it facilitates the propagation of information across the modelled dependencies. This operation allows a token to be compared with all the other tokens. The output token is computed as a weighted sum of all the tokens based on a similarity comparison, *i.e.* ,

$$\mathbf{Y} = \text{Softmax}\left(\mathbf{Q}\cdot\mathbf{K}^T/\sqrt{d}\right)\cdot\mathbf{V}, \tag{1}$$

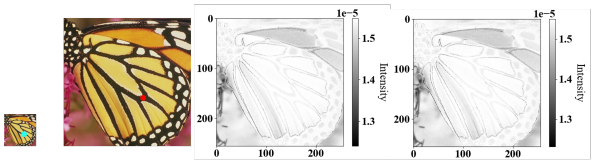
where $\mathbf{Q} = \mathbf{X}\cdot\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\cdot\mathbf{W}_K$, $\mathbf{V} = \mathbf{X}\cdot\mathbf{W}_V$, $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$, and $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N\times d}$. N and d denote the number of tokens and the dimension of one token, respectively. Additionally, $\mathbf{M}$ denotes the attention map, *i.e.* $\mathbf{M} = \text{Softmax}(\mathbf{Q}\cdot\mathbf{K}^T/\sqrt{d})$.

The time complexity of self-attention is $\mathcal{O}(N^2d)$ and the space complexity is dominated by the term $\mathcal{O}(N^2)$ of the attention map $\mathbf{M}$. The computational complexity and memory footprint of self-attention grow quadratically with the number of tokens. Thus, self-attention can easily become a computation bottleneck for images where the number of tokens is the multiplication of the two dimensions of the feature map. To overcome this problem, it is proposed to apply self-attention within a window. In this way, the number of tokens that participate in self-attention is significantly reduced and the computational burden is also lifted.

The problem of window self-attention is that the modelling capacity of the operation is limited to a regional range due to the small window size (8×8 [43]). On the other hand, it is shown in recent works [6, 23] that even a slight increase in window size can lead to better image restoration. Thus, it can be conjectured that modelling dependencies beyond the regional range is still important for image restoration. Hence, it remained to be investigated how to maintain the ability for long-range dependency modelling under a controlled computational budget.

3.2. Motivation I: cross-scale similarity

The attention map $\mathbf{M}$ plays an essential role in self-attention as it captures the similarity between every paired pixels in the image. Thus, improving the efficiency of the self-attention in Eq. (1) needs one to analyze the property of the attention map. And we are inspired by a property of images, *i.e.* cross-scale similarity. That is, the basic structure such as lines and edges of an image is kept in the different versions of the image with different scaling factors. In Fig. 3, the attention map between pixels in an image is shown. Particularly, the attention map between a pixel and the whole image is visualized as a gray-scale heat map. As shown, no matter whether the pixel comes from the high-resolution image or the down-scaled version, the heat map between the pixel and the high-resolution image shows the basic structure of the image. And the heat maps in Fig. 3(c) and Fig. 3(d) are very similar to each other.



(a) LR image (b) HR image (c) Attention between the cyan and other pixels in (b). (d) Attention between the red pixel and other pixels in (b).

Figure 3. Cross-scale similarity. (c) and (d) shows the attention map between the selected pixels and the example high-resolution image. Although the cyan pixel in (a) and the red pixel in (b) are from images with different resolutions, their attention map with respect to the high-resolution image shows very similar structures.

Anchored self-attention. Inspired by the cross-scale similarity shown in Fig. 3, we try to reduce the complexity of the global self-attention in Eq. (1) by operating on images with different resolutions and manipulating the number of tokens, *i.e.* the N^2 term in $\mathcal{O}(N^2d)$. To achieve that, we introduce an additional concept named anchors besides the triplets of queries, keys, and values. The set of anchors is a summary of the information in the image feature map and has a lower dimensionality. Instead of conducting the similarity comparison between the queries and keys directly, the anchors act as an intermediate for the similarity comparison. Formally, the anchored self-attention is proposed as in the following equation

$$\mathbf{Y} = \mathbf{M}_e\cdot\mathbf{Z} = \mathbf{M}_e\cdot(\mathbf{M}_d\cdot\mathbf{V}), \tag{2}$$

$$\mathbf{M}_d = \text{Softmax}(\mathbf{A}\cdot\mathbf{K}^T/\sqrt{d}), \tag{3}$$

$$\mathbf{M}_e = \text{Softmax}(\mathbf{Q}\cdot\mathbf{A}^T/\sqrt{d}), \tag{4}$$

where $M \ll N$, $\mathbf{A} \in \mathbb{R}^{M\times d}$ is the anchor, $\mathbf{M}_e \in \mathcal{R}^{N\times M}$ and $\mathbf{M}_d \in \mathcal{R}^{M\times N}$ denotes the attention map between the query-anchor pair and anchor-key pair. The choice of the operations to derive the anchors is investigated in the implementation details of the ablation study of the paper.

Since the number of anchors is much smaller than the number of the other tokens, the size of the resulting two attention maps $\mathbf{M}_e$ and $\mathbf{M}_d$ are much smaller than the orig-

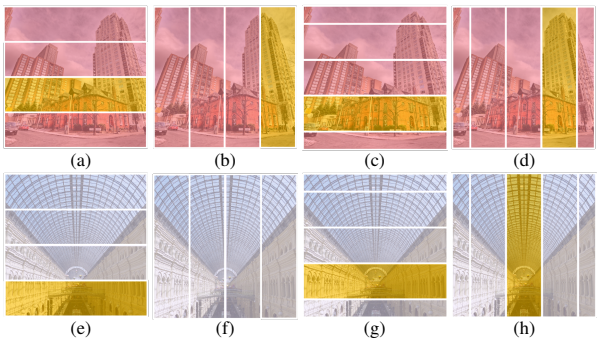


Figure 4. The image features in natural images are anisotropic. Thus, it is not always necessary to employ the uniform global range attention in all parts of the image.

inal attention map $\mathbf{M}$ in Eq. (1). Then the matrix multiplication in Eq. (2) is computed from the right hand. The self-attention is first done for the anchors and keys. The attention map $\mathbf{M}_d$ distills the tokens $\mathbf{V}$ into an intermediate feature $\mathbf{Z}$. Then the self-attention is done between the queries and the anchors. The second attention map $\mathbf{M}_e$ expands the size of the feature $\mathbf{Z}$ and recovers the information in $\mathbf{V}$. The computational complexity of the anchored self-attention is reduced to $\mathcal{O}(NMd)$. And the space complexity is reduced to $\mathcal{O}(NM)$.

3.3. Motivation II: anisotropic image features

The anchored self-attention could reduce the space and time complexity of the self-attention in Eq. (1) significantly by removing the quadratic term N^2. Yet, for image restoration tasks, the remaining term N is the multiplication of the width and height of the image. Thus, the complexity of the anchored self-attention in Eq. (2) could still be unaffordable due to the large term N. Thus, it is desirable to further reduce the complexity of the anchored self-attention.

To achieve that goal, we resort to another characteristic of natural images, *i.e.* , the anisotropic image features. As shown in Fig. 4, the natural image features such as the single object in Fig. 4(c)&(d), the multi-scale similarity in Fig. 4(h), symmetry in Fig. 4(e)&(g) span in an anisotropic manner. Thus, isotropic global range attention across the entire image is redundant to capture the anisotropic image features. And in response to that, we propose to conduct attention within the anisotropic stripes shown in Fig. 4.

Stripe attention mechanism. The proposed stripe attention mechanism consists of four modes including the horizontal stripe, the vertical stripe, the shifted horizontal stripe, and the shifted vertical stripe. The horizontal and vertical stripe attention mechanisms could be employed alternately across a transformer network. In this way, a trade-off is made between maintaining the global range modelling capacity and controlling the computation complexity of global self-attention. Thus, in combination with the concept of anchors, we propose the **anchored stripe self-attention**.

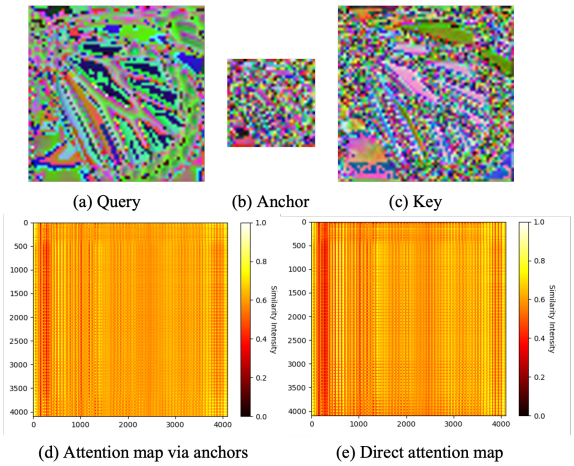


Figure 5. The visualization of the (a) queries, (b) anchors, and (c) keys from the different layers of the proposed network. (d) shows the attention map approximated by Eq. (2), *i.e.* $\mathbf{M}_e \cdot \mathbf{M}_d$. (e) shows the exact attention map $\mathbf{M}$ computed in Eq. (1).

For this attention mechanism, efficient self-attention is conducted inside the vertical and horizontal stripes with the help of the introduced anchors.

3.4. Discussion

The proposed anchored stripe self-attention mechanism is closely related to two other concepts including low-rankness and similarity propagation. And we detail the relationship in this subsection as follows.

Low-rankness of attention map. By comparing the self-attention mechanisms in Eq. (1) and Eq. (2), we can easily found out that the original attention map $\mathbf{M}$ is decomposed into small attention maps $\mathbf{M}_d$ and $\mathbf{M}_e$ whose rank is no larger than M. And the essence here is to provide the low-rank approximation without calculating the original attention map first. For the success of the anchored self-attention, it is important to ensure that with the anchors as the intermediate, the approximated attention map is similar to the original attention map. Thus, an additional analysis is provided in Fig. 5.

First, by observing the queries, anchors, and keys, we can conclude that the anchors have a very similar structure to the query and key. Thus, the anchors are a good summary of the information in the queries and keys. And approximating self-attention with anchors as intermediate seems to be plausible. Additionally, the approximate attention map $\mathbf{M}_e \cdot \mathbf{M}_d$ and the exact attention map $\mathbf{M}$ are also compared in Fig. 5. As shown, the approximate attention map keeps the major structure in the exact attention map, which is confirmed by the large Pearson correlation coefficients (0.9505) between the two attention maps. Thus, the quality of the anchored self-attention is guaranteed.

Metric and similarity propagation. From another perspective, in the proposed anchored self-attention, the queries and keys are first compared with the anchors and

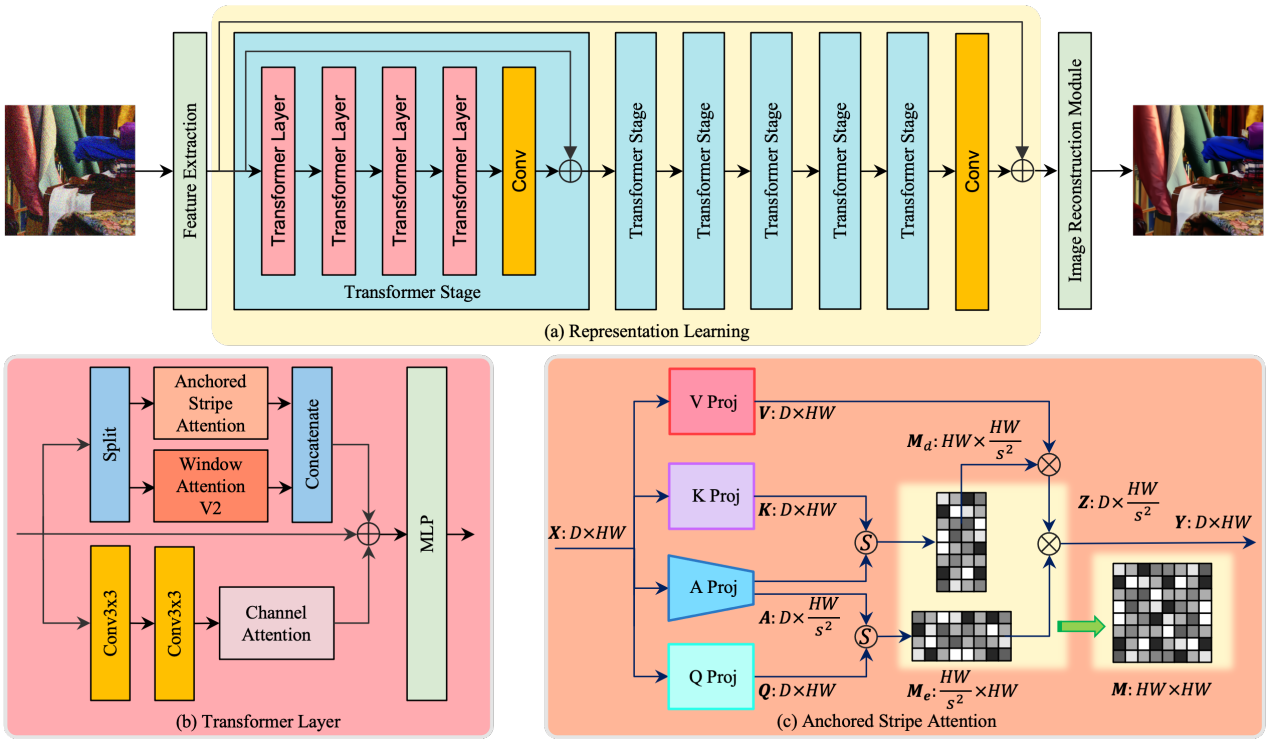


Figure 6. Network architecture. (a) The representation learning module contains stages of transformer layers. (b) The transformer layer is equipped with global, regional, and local modelling blocks. (c) The anchored stripe attention helps to attend beyond regional ranges.

then the query-key similarity is computed. Thus, this computation procedure needs to propagate the query-anchor and key-anchor similarity to the query-key pair. And similarity propagation is related to the triangle inequality in a metric space [19, 27, 73]. A mathematical metric needs to satisfy several conditions including the essential triangle inequality, $d(\mathbf{q}, \mathbf{k}) \leq d(\mathbf{a}, \mathbf{q}) + d(\mathbf{a}, \mathbf{k})$, where $d(\cdot, \cdot)$ defines a metric between two entities. Thus, the $\mathbf{q}$ / $\mathbf{k}$ distance is upper-bounded by the sum of the $\mathbf{a}$ / $\mathbf{q}$ distance and the $\mathbf{a}$ / $\mathbf{k}$ distance. This implies that if $\mathbf{a}$ is similar (close) to both $\mathbf{q}$ and $\mathbf{k}$, then $\mathbf{q}$ and $\mathbf{k}$ should also be similar (close) to each other. Yet, the similarity measure in Eq. (1) and Eq. (2) is defined by the dot product instead of the distance between tokens, which does not satisfy the triangle inequality. Thus, similarity propagation could not be theoretically guaranteed. To study the influence of the similarity measure, an ablation study is conducted and the results are shown in Sec. 5. Dot product and distance are compared as a similarity measure. According to the results, although the dot product does not strictly obey the triangle inequality, it still guarantees better image restoration results. Thus, we can conclude empirically that the dot product is enough for similarity propagation.

4. Modelling Image Hierarchies

In this section, we answer the second research question described in the introduction, that is, how to explicitly model image hierarchies by a single computational module. In response to that, we propose the GRL network architecture that incorporates global range, regional range, and local range image modelling capacities.

Network architecture. The overall architecture of the proposed network is shown in Fig. 6. The network takes a degraded low-quality image as input, processes the image inside the network, and outputs a recovered high-quality image. In detail, the network contains three parts. 1) The feature extraction layer is implemented as a simple convolution and converts the input image into feature maps. 2) The representation learning component enriches the information extracted in the previous operation. The transformer stage consists of several transformer layers and ends with a convolution layer. The dimension of the feature map is maintained across the whole representation learning module. Skip connection is applied to both the transformer stage and the representation learning module. 3) The image reconstruction module takes the rich features calculated by the previous operations and estimates a recovered image.

Transformer Layer. This layer in Fig. 6b is the key component that provides the hierarchical image modelling capacity in the global, regional, and local range. This layer first processes the input feature map by the parallel self-attention module and channel attention enhanced convolutions. The convolution branch serves to capture local structures in the input feature map. On the other hand, the self-attention module contains the window attention proposed in Swin transformer V2 [47] and the anchored stripe atten-

Table 1. ***Defocus deblurring*** results. **S:** single-image defocus deblurring. **D:** dual-pixel defocus deblurring.

Method	Indoor Scenes				Outdoor Scenes				Combined			
	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓
EBDB$_S$ [30]	25.77	0.772	0.040	0.297	21.25	0.599	0.058	0.373	23.45	0.683	0.049	0.336
DMENet$_S$ [40]	25.50	0.788	0.038	0.298	21.43	0.644	0.063	0.397	23.41	0.714	0.051	0.349
JNB$_S$ [60]	26.73	0.828	0.031	0.273	21.10	0.608	0.064	0.355	23.84	0.715	0.048	0.315
DPDNet$_S$ [1]	26.54	0.816	0.031	0.239	22.25	0.682	0.056	0.313	24.34	0.747	0.044	0.277
KPAC$_S$ [62]	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227
IFAN$_S$ [41]	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217
Restormer$_S$ [76]	28.87	0.882	0.025	0.145	23.24	0.743	0.050	0.209	25.98	0.811	0.038	0.178
GRL$_S$-B	29.06	0.886	0.024	0.139	23.45	0.761	0.049	0.196	26.18	0.822	0.037	0.168
DPDNet$_D$ [1]	27.48	0.849	0.029	0.189	22.90	0.726	0.052	0.255	25.13	0.786	0.041	0.223
RDPD$_D$ [2]	28.10	0.843	0.027	0.210	22.82	0.704	0.053	0.298	25.39	0.772	0.040	0.255
Uformer$_D$ [74]	28.23	0.860	0.026	0.199	23.10	0.728	0.051	0.285	25.65	0.795	0.039	0.243
IFAN$_D$ [41]	28.66	0.868	0.025	0.172	23.46	0.743	0.049	0.240	25.99	0.804	0.037	0.207
Restormer$_D$ [76]	29.48	0.895	0.023	0.134	23.97	0.773	0.047	0.175	26.66	0.833	0.035	0.155
GRL$_D$-B	29.83	0.903	0.022	0.114	24.39	0.795	0.045	0.150	27.04	0.847	0.034	0.133

Table 2. ***Single-image motion deblurring*** results. GoPro dataset [51] is used for training.

Method	**GoPro** [51] PSNR↑ / SSIM↑	**HIDE** [59] PSNR↑ / SSIM↑	Average PSNR↑ / SSIM↑
DeblurGAN [37]	28.70 / 0.858	24.51 / 0.871	26.61 / 0.865
Nah *et al.* [51]	29.08 / 0.914	25.73 / 0.874	27.41 / 0.894
DeblurGAN-v2 [38]	29.55 / 0.934	26.61 / 0.875	28.08 / 0.905
SRN [64]	30.26 / 0.934	28.36 / 0.915	29.31 / 0.925
Gao *et al.* [20]	30.90 / 0.935	29.11 / 0.913	30.01 / 0.924
DBGAN [81]	31.10 / 0.942	28.94 / 0.915	30.02 / 0.929
MT-RNN [53]	31.15 / 0.945	29.15 / 0.918	30.15 / 0.932
DMPHN [79]	31.20 / 0.940	29.09 / 0.924	30.15 / 0.932
Suin *et al.* [63]	31.85 / 0.948	29.98 / 0.930	30.92 / 0.939
SPAIR [55]	32.06 / 0.953	30.29 / 0.931	31.18 / 0.942
MIMO-UNet+ [8]	32.45 / 0.957	29.99 / 0.930	31.22 / 0.944
IPT [5]	32.52 / -	- / -	- / -
MPRNet [77]	32.66 / 0.959	30.96 / 0.939	31.81 / 0.949
Restormer [76]	32.92 / 0.961	31.22 / 0.942	32.07 / 0.952
GRL-B (ours)	33.93 / 0.968	31.65 / 0.947	32.79 / 0.958

Table 3. ***Single-image motion deblurring*** results on RealBlur [57] dataset. The networks are trained and tested on RealBlur dataset.

Method	**RealBlur-R** [57] PSNR↑ / SSIM↑	**RealBlur-J** [57] PSNR↑ / SSIM↑	Average PSNR↑ / SSIM↑
DeblurGAN-v2 [38]	36.44 / 0.935	29.69 / 0.870	33.07 / 0.903
SRN [64]	38.65 / 0.965	31.38 / 0.909	35.02 / 0.937
MPRNet [77]	39.31 / 0.972	31.76 / 0.922	35.54 / 0.947
MIMO-UNet+ [8]	- / -	32.05 / 0.921	- / -
MAXIM-3S [67]	39.45 / 0.962	32.84 / 0.935	36.15 / 0.949
BANet [66]	39.55 / 0.971	32.00 / 0.923	35.78 / 0.947
MSSNet [34]	39.76 / 0.972	32.10 / 0.928	35.93 / 0.950
Stripformer [65]	39.84 / 0.974	32.48 / 0.929	36.16 / 0.952
GRL-B (ours)	40.20 / 0.974	32.82 / 0.932	36.51 / 0.953

tion proposed in this paper. The feature map is split equally along the channel dimension and concatenated along the channel dimension again after the parallel processing within the two attention modules. The window attention provides the mechanism to capture the regional range dependencies. Then the feature maps outputted by the convolution module and the attention module are added to the input feature map, which is processed by the following MLP module.

Anchored stripe self-attention. The operation of the proposed anchored stripe attention is conducted according to Eq. (2) and visualized in Fig. 6c. The dimension of different features is also shown. The triplet of $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ is derived by plain linear projections. To summarize the information into anchors, the anchor projection is implemented as an average pooling layer followed by a linear projection. After the anchor projection, the resolution of the image feature map is down-scaled by a factor of s along both directions. As shown in Fig. 6, the two attention maps $\mathbf{M}_d$ and $\mathbf{M}_e$ play a similar role as the original attention map $\mathbf{M}$ but with less space and time complexity.

5. Experimental Results

The experimental results are shown in this section. We answer the third research question raised in the introduction by investigating the performance of the proposed network on different image restoration tasks. Based on the data type, the investigated tasks are classified into three commonly used settings including 1) real image restoration (single-image motion deblurring, defocus deblurring), 2) image restoration based on synthetic data (image denoising, single image SR, JPEG compression artifact removal, demosaicking), and 3) real image restoration based on data synthesis. We provide three networks with different model sizes including the tiny, small, and base versions (GRL-T, GRL-S, GRL-B). For real and synthetic image restoration, Adam optimizer and L_1 loss are used to train the network with an initial learning rate 2×10^{-4}. More details about the training dataset, training protocols, and additional visual results are shown in the *supplementary material*.

5.1. Image deblurring

We first investigate the performance of the proposed network on two real image restoration tasks including single-image motion deblurring, and motion deblurring.

Single image motion deblurring. Tab. 2 and Tab. 3 shows the experimental results for single image motion deblurring on synthetic datasets (GoPro [51], HIDE [59]) and real dataset (RealBlur-R [57]), respectively. Compared with the previous state-of-the-art Restormer [76], the proposed GRL achieves significant PSNR improvement of 1.01 dB on the GoPro dataset. On the HIDE dataset, the PSNR improvement is 0.43 dB. Please note that the improvement is achieved under fewer parameter budget. As shown in Tab. 4,

Table 4. ***Color and grayscale image denoising*** results. Model complexity and prediction accuracy are shown for better comparison.

Method	# Params [M]	Color												Grayscale								
		CBSD68 [49]			Kodak24 [16]			McMaster [83]			Urban100 [28]			Set12 [82]			BSD68 [49]			Urban100 [28]		
		σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
DnCNN [32]	0.56	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59	32.86	30.44	27.18	31.73	29.23	26.23	32.64	29.95	26.26
RNAN [85]	8.96	-	-	28.27	-	-	29.58	-	-	29.72	-	-	29.08	-	-	27.70	-	-	26.48	-	-	27.65
IPT [5]	115.33	-	-	28.39	-	-	29.64	-	-	29.98	-	-	29.71	-	-	-	-	-	-	-	-	-
EDT-B [42]	11.48	34.39	31.76	28.56	35.37	32.94	29.87	35.61	33.34	30.25	35.22	33.07	30.16	-	-	-	-	-	-	-	-	-
DRUNet [80]	32.64	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61	33.25	30.94	27.90	31.91	29.48	26.59	33.44	31.11	27.96
SwinIR [43]	11.75	34.42	31.78	28.56	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82	33.36	31.01	27.91	31.97	29.50	26.58	33.70	31.30	27.98
Restormer [76]	26.13	34.40	31.79	28.60	35.47	33.04	30.01	35.61	33.34	30.30	35.13	32.96	30.02	33.42	31.08	28.00	31.96	29.52	26.62	33.79	31.46	28.29
GRL-T	0.88	34.30	31.66	28.45	35.24	32.78	29.67	35.49	33.18	30.06	35.08	32.84	29.78	33.29	30.92	27.78	31.90	29.43	26.49	33.66	31.23	27.89
GRL-S	3.12	34.36	31.72	28.51	35.32	32.88	29.77	35.59	33.29	30.18	35.24	33.07	30.09	33.36	31.02	27.91	31.93	29.47	26.54	33.84	31.49	28.24
GRL-B	19.81	34.45	31.82	28.62	35.43	33.02	29.93	35.73	33.46	30.36	35.54	33.35	30.46	33.47	31.12	28.03	32.00	29.54	26.60	34.09	31.80	28.59

Table 5. ***Classical image SR*** results. Results of both lightweight models and accurate models are summarized.

Method	Scale	# Params [M]	Set5 [3]		Set14 [78]		BSD100 [49]		Urban100 [28]		Manga109 [50]	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
RCAN [84]	×2	15.44	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [11]	×2	15.71	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
HAN [52]	×2	63.61	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
IPT [5]	×2	115.48	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [43]	×2	0.88	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
SwinIR [43]	×2	11.75	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
EDT [42]	×2	0.92	38.23	0.9615	33.99	0.9209	32.37	0.9021	32.98	0.9362	39.45	0.9789
EDT [42]	×2	11.48	38.63	0.9632	34.80	0.9273	32.62	0.9052	34.27	0.9456	40.37	0.9811
GRL-T (ours)	×2	0.89	38.27	0.9627	34.21	0.9258	32.42	0.9056	33.60	0.9411	39.61	0.9790
GRL-S (ours)	×2	3.34	38.37	0.9632	34.64	0.9280	32.52	0.9069	34.36	0.9463	39.84	0.9793
GRL-B (ours)	×2	20.05	38.67	0.9647	35.08	0.9303	32.67	0.9087	35.06	0.9505	40.67	0.9818
RCAN [84]	×4	15.59	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [11]	×4	15.86	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
HAN [52]	×4	64.20	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IPT [5]	×4	115.63	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [43]	×4	0.90	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
SwinIR [43]	×4	11.90	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
EDT [42]	×4	0.92	32.53	0.8991	28.88	0.7882	27.76	0.7433	26.71	0.8051	31.35	0.9180
EDT [42]	×4	11.63	33.06	0.9055	29.23	0.7971	27.99	0.7510	27.75	0.8317	32.39	0.9283
GRL-T (ours)	×4	0.91	32.56	0.9029	28.93	0.7961	27.77	0.7523	27.15	0.8185	31.57	0.9219
GRL-S (ours)	×4	3.49	32.76	0.9058	29.10	0.8007	27.90	0.7568	27.90	0.8357	32.11	0.9267
GRL-B (ours)	×4	20.20	33.10	0.9094	29.37	0.8058	28.01	0.7611	28.53	0.8504	32.77	0.9325

GRL-B saves 24% parameters compared with Restormer. As shown in Tab. 3, GRL-B sets the new state-of-the-art performance of 40.20 PSNR on RealBlur-R dataset.

Defocus deblurring. Tab. 1 shows the experimental results for defocus deblurring using single image and dual-pixel images. Our GRL outperforms the previous methods for all three scene types. Compared with Restormer on the combined scenes, our GRL achieves an elegant performance boost of 0.20 dB and 0.38 dB for single and dual-pixel defocus deblurring. Compared with Uformer [74] and IFAN [41], GRL achieves PSNR gain of 1.39 dB and 1.05 dB for the dual-pixel setting.

5.2. Image restoration based on synthetic data

Investigating image restoration with synthetic data is also valuable to reveal the network capacity of restoration methods. Besides the experiments on the real data, we also study the performance of the network on synthetic data.

Image denoising. First, the experimental results on Gaussian image denoising are shown in Tab. 4. For a fair comparison between different models, both the network complexity and accuracy are shown in the table. And several key findings are observed. **I.** The tiny version GRL-T is extremely efficient, reducing model complexity by two orders of magnitude (only 0.76% of [5] and 2.7% of DRUNet [80]) while not sacrificing network accuracy. **II.** The small version GRL-S performs competitive with the previous state-of-the-art SwinIR [43] and Restormer [76]. **II.** On Urban100, the base version outperforms Restormer by a large margin (*e.g.* 0.44dB PSNR gain for color image and noise level 50).

Image SR. Experimental results for classical images are shown in Tab. 5. Both lightweight models and accurate SR models are summarized. A similar conclusion could be drawn from the results. **I.** Among the lightweight networks, GRL-T outperforms both convolution and self-attention-based networks including DBPN [25], SwinIR [43] and EDT [42]. Compared with EDT, Significant improvements are obtained on Urban100 and Manga109 datasets (0.44 dB and 0.22 dB for $\times 4$ SR). **II.** GRL-B sets the new state-of-the-art for accurate image SR. **III.** GRL-S achieves a good balance between network complexity and SR image quality.

JPEG compression artifact removal. The experimental results for color and grayscale images are shown in Tab. 6. Four image quality factors ranging from 10 to 40 for JPEG

Table 6. ***Grayscale image JPEG compression artifact removal*** results. As a comparison metric, the parameter count of FBCNN [29] GRL-S are 71.92M and 3.12M.

Set	QF	JPEG		DnCNN [82]		DCSC [18]		QGAC [14]		MWCNN [46]		FBCNN [29]		GRL-S	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Classic5 [15]	10	27.82	0.760	29.40	0.803	29.62	0.810	29.84	0.812	30.01	0.820	30.12	0.822	30.20	0.829
	20	30.12	0.834	31.63	0.861	31.81	0.864	31.98	0.869	32.16	0.870	32.31	0.872	32.49	0.878
	30	31.48	0.867	32.91	0.886	33.06	0.888	33.22	0.892	33.43	0.893	33.54	0.894	33.72	0.899
	40	32.43	0.885	33.77	0.900	33.87	0.902	34.05	0.905	34.27	0.906	34.35	0.907	34.53	0.911
BSD500 [49]	10	27.80	0.768	29.21	0.809	29.32	0.813	29.46	0.821	29.61	0.820	29.67	0.821	29.74	0.823
	20	30.05	0.849	31.53	0.878	31.63	0.880	31.73	0.884	31.92	0.885	32.00	0.885	32.05	0.885
	30	31.37	0.884	32.90	0.907	32.99	0.908	33.07	0.912	33.30	0.912	33.37	0.913	33.43	0.912
	40	32.30	0.903	33.85	0.923	33.92	0.924	34.01	0.927	34.27	0.928	34.33	0.928	34.38	0.928

Table 7. ***Color image JPEG compression artifact removal*** *results.*

Set	QF	JPEG		QGAC [14]		FBCNN [29]		GRL-S	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LIVE1 [58]	10	25.69	0.743	27.62	0.804	27.77	0.803	28.13	0.814
	20	28.06	0.826	29.88	0.868	30.11	0.868	30.49	0.878
	30	29.37	0.861	31.17	0.896	31.43	0.897	31.85	0.905
	40	30.28	0.882	32.05	0.912	32.34	0.913	32.79	0.920
BSD500 [49]	10	25.84	0.741	27.74	0.802	27.85	0.799	28.26	0.808
	20	28.21	0.827	30.01	0.869	30.14	0.867	30.57	0.875
	30	29.57	0.865	31.33	0.898	31.45	0.897	31.92	0.903
	40	30.52	0.887	32.25	0.915	32.36	0.913	32.86	0.919

Table 8. ***Image demosaicking*** results.

Datasets	Matlab	DDR [75]	DeepJoint [21]	MMNet [36]	RLDD [24]	RNAN [85]	DRUNet [80]	GRL-S (ours)
Kodak [16]	35.78	41.11	42.00	40.19	42.49	43.16	42.68	43.57
McMaster [83]	34.43	37.12	39.14	37.09	39.25	39.70	39.39	40.22

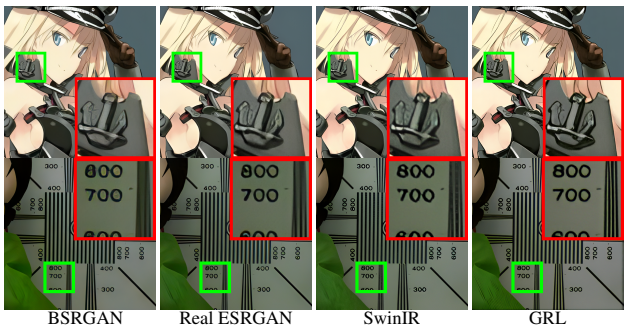


Figure 7. Visual results for real-world image SR.

compression are studied. As shown in the table, the proposed GRL-S network outperforms the previous state-of-the-art method elegantly across different datasets and quality factors. Notably, GRL-S has a much smaller model complexity than FBCNN.

Demosaicking. Results for image demosaicking is shown in Tab. 8. The proposed method outperforms the previous methods RNAN [85] and DRUNet [80] significantly.

5.3. Real image restoration based on data synthesis

Finally, we also investigate the performance of the network for real-world image restoration. The aim is to super-resolve a low-quality image by an upscaling factor of 4. Since there are no ground-truth images for this task, only the visual comparison is given in Fig. 7. Compared with the other methods, the proposed GRL is able to remove more artifacts in the low-resolution images.

5.4. Ablation study

Influence of the similarity comparison method. As mentioned in Sec. 3.4, for theoretical guarantee of similarity propagation, a mathematical metric rather than a dot product should be used. To study the difference between, image restoration with the two operations are compared and the results are shown in Tab. 9. As revealed by the table, the dot product is very competitive compared with a metric and it outperforms a distance metric for a couple of settings. Considering this, the dot product is still used.

Table 9. Ablation study on similarity comparison operation.

Test set	Metric	Color DN			Gray DN			Image SR		
		σ15	σ25	σ50	σ15	σ25	σ50	$\times$2	$\times$3	$\times$4
BSD68 or BSD100	Euclidean	35.02	32.56	29.42	31.84	29.36	26.43	32.30	29.19	27.67
	Dot product	35.10	32.64	29.54	31.85	29.39	26.44	32.33	29.22	27.70
Urban100	Euclidean	34.63	32.28	28.94	33.25	30.64	27.17	32.76	28.62	26.50
	Dot product	34.77	32.41	29.19	33.28	30.75	27.26	32.88	28.78	26.67

Table 10. Ablation study on anchor projection operation.

Anchor projection operation	# Params [m]	PSNR on Set 5
Depthwise Conv	3.17	35.03
Conv	4.19	35.03
Patch merging	3.53	34.98
Maxpool + Linear Projection	3.12	35.02
Avgpool + Linear Projection	3.12	35.03

Influence of the anchor projections. The anchor projection operation helps to summarize the information in the feature map. The ablation study is shown in Tab. 10. Considering both the accuracy performance and parameter budget, Avgpool followed by linear projection is finally used.

6. Conclusion

In this paper, we proposed GRL, a network with efficient and explicit hierarchical modelling capacities for image restoration. The proposed network was mainly inspired by two image properties including cross-scale similarity and anisotropic image features. Based on that, we proposed the efficient anchored stripe self-attention module for long-range dependency modelling. Then a versatile network architecture was proposed for image restoration. The proposed network can model image hierarchies in the global, regional, and local ranges. Owing to the advanced computational mechanism, the proposed network architecture achieves state-of-the-art performances for various image restoration tasks.

Acknowledgements. This work was partly supported by ETH Zürich General Fund (OK), Meta Reality Labs and the Alexander von Humboldt Foundation.

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pages 111–126. Springer, 2020. 6

[2] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S. Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *ICCV*, 2021. 6

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, 2012. 7

[4] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 60–65. IEEE, 2005. 2

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 6, 7

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 3

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 2

[8] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, 2021. 6

[9] 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. 2

[10] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 2

[11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 7

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceeding of the European Conference on Computer Vision*, pages 184–199. Springer, 2014. 2

[14] Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. Quantization guided jpeg artifact correction. In *European Conference on Computer Vision*, pages 293–309. Springer, 2020. 8

[15] Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *IEEE Transactions on Image Processing*, 16(5):1395–1411, 2007. 8

[16] Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 4(2), 1999. 7, 8

[17] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002. 2

[18] Xueyang Fu, Zheng-Jun Zha, Feng Wu, Xinghao Ding, and John Paisley. Jpeg artifacts reduction via deep convolutional sparse coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2501–2510, 2019. 8

[19] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, and Balaji Krishnamurthy. Simpropnet: Improved similarity propagation for few-shot image segmentation. *arXiv preprint arXiv:2004.15014*, 2020. 5

[20] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, 2019. 6

[21] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 8

[22] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *Proceedings of the International Conference on Computer Vision*, pages 349–356. IEEE, 2009. 2

[23] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 3

[24] Yu Guo, Qiyu Jin, Gabriele Facciolo, Tieyong Zeng, and Jean-Michel Morel. Residual learning for effective joint demosaicing-denoising. *arXiv preprint arXiv:2009.06205*, 2020. 8

[25] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2018. 7

[26] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep backprojection networks for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[27] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. Collaborative metric learning. In *Proceedings of the 26th international conference on world wide web*, pages 193–201, 2017. 5

[28] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 7

[29] Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2021. 8

[30] Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *TIP*, 2017. 6

[31] 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. 2

[32] Daisuke Kiku, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Beyond color difference: Residual interpolation for color image demosaicking. *IEEE Transactions on Image Processing*, 25(3):1288–1300, 2016. 7

[33] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 2

[34] Kiyeon Kim, Seungyong Lee, and Sunghyun Cho. MSSNet: Multi-scale-stage network for single image deblurring. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 524–539. Springer, 2022. 6

[35] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 2

[36] Filippos Kokkinos and Stamatios Lefkimmiatis. Iterative joint image demosaicking and denoising using a residual denoising network. *IEEE Transactions on Image Processing*, 28(8):4177–4188, 2019. 8

[37] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. DeblurGAN: Blind motion deblurring using conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[38] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 6

[39] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017. 2

[40] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *CVPR*, 2019. 6

[41] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, 2021. 6, 7

[42] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 3, 7

[43] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1833–1844, 2021. 3, 7

[44] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1132–1140, 2017. 2

[45] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[46] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018. 8

[47] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 3, 5

[48] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[49] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 416–423. IEEE, 2001. 7, 8

[50] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 7

[51] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017. 6

[52] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*, pages 191–207, 2020. 7

[53] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*, 2020. 6

[54] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 3

[55] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *ICCV*, 2021. 6

[56] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[57] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking de-

blurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020. 6

[58] HR Sheikh. Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 8

[59] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019. 6

[60] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *CVPR*, 2015. 6

[61] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 2

[62] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *ICCV*, 2021. 6

[63] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, 2020. 6

[64] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 6

[65] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *Proceedings of the European Conference on Computer Vision*, pages 146–162. Springer, 2022. 6

[66] Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Yen-Yu Lin, and Chia-Wen Lin. BANet: A blur-aware attention network for dynamic scene deblurring. *IEEE Transactions on Image Processing*, 31:6789–6799, 2022. 6

[67] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022. 6

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[69] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2

[70] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[71] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 2

[72] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[73] Xin-Jing Wang, Wei-Ying Ma, Gui-Rong Xue, and Xing Li. Multi-model similarity propagation and its application for web image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 944–951, 2004. 5

[74] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 6, 7

[75] Jiqing Wu, Radu Timofte, and Luc Van Gool. Demosaicing based on directional difference regression and efficient regression priors. *IEEE Transactions on Image Processing*, 25(8):3862–3874, 2016. 8

[76] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 6, 7

[77] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 6

[78] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proceedings of International Conference on Curves and Surfaces*, pages 711–730. Springer, 2010. 7

[79] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 2019. 6

[80] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 7, 8

[81] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, 2020. 6

[82] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 7, 8

[83] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging*, 20(2):023016, 2011. 7, 8

[84] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018. 7

[85] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2, 7, 8

[86] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2