KNN Local Attention for Image Restoration

Hunsang Lee[1], Hyesong Choi[2], Kwanghoon Sohn[1], Dongbo Min[2†]
[1]Yonsei University, Korea, [2]Ewha W. University, Korea
hslee91@yonsei.ac.kr, hyesongchoi2010@gmail.ac.kr, khsohn@yonsei.ac.kr, dbmin@ewha.ac.kr

Abstract

Recent works attempt to integrate the non-local operation with CNNs or Transformer, achieving remarkable performance in image restoration tasks. The global similarity, however, has the problems of the lack of locality and the high computational complexity that is quadratic to an input resolution. The local attention mechanism alleviates these issues by introducing the inductive bias of the locality with convolution-like operators. However, by focusing only on adjacent positions, the local attention suffers from an insufficient receptive field for image restoration. In this paper, we propose a new attention mechanism for image restoration, called k-NN Image Transformer (KiT), that rectifies the above mentioned limitations. Specifically, the KiT groups k-nearest neighbor patches with locality sensitive hashing (LSH), and the grouped patches are aggregated into each query patch by performing a pair-wise local attention. In this way, the pair-wise operation establishes non-local connectivity while maintaining the desired properties of the local attention, i.e., inductive bias of locality and linear complexity to input resolution. The proposed method outperforms state-of-the-art restoration approaches on image denoising, deblurring and deraining benchmarks. The code will be available soon.

1. Introduction

Image restoration aims to recover a clean image from various type of degradations (e.g. noise, blur, rain, and compression artifacts), which has a huge impact on the performance of downstream tasks such as image classification [14, 56], object detection [22, 46], segmentation [4, 10], and to name a few. It is a highly ill-posed inverse problem as there may exist multiple number of solutions for a single degraded image. Recent restoration works [17, 36, 76] attempt to establish a mapping relation between clean and degraded images by leveraging the representation power of the convolutional neural networks (CNNs). A series of local operations used in the CNNs is, however, inherently less capable of capturing a long-range dependency, exhibiting certain limitations in deliberating global information over an entire image. To enlarge the receptive field, increasing network depth [51], dilated convolution [66], and hierarchical architecture [40] have been proposed, but the receptive field still does not secure global information as it is limited to local regions. Recently, *non-local* operation, which mostly contributed to non-learning based restoration approaches [5, 15], has again emerged as a promising solution with the success of non-local neural networks [58]. As similar patterns tend to repeat within a natural image, non-local self-similarity of computing the response at a single position by weighted sum of all positions has served as an important cue for an image restoration [16, 28, 32, 37, 38, 43, 53, 77, 78]. A non-local self-similarity of [58] could capture the long-range dependency within deep networks, but the quadratic complexity with respect to the input feature resolution limits the network capacity. Consequently, it is employed only in relatively low-resolution feature maps of specific layers [16, 32, 77].

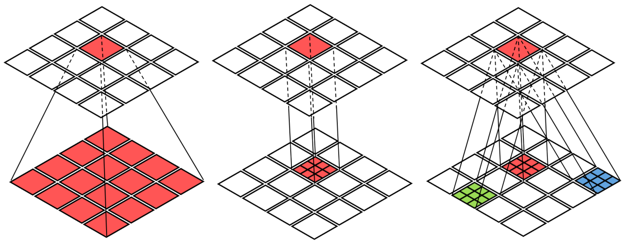


Figure 1. Comparisons of different attention approaches: (a) Global attention [18, 45, 57] computes self-similarity between patches globally, (b) Local attention [33, 59] measures self-similarity within a single patch at the pixel-level, and (c) the proposed method aggregates similar k patches with a pair-wise local attention at the pixel-level.

More recently, Vision Transformer (ViT) [18] proposed a new approach to apply the *global* attention mechanism, which can be viewed as the non-local operation, of the Transformer [55] to vision tasks by splitting an image into

*This work was supported by the Mid-Career Researcher Program through the NRF of Korea (NRF-2021R1A2C2011624 and NRF-2021R1A2C2006703) and the Yonsei University Research Fund of 2021 (2021-22-0001).

[†] Corresponding author: dbmin@ewha.ac.kr

a set of non-overlapping patches, embedding into the feature space, and feeding them into multiple transformer layers to model global self-similarities among patches (Fig. 1 (a)). ViT achieved a pleasing trade-off between accuracy and computational complexity in the image classification task, but the quadratic complexity with respect to the input feature resolution still makes it nearly infeasible to apply the transformer to dense prediction tasks. To overcome this limitation, different from ViT that maintains feature resolutions across the entire network, some approaches [45, 57] proposed a hierarchical architecture to exploit multi-scale feature maps that are suitable for dense prediction tasks. However, they focus only on capturing global self-similarity, and their capability in exploring locality that is essential for image restoration is inferior to that of CNNs.

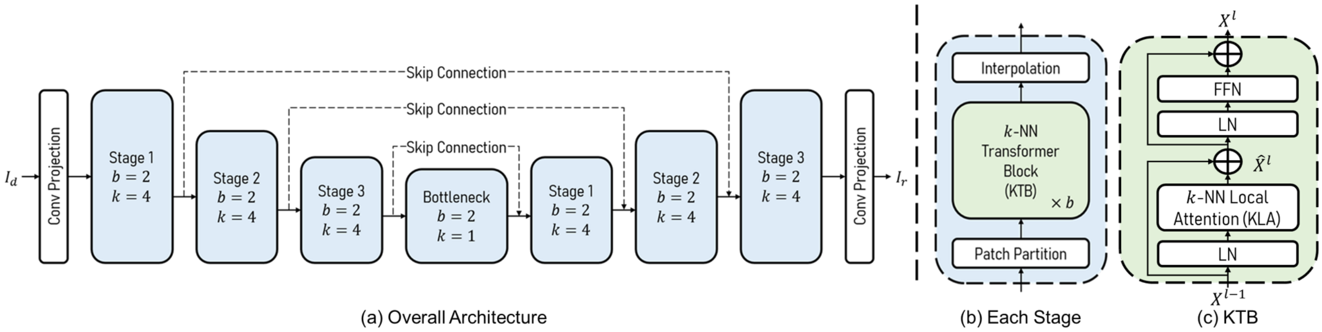
In this context, numerous methods have been proposed to introduce the inductive bias of locality into transformer architectures [30,33,59,61,63]. Among them, *local* attention is considered in recent works [31,33,54,59,67] at the cost of restricting the receptive field in the transformer. These approaches propose the local self-attention module, achieving a linear complexity to the input feature resolution (Fig. 1 (b)). Since they constrain the self-attention computation only within a local patch, a shifting approach [31, 33, 59] is additionally applied to exchange information across non-overlapping patches. However, it considers only neighboring patches and thus still has insufficient receptive field.

In this paper, we propose a novel non-local image restoration method, called k-NN Image Transformer (KiT), that successfully captures locality while explicitly establishing non-local connectivity by considering the local attention of k nearest neighbor (k-NN) patches. To remedy the lack of the long-range dependency inherent in the local attention, the proposed method considers k matched patches that generate non-local connectivity between patches of different positions. To be specific, the KiT first searches a set of similar patches for each base patch with k-NN matching, and then sets the base patch as query and k matched patches as key and value for applying *pair-wise attention* locally, as shown in Fig. 1 (c). This enables our method to apply the local attention over an entire image while maintaining a linear complexity with respect to the feature resolution. Additionally, the inductive bias of locality enhances local feature extraction capability. As shown in Fig. 2, our method consists of a series of k-NN transformer block (KTB), and adopts U-shaped hierarchical architecture for efficiently leveraging multi-scale features. Comprehensive experiments on various image restoration tasks demonstrate the effectiveness of the proposed method over state-of-the-art methods.

2. Related Works

Non-local image restoration. Non-local operation has been widely used in the image restoration. In classical approaches [15, 35], a set of pixels grouped by self-similarity contributes to an output filtered response. Recently, with the success of non-local neural networks [58], some methods [16, 32, 77] attempted to integrate the non-local operation into CNNs for image restoration tasks by establishing the long-range dependency with the global self-attention. However, its expensive computational cost limits the spatial resolution of feature maps or network depth. To reduce computational cost, sparse connections were used in [28, 37, 38, 43, 53, 78] instead of full connections within the input feature map. N^3Net [43] and GCDN [53] find k-nearest neighbors that are close in the embedding space in a learnable manner, and aggregate them for an efficient computation. DAGL [38] dynamically selects the number of neighbors for each query which has distinct distributions according to an image content. IGNN [78] and CPNet [28] finds k-NN patches among cross-scale feature maps by considering both sparseness and cross-scale patch recurrency. Nevertheless, aforementioned approaches have the quadratic complexity for k-NN matching that heavily slows the entire process. NLSN [37] reduces the complexity of k-NN matching process to be asymptotic linear by performing non-local sparse attention with locality sensitive hashing (LSH). But, as the NLSN [37] approximated full connection of the global attention in pixel-level, local information can not be captured in their attention module.

Vision Transformer. In [18], the Transformer architecture [55], originally proposed for natural language processing, was applied to the image classification task. This method, called Vision Transformer (ViT), is remarkable at capturing the long range dependencies by applying *global* attention to image patches, but is not suitable for dense predictions due to the quadratic complexity to an input spatial resolution. Unlike ViT that maintains a fixed spatial resolution across the entire architecture, the hierarchical architecture, where feature resolutions are progressively reduced, is adopted for conducting the dense prediction more effectively [9,45,57,62]. PvT [57] builds pyramid feature maps with spatial reduction attention (SPA) layer. IPT [9] and DPT [45] propose an encoder-decoder architecture to recover fine-grained predictions. However, these approaches based on the global attention lacks the capability that explores locality essential for image restoration. Lately, Swin Transformer [33] leverages the local attention with a shifting approach for patch connection and achieves a competitive performance on object detection and segmentation with low complexity. As the local attention module generates attention weights among adjacent elements only, the computational complexity is linear to the spatial resolution and the inductive bias of locality is injected into the attention. Uformer [59] and SwinIR [31] adopt the local attention for image restoration tasks, demonstrating impressive results. However, the shifting approach still has a limited recep-



(a) Overall Architecture

(b) Each Stage

(c) KTB

Figure 2. Overall architecture of the k-NN Image Transformer (KiT): (a) U-shaped hierarchical architecture is adopted for image restoration. (b) Each stage has two k-NN Transformer Blocks (KTBs) and an interpolation layer. For a skip-connection, the output feature of i-th stage in the encoder is concatenated with $(4 - i)$-th stage in the decoder. (c) The KTB consists of layer normalization (LN), k-NN local attention (KLA), and feed-forward network (FFN) consisting of depth-wise convolution (DW) and multi-layer perceptron (MLP).

tive field due to the nature of considering neighbor patches, thereby losing the non-local connectivity. In contrast, the proposed method establishes non-local connectivity by performing the pair-wise local attention with k-NN patches. This enables us to impose the non-local connectivity with a linear complexity with respect to the spatial resolution, while capturing locality in the attention module.

3. Proposed Method

3.1. Problem Statement and Overview

The non-local self-similarity is known to be effective in the image restoration task [9, 16, 32, 37, 43, 77]. As similar patterns are globally distributed within the image, this requires the capability to capture a long-range dependency. The ViT [18] applies the attention mechanism of an original Transformer [55] directly to sequences of image patches. For a given input $X \in \mathbb{R}^{HW \times C_{in}}$, they split it into non-overlapping patches, and reshape into a sequence of flattened 2D patches $X_p \in \mathbb{R}^{N \times r^2 C_{in}}$, where HW is the spatial resolution of the input feature map, C_{in} is the channel of input feature map, $N = HW/r^2$, and r is the patch size. The global attention with dot-product between split patches is represented as:

$$O = softmax\Big(\frac{\phi(X_p)\theta(X_p)^\mathsf{T}}{\sqrt{C}}\Big)\psi(X_p). \tag{1}$$

The learnable projection functions $\phi, \theta : \mathbb{R}^{N \times r^2 C_{in}} \rightarrow \mathbb{R}^{N \times r^2 C}$, and $\psi : \mathbb{R}^{N \times r^2 C_{in}} \rightarrow \mathbb{R}^{N \times r^2 C_{out}}$ project X_p into the *query*, *key*, and *value*, respectively. The output $O \in \mathbb{R}^{N \times C_{out}}$, where C_{out} is an output channel size, is obtained as an weighted sum of the projected values using the affinity matrix computed between the projected query and key. As C, C_{in} and C_{out} are usually set the same, we denote them as C. Although the global attention mechanism establishes the long-range dependency well, the quadratic complexity to the input feature resolution, $\mathcal{O}(r^2 N^2 C)$, makes it hard to take advantage of global attention for dense prediction tasks.

The local attention mechanism [31, 33, 54, 59, 67] reduces the complexity by computing attention within a local patch. An input feature map X is split into non-overlapping patches, satisfying $X = \{x_i \in \mathbb{R}^{r^2 \times C} \mid i = 0, \ldots, N-1\}$. The local attention is computed within each patch individually

$$o_i = softmax\Big(\frac{\phi(x_i)\theta(x_i)^\mathsf{T}}{\sqrt{C}}\Big)\psi(x_i), \tag{2}$$

where o_i is an output patch corresponding to x_i. Note that the learnable projection functions ϕ, θ and ψ project r^2 elements with a size of C, unlike ViT projecting N elements with a size of $r^2 C$, and are shared for all patches. The local attention achieves the linear complexity $\mathcal{O}(r^4 N C)$ to the input feature resolution. However, as Eq. (2) is applied to each patch separately, no information is exchanged across patches. Thus, a shifting approach [31, 33, 59] is sequentially applied for imposing patch connectivity among neighbor patches with an enlarged receptive field. Nevertheless, as only neighbor patches contributes to the query patch, the receptive field is still limited.

We overcome this limitation by leveraging k-NN in the computation of the local attention, termed k-NN local attention. In order to impose the non-local connectivity in computing the local attention, we utilize k-NN search to seek a set of patch candidates used for computing the local attention. By conducting the pair-wise local attention between a query patch and k matched patches, the proposed method captures locality efficiently while establishing non-local connectivity essential for image restoration.

3.2. Overall Pipeline

The overall framework for image restoration is shown in Fig. 2. To restore a degraded image, we first conduct three convolutions to a degraded input image I_d, and then pass it through three stages of the encoder network and the decoder

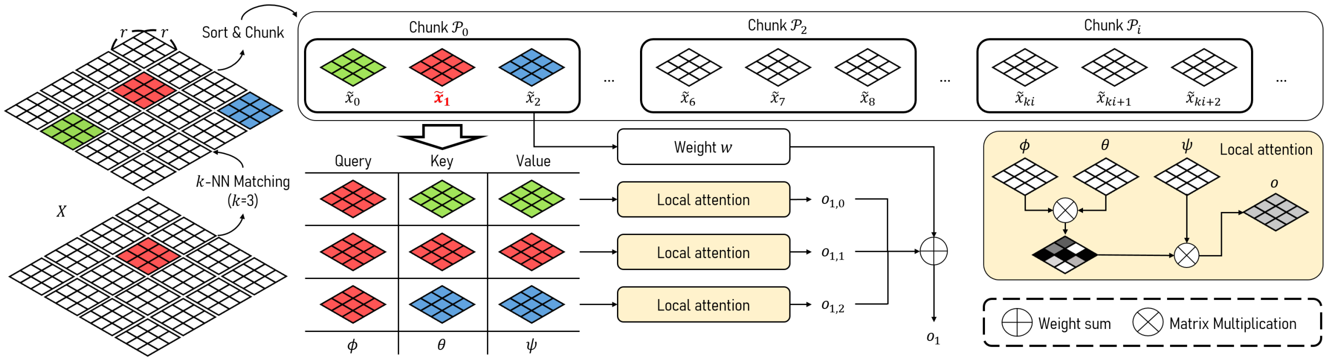


Figure 3. The proposed k-NN local attention (KLA): k-NN search is first conducted for finding k similar patches within an entire feature map. Here, we use locality sensitive hashing (LSH) for an efficient k-NN search. The LSH assigns hash values to patches, and the patches are sorted by the hash values in ascending order. The sorted patches are then partitioned with a chunk size of k. In each chunk, the pair-wise local attention among patches are executed, and the final output is then computed by an weighted sum of the local attention outputs. Here, the KLA for $\tilde{x}_1$ is illustrated when a single chunk is used. In pratice, the previous chunk also contributes to the current chunk containing the query patch. For more details, please refer to Sec. 3.3.

network. Each stage is comprised of the patch partition, k-NN Transformer Blocks (KTB), and an interpolation layer. The patch partition operation splits the input feature map X into non-overlapping patches with the patch size r, satisfying $X = \{x_i \in \mathbb{R}^{r^2 \times C} \mid i = 0, \ldots, N-1\}$.

In the KTB, the split patches are normalized and fed to k-NN local attention (KLA) for non-local aggregation. The KLA first establishes k patches based on the fact that patches with similar patterns frequently appear within an image and aggregating them is beneficial to the image restoration. Note that the existing sparse attention approaches [26, 37] cluster pixels into separate groups to approximate global attention, and thus naturally the lack of the locality, which is a drawback of the global attention, still remains in their attention module. On the other hand, the proposed method computes the sparse similarity between query patch and k patches with the pair-wise local attention, and the inductive bias of the locality is reflected in the proposed attention module.

The proposed network has a U-shaped hierarchical architecture for taking patterns of various scales into account. The aggregated features pass through an interpolation layer (downsampling for encoder and upsampling for decoder). In the each stage of the decoder, input feature maps are concatenated with corresponding encoder features for recovering fine details. At the end of the network, three convolutions are conducted to predict a restored image from the output feature map.

3.3. KTB: k-NN Transformer Block

A layer normalization (LN) is applied to each patch, and then the k-NN local attention (KLA) conducts local attention with *query* patch and matched patches as *key* and *value*. To enhance locality of the network, depth-wise convolution (DW) [13] is employed together with multi-layer perceptron (MLP) in the feed-forward network (FFN) [30], as depicted in Fig. 2 (c). Formally, the KTB is written as:

$$\hat{X}^l = \text{KLA}(\text{LN}(X^{l-1})) + X^{l-1}, \tag{3}$$

$$X^l = \text{FFN}(LN(\hat{X}^l)) + \hat{X}^l, \tag{4}$$

where $\text{FFN}(X) = \text{MLP}(\text{DW}(\text{MLP}(X)))$. In l-th block of each stage ($l = 0, \ldots, b-1$), the output feature map X^{l-1} from the previous block is normalized and fed into KLA. The intermediate feature $\hat{X}^l$ is computed via a non-local aggregation of k similar patch features and residual connection. The bottleneck stage is identical to the KTB, except that the interpolation layer is not used and k is set to 1.

k-nearest neighbor matching. A brute-force k-NN matching requires computing a pair-wise distance between two patches. As this pair-wise distance involves the quadratic complexity to an input length, we leverage the locality sensitive hashing (LSH) [2] that has linear computational complexity. The LSH projects split patches into an unit hypersphere to establish buckets. Assuming there are m hash buckets, a hash value $L(x)$ is assigned by multiplying random rotation matrix $R \in \mathbb{R}^{N \times m/2}$ to a spherically projected patch x as:

$$L(x) = arg\max([xR; -xR]), \tag{5}$$

where $[\cdot;\cdot]$ indicates the concatenation of two elements. With this hashing operation, patches with high correlation are very likely to receive the same hash value (in the same hash bucket), and vice versa. However, as LSH depends on random rotation matrix, similar patches may occasionally fall in different hash buckets. To cope with this issue, multi-round LSH is adopted where LSH is applied with different random rotation matrix h times.

KLA: k-NN local attention. As shown in Fig. 3, similar patches are grouped according to the assigned hash values. To make only patches with the same hash value contribute to a query patch efficiently, we first sort patches according to hash values, and then partition the sorted patches into chunks each involving k patches (equal to the number of NN patches) for batching purpose, so that the local attention is performed on only patches in the same chunk. We denote $\pi : n \rightarrow n$ be a permutation that sorts the patches in ascending order of hash values:

$$\pi(x_p) < \pi(x_q) \Rightarrow L(x_p) \leq L(x_q). \tag{6}$$

For the sake of a simplicity, we define $\tilde{x}$ as a sorted patch where $\tilde{x}_p$ is equal to $x_{\pi(p)}$. Then, i-th chunk $\mathcal{P}_i$ for $i = 0, ..., N/k$ contains k patches,

$$\mathcal{P}_i = \{\tilde{x}_{ki}, \tilde{x}_{ki+1}, \tilde{x}_{ki+2}, ..., \tilde{x}_{ki+k-1}\}. \tag{7}$$

The local attention is then conducted within patch pairs in the chunk for non-local aggregation. Sorted input patches $\tilde{X}$ are projected into query, key and value with the learnable projection functions ϕ, θ and $\psi : \mathbb{R}^{r^2 \times C} \rightarrow \mathbb{R}^{r^2 \times C}$, respectively. As there are k patches in a chunk, the local attention is conducted k^2 times. The pair-wise local attention output for p-th patch as a query where q-th patch is used as key and value is defined as:

$$o_{p,q} = softmax(\frac{\phi(\tilde{x}_p)\theta(\tilde{x}_q)^\intercal}{\sqrt{C}})\psi(\tilde{x}_q). \tag{8}$$

For instance, denoting $\phi(\tilde{x}_1)$ as a query patch in chunk $\mathcal{P}_0$ of Fig. 3, there are k output patches $\{o_{1,j} | j = 0, ..., k-1\}$. Different from ViT [18] that performs self-attention of all patches, we perform the local attention, but k times and in the pair-wise manner that computes affinity matrix between two patches (query and key) for enhancing locality as described in Fig. 1 (c).

As k output patches for a query patch are computed, the pair-wise outputs should be aggregated into the query patch. The output patch o_p for the input patch $\tilde{x}_p$ is computed by weighted sum as:

$$o_p = \sum_{j \in \mathcal{N}_p} w_{p,j} \cdot o_{p,j}, \tag{9}$$

where $w_{p,j}$ is a pair-wise relative similarity between patches, and $\mathcal{N}_p$ is a set of patch indices of a chunk to which the query patch $\tilde{x}_p$ belongs,

$$w_{p,q} = \frac{\phi(\overrightarrow{\tilde{x}_p}) \cdot \theta(\overrightarrow{\tilde{x}_q})}{\sum_{j \in \mathcal{N}_p} \phi(\overrightarrow{\tilde{x}_p}) \cdot \theta(\overrightarrow{\tilde{x}_j})}. \tag{10}$$

Here, $\overrightarrow{x_p} \in \mathbb{R}^{r^2 C}$ represents the flatten patch of $\tilde{x}_p$. As the number of patches in a hash bucket is often indivisible by chunk size in practice, the patches with the same hash value may fall into nearby chunks. To deal with it, similar to [26], we allow the previous chunks to contribute to the current chunk containing the query patch, e.g. $\mathcal{P}_{i-1}$ for $\mathcal{P}_i$. Thus, the local attention is conducted $2k$ times for each query patch.

3.4. Training Loss

Following existing image restoration approaches [32, 76], the proposed network also predicts a residual image I_r from the degraded input image I_d. The objective is to recover clean image I satisfying $I = I_d + I_r$. We leverage Charbonnier loss [8] $\mathcal{L}_{char}$ and an edge loss $\mathcal{L}_{edge}$ for optimizing the network,

$$\begin{aligned} &\mathcal{L}_{char} = \sqrt{\|I - (I_d + I_r)\|^2 + \epsilon^2}, \\ &\mathcal{L}_{edge} = \sqrt{\|\triangle I - \triangle(I_d + I_r)\|^2 + \epsilon^2}, \\ &\mathcal{L} = \mathcal{L}_{char} + \lambda \mathcal{L}_{edge}. \end{aligned} \tag{11}$$

where ϵ is empirically set to 10^{-3} for all experiments and $\triangle$ represents the Laplacian function. The total loss $\mathcal{L}$ is defined with $\mathcal{L}_{char}$ and $\mathcal{L}_{edge}$, where a hyper-parameter λ controls the ratio of the two losses.

4. Experiments

4.1. Implementation Details

The proposed KiT was implemented in PyTorch. We trained the whole networks on the batches of 16 images cropped to 128×128 for 300 epochs using AdamW optimizer [34]. The learning rate was set to 1×10^{-4} initially, and the linear warm-up strategy and cosine annealing for decreasing the learning rate were adopted. The chunk size k (equal to the number of NN patches) and patch size r were set to 4 by default. In the bottleneck stage, k is set to 1 since there are only a few patches (e.g. the number of patches is 4×4 when HW is 256×256). The number of KTB in each stage, b, was set to 2 in all stages. In the KLA, the number of hashes, h, was set to 4 for multi-round LSH. We validated the performance of the proposed method on various image restoration tasks such as image denoising, debluring and deraining. For the performance evaluation, the PSNR and SSIM were measured on the RGB space for denoising and deblurring. In deraining, the evaluation was done on the Y channel of the YCbCr color space, following previous works [24, 72]. More results are provided in the supplementary materials.

4.2. Image Denoising

We trained the KiT with the SIDD [1] dataset containing 320 high-resolution images with realistic noise. Tab. 1 shows the quantitative evaluation of real noise removal on the SIDD [1] and DND [42] datasets. The evaluation results include the classical denoisng method [15], CNN-based methods [3, 6, 11, 23, 25, 68, 70–72, 76], self-attention based methods [38] and transformer-based methods [59]. As DND [42] dataset does not provide ground-truth labels, the results were obtained from official benchmark. The proposed method outperforms the state-of-the-art methods both

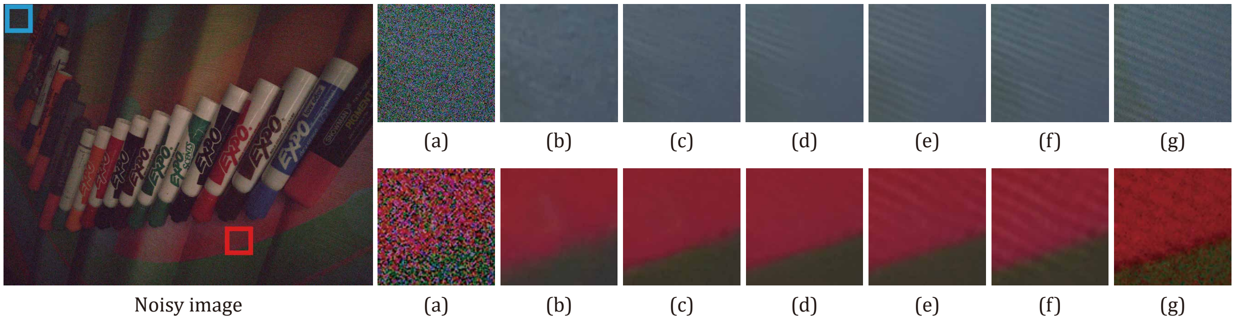


Figure 4. Visual comparisons on the SIDD [1] dataset: (a) Cropped image, (b) RIDNet [3], (c) CycleISP [70], (d) MPRNet [72], (e) Uformer [59], (f) KiT, and (g) ground truth.

Method	SIDD		DND	
	PSNR	SSIM	PSNR	SSIM
BM3D [15]	25.65	0.685	34.51	0.851
DnCNN [76]	23.66	0.583	32.43	0.790
MLP [6]	24.71	0.641	34.23	0.833
CBDNet [23]	30.78	0.801	38.06	0.942
RIDNet [3]	38.71	0.951	39.26	0.953
AINDNet [25]	38.95	0.952	39.37	0.951
VDN [68]	39.28	0.956	39.38	0.952
SADNet [7]	39.46	0.957	39.59	0.952
DANet [69]	39.47	0.957	39.58	0.955
CycleISP [70]	39.52	0.957	39.56	<u>0.956</u>
MPRNet [72]	39.71	0.958	39.80	0.954
MIRNet [71]	39.72	0.958	39.88	<u>0.956</u>
NBNet [11]	39.75	0.973	<u>39.89</u>	0.955
DAGL [38]	-	-	39.83	**0.957**
Uformer [59]	<u>39.77</u>	<u>0.970</u>	**39.96**	<u>0.956</u>
KiT	**39.80**	**0.972**	**39.96**	<u>0.956</u>

Table 1. The quantitative results on SIDD [1] and DND [42] dataset. The bold and underlined numbers indicate the best and the second best results, respectively.

on SIDD datasets and achieves competitive performance on the DND dataset. As the DND dataset does not provide any training data, the performance in Tab. 1 is achieved using the network trained on SIDD dataset, proving the robustness of the proposed method. Fig. 4 shows the denoised images with various state-of-the-art methods. While existing methods output restored images with loss of details, the proposed method successfully restores degraded images with fine-detailed structures thanks to the capability of capturing locality with non-local connectivity.

4.3. Image Deblurring

Tab. 2 reports the image deblurring performance on the GoPro dataset [39]. The GoPro dataset provides synthetic blurry images where each image is obtained by averaging successive sharp images. For training, 2,103 images of the GoPro [39] dataset were used, and 1,111 images of the GoPro [39], 3,758 images of RealBlur [48] and 2,025 images of the HIDE [49] datasets were evaluated. The outstanding performance on the PSNR and SSIM metrics validated that the proposed method is also beneficial to restoring blurry images. Fig. 5 shows restored images from blur artifacts in the GoPro [39] dataset. It is easily found that our results capture sharp and fine details whereas other methods are unable to deal with high-frequency details.

Method	GoPro		HIDE	
	PSNR	SSIM	PSNR	SSIM
DeepDeblur [39]	29.23	0.916	25.73	0.874
SRN [52]	30.26	0.934	28.36	0.915
PSS-NSC [21]	30.92	0.942	29.11	0.913
DMPHN [73]	31.20	0.945	29.09	0.924
SAPHN [50]	32.02	0.953	29.98	0.930
MT-RNN [41]	31.15	0.945	29.15	0.918
SPAIR [44]	32.06	0.953	30.29	0.931
MPRNet [72]	32.66	0.959	<u>30.96</u>	<u>0.939</u>
MIMO-UNet [12]	<u>32.68</u>	<u>0.959</u>	-	-
KiT	**32.70**	**0.959**	**30.98**	**0.942**

Table 2. The quantitative results on GoPro [39] and HIDE [49] dataset. The network was trained on GoPro dataset.

4.4. Image Deraining

Following the experimental setup of [24], 13,712 clean-rain image pairs sampled from multiple datasets [20, 29, 64, 74, 75] were used to train the network for image deraining. We evaluated the deraining results on five datasets, Test100 [75], Rain100H [64], Rain100L [64], Test2800 [20], and Test1200 [74]. While SPAIR [44] that leverages extra distortion-guided networks shows competitive results with the proposed method in terms of PSNR, the proposed KIT achieves a higher SSIM, demonstrating that fine-details can be better restored. As shown in the results of Fig. 4 and Fig. 6, the proposed method has an advantage of dealing with repeated textures thanks to the KLA based aggregation. In rainy images, as many patches with similar patterns exist in the image, the proposed method shows out-

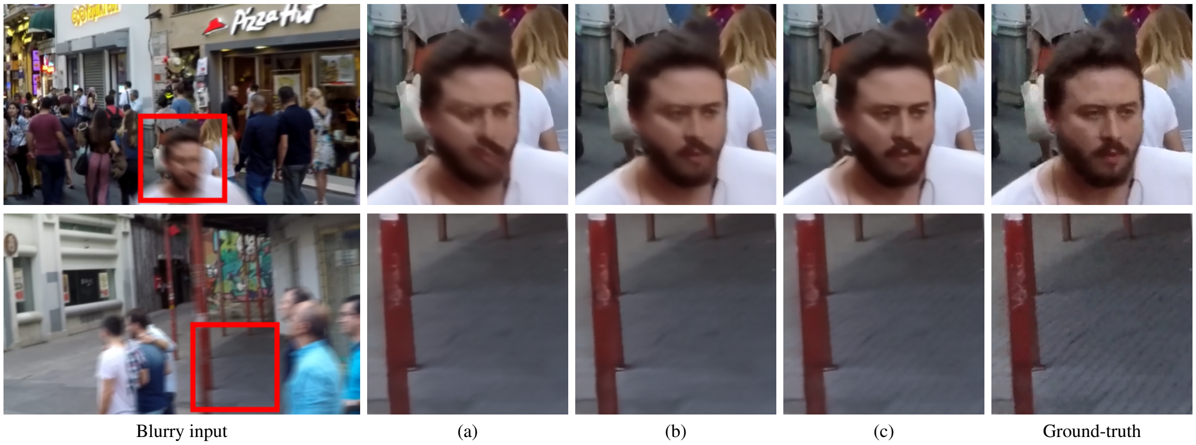


Figure 5. Visual comparisons on the GoPro [39] dataset: (a) MPRNet [72], (b) MIMO-UNet [12], and (c) KiT.

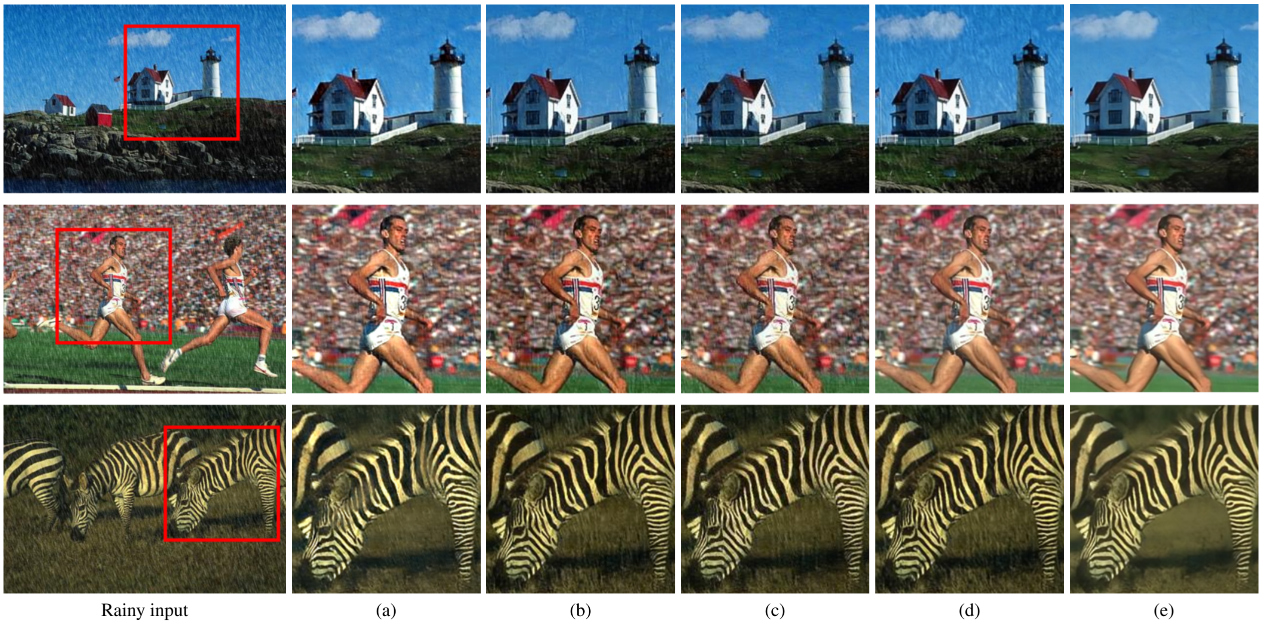


Figure 6. Visual results on the Test100 [75] dataset: (a) DerainNet [19], (b) PreNet [47], (c) RESCAN [27], (d) MPRNet [72], and (e) KiT.

standing performance quantitatively and qualitatively compared with evaluated methods.

4.5. Ablation Study

We conducted the ablation studies to analyze the performance of our method at various aspects. All experiments were conducted on SIDD [1] for image denoising task.

Computational complexity. The proposed method is comprised of multi-round hashing for k-NN search, feature projection, and pair-wise local attention. The multi-round hashing is performed to the input feature patches by multiplying random rotation matrix R, which has $\mathcal{O}(hNCm)$ complexity. Then, each patch is projected to the query, key, and value with the learnable projection functions ϕ, θ and $\psi : \mathbb{R}^{r^2 \times C} \rightarrow \mathbb{R}^{r^2 \times C}$, whose complexity is $\mathcal{O}(NC^2)$. The complexity of computing the local attention between all patch pairs is $\mathcal{O}(khr^4N)$. Thus, all operations takes linear computation with respect to the input feature resolution.

Visualization of the k-NN patches. Our method aims to preserve fine details while achieving non-local connectivity efficiently, achieved by aggregating patches with similar characteristics. To visually validate this, we further visualize the patches used for KLA in Fig. 7. The left most images are divided into non-overlapping patches, where the patches marked with color boxes represent query patches for visualization. As the KLA leverages LSH for k-NN search, similar k patches are grouped with a chunk in the right figures. The k patches belonging to the same chunk are also marked with the same color boxes. With red and green boxes have similar patterns, while the patches with blue boxes include non-textured areas, proving that the LSH finds visually similar patches effectively.

Method	Test100 [75]		Rain100H [64]		Rain100L [64]		Test2800 [20]		Test1200 [74]		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DerainNet [19]	22.77	0.810	14.92	0.592	27.03	0.884	24.31	0.861	23.38	0.835	22.48	0.796
SEMI [60]	22.35	0.788	16.56	0.486	25.03	0.842	24.43	0.782	26.05	0.822	22.88	0.744
DIDMDN [74]	22.56	0.818	17.35	0.524	25.23	0.741	28.13	0.867	29.65	0.901	24.58	0.770
UMRL [65]	24.41	0.829	26.01	0.832	29.18	0.923	29.97	0.905	30.55	0.910	28.02	0.880
RESCAN [27]	25.00	0.835	26.36	0.786	29.80	0.881	31.29	0.904	30.51	0.882	28.59	0.857
PreNet [47]	24.81	0.851	26.77	0.858	32.44	0.950	31.75	0.916	31.36	0.911	29.42	0.897
MSPFN [24]	27.50	0.876	28.66	0.860	32.40	0.933	32.82	0.930	32.39	0.916	30.75	0.903
MPRNet [72]	<u>30.27</u>	0.897	30.41	0.890	36.40	<u>0.965</u>	<u>33.64</u>	<u>0.938</u>	<u>32.91</u>	0.916	32.73	0.921
SPAIR [44]	**30.35**	**0.909**	**30.95**	<u>0.892</u>	**36.93**	**0.969**	33.34	0.936	**33.04**	**0.922**	**32.91**	<u>0.926</u>
KiT	30.26	<u>0.904</u>	<u>30.47</u>	**0.897**	<u>36.65</u>	**0.969**	**33.85**	**0.941**	32.81	<u>0.918</u>	<u>32.81</u>	**0.929**

Table 3. The quantitative results of image deraining. The widely used five datasets [20, 64, 74, 75] are used for evaluation.

PSNR		h				
		1	2	4	8	16
	1	38.79	38.92	38.96	38.96	38.96
	2	39.58	39.69	39.75	39.76	39.78
k	4	39.69	39.76	39.80	39.80	<u>39.81</u>
	8	39.74	39.78	39.80	<u>39.81</u>	**39.82**
	16	39.75	39.78	39.80	<u>39.81</u>	**39.82**

Table 4. Ablation study of the number of patches k and hash rounds h.

The number of k and h. The chunk size k determines the maximum number of patches used for performing the local attention with the query patch. The h hash rounds are used to reduce the probability that similar patches fall into different hash buckets. As the two hyper-parameters do not affect the number of network parameters and are only related to the memory and computational complexity, the network capacity can be flexibly adjusted according to computational resources. Tab. 4 shows the denoising performance of the proposed method according to the two hyper-parameters. The best performance was achieved when the two hyper-parameters are set to 16, but, we set k and h to 4 as it has comparable performance with relatively low computation.

Sharing query and key. In the existing methods [26, 37] that leverage LSH for sparse global attention, the projection functions for query and key should be shared, *i.e.*, $\phi = \theta$, and thus the shared attention masks out the query as the dot-product of a query with itself almost always overwhelms the dot product of a query with a key at other positions. Contrarily, our method has no such a constraint for sharing the projection functions. For the purpose of ablation study, we conducted additional experiments of sharing the projection functions for query and key in Tab. 5, denoted as KiT$_S$. When using the shared projection, the overall performance was slightly reduced.

5. Conclusion

In this paper, we have presented a novel non-local image restoration method. Specifically, the k-NN local attention (KLA) conducts the pair-wise local attention among similar patches with k-NN matching. The KLA holds the inductive bias of locality while establishing the non-local connectivity with the linear computational complexity to the input spatial resolution. The proposed method outperforms the state-of-the-art methods on various image tasks, in terms of quantitative/qualitative performance. As the number of NN patches and hash rounds that determine the network capacity are independent of the network parameters, the flexible adjustment of the network capacity is feasible.

Limitations. The proposed method only considers the pair-wise local attention between patches of the same scale. The cross-scale attention can be an interesting methodology that further improves the restoration performance. We will continue to investigate network that integrates in cross-scale attention for our cross-position based model.

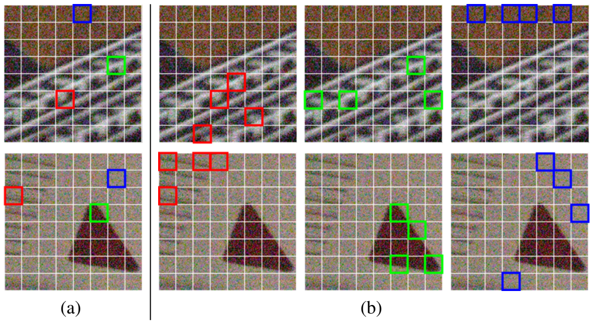


Figure 7. Visualization of the k-NN patches: (a) input image and (b) k-NN patches. k-NN patches are discovered by LSH with k = 4. Patches belonging to the same chunk are marked with boxes of the same color.

Method	SIDD	
	PSNR	SSIM
KiT	39.80	0.972
KiT$_S$	39.75	0.969

Table 5. Ablation study of the shared projection $\phi = \theta$ (KIT$_S$).

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 5, 6, 7

[2] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. *arXiv preprint arXiv:1509.02897*, 2015. 4

[3] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3155–3164, 2019. 5, 6

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1

[5] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. 1

[6] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *2012 IEEE conference on computer vision and pattern recognition*, pages 2392–2399. IEEE, 2012. 5, 6

[7] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. In *European Conference on Computer Vision*, pages 171–187. Springer, 2020. 6

[8] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172. IEEE, 1994. 5

[9] 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. 2, 3

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[11] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4896–4906, 2021. 5, 6

[12] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4641–4650, 2021. 6, 7

[13] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4

[14] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012. 1

[15] 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. 1, 2, 5, 6

[16] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 1, 2, 3

[17] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 1

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 5

[19] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 7, 8

[20] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3855–3863, 2017. 6, 8

[21] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya 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*, pages 3848–3856, 2019. 6

[22] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[23] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1712–1722, 2019. 5, 6

[24] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020. 5, 6, 8

[25] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3482–3492, 2020. 5, 6

[26] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2019. 4, 5, 8

[27] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 254–269, 2018. 7, 8

[28] Yao Li, Xueyang Fu, and Zheng-Jun Zha. Cross-patch graph convolutional network for image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4651–4660, 2021. 1, 2

[29] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2744, 2016. 6

[30] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 2, 4

[31] 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*, pages 1833–1844, 2021. 2, 3

[32] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *arXiv preprint arXiv:1806.02919*, 2018. 1, 2, 3, 5

[33] 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. *arXiv preprint arXiv:2103.14030*, 2021. 1, 2, 3

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[35] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *2009 IEEE 12th international conference on computer vision*, pages 2272–2279. IEEE, 2009. 2

[36] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29:2802–2810, 2016. 1

[37] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 1, 2, 3, 4, 8

[38] Chong Mou, Jian Zhang, and Zhuoyuan Wu. Dynamic attentive graph learning for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4328–4337, 2021. 1, 2, 5, 6

[39] 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, 7

[40] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1

[41] 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 *European Conference on Computer Vision*, pages 327–343. Springer, 2020. 6

[42] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1586–1595, 2017. 5, 6

[43] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. *Advances in Neural Information Processing Systems*, 31:1087–1098, 2018. 1, 2, 3

[44] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2309–2319, 2021. 6, 8

[45] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1, 2

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[47] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3937–3946, 2019. 7, 8

[48] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020. 6

[49] 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

[50] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3615, 2020. 6

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[52] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 6

[53] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deep graph-convolutional image denoising. *IEEE Transactions on Image Processing*, 29:8226–8237, 2020. 1, 2

[54] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021. 2, 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2, 3

[56] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 1

[57] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 1, 2

[58] 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. 1, 2

[59] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 1, 2, 3, 5, 6

[60] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2019. 8

[61] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 2

[62] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 2

[63] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers, 2021. 2

[64] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017. 6, 8

[65] Rajeev Yasarla and Vishal 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*, pages 8405–8414, 2019. 8

[66] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1

[67] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *arXiv preprint arXiv:2106.13112*, 2021. 2, 3

[68] Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational denoising network: Toward blind noise modeling and removal. *arXiv preprint arXiv:1908.11314*, 2019. 5, 6

[69] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *European Conference on Computer Vision*, pages 41–58. Springer, 2020. 6

[70] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2696–2705, 2020. 5, 6

[71] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. 5, 6

[72] 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. 5, 6, 7, 8

[73] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019. 6

[74] He Zhang and Vishal 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*, pages 695–704, 2018. 6, 8

[75] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019. 6, 7, 8

[76] 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. 1, 5, 6

[77] 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. 1, 2, 3

[78] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *Advances in Neural Information Processing Systems*, 2020. 1, 2