IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 December 2022, accepted 26 January 2023, date of publication 3 February 2023, date of current version 10 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242556

RESEARCH ARTICLE

Cross-Scale KNN Image Transformer for Image Restoration

HUNSANG LEE[1], (Student Member, IEEE), HYESONG CHOI[2], (Student Member, IEEE), KWANGHOON SOHN[1], (Senior Member, IEEE), AND DONGBO MIN[2], (Senior Member, IEEE)

[1]School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
[2]Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, South Korea

Corresponding author: Dongbo Min (dbmin@ewha.ac.kr)

This work was supported in part by the Mid-Career Researcher Program, and in part by the Basic Research Laboratory Program through the National Research Foundation (NRF) of Korea under Grant 2021R1A2C2011624 and Grant 2021R1A4A1032582.

ABSTRACT Numerous image restoration approaches have been proposed based on attention mechanism, achieving superior performance to convolutional neural networks (CNNs) based counterparts. However, they do not leverage the attention model in a form fully suited to the image restoration tasks. In this paper, we propose an image restoration network with a novel attention mechanism, called cross-scale k-NN image Transformer (CS-KiT), that effectively considers several factors such as locality, non-locality, and cross-scale aggregation, which are essential to image restoration. To achieve locality and non-locality, the CS-KiT builds k-nearest neighbor relation of local patches and aggregates similar patches through local attention. To induce cross-scale aggregation, we ensure that each local patch embraces different scale information with scale-aware patch embedding (SPE) which predicts an input patch scale through a combination of multi-scale convolution branches. We show the effectiveness of the CS-KiT with experimental results, outperforming state-of-the-art restoration approaches on image denoising, deblurring, and deraining benchmarks.

INDEX TERMS Image restoration, denoising, deblurring, deraining, transformer, self-attention, k-nn search, transformer, low-level vision.

I. INTRODUCTION

Image restoration, which has the intention of recovering a pure image from various types of degradations including noise, blur, rain, and compression artifacts, exerts a strong influence on the performance of downstream tasks such as image classification [1], [2], object detection [3], [4], segmentation [5], [6], and to name a few. As numerous solutions can exist for a single degraded input, image restoration is an ill-posed inverse problem.

Although numerous methods have been proposed for image restoration over the past few decades, several challenges have still remained due to various factors to consider in image restoration: locality, non-locality, and cross-scale aggregation. The locality (local textures or edges) is a key factor in restoring the degraded image in that neighboring pixels are highly correlated. With the advances of convolutional neural networks (CNNs), recent restoration works [7], [8], [9] tried to establish a mapping relation between clean and degraded images by leveraging the representation power of the CNNs. The inductive bias of locality is well driven into the network by a virtue of the local operation in CNNs, but it inherently lacks the ability to capture long-range dependency, thus disregarding the knowledge of global image statistics. This limitation may be overcome by enlarging the receptive field of the convolution operation such as increasing network depth [10], dilated convolution [11], and hierarchical architecture [12]. Despite the large receptive field, it is still insufficient to model non-locality, considering that aggregating resemble patterns commonly come into existence within an entire image boosts the restoration performance significantly. In this context, *non-local* operation, which primarily contributed to traditional non-learning restoration methods [13], [14], has once more become a promising

The associate editor coordinating the review of this manuscript and approving it for publication was Bing Li.

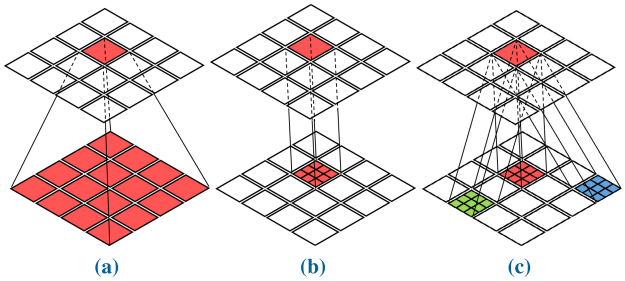


FIGURE 1. Comparisons of different attention approaches: (a) Global attention [19], [20], [21] computes self-similarity between patches globally, (b) Local attention [22], [23] measures self-similarity within a single patch at the pixel-level, and (c) the proposed method aggregates similar k patches through local attention at the pixel-level.

solution as a result of the development of non-local neural networks [15]. By computing the response at a single position with a weighted sum of the features at all positions, they capture the long-range dependency within deep networks, but their capacity is limited by quadratic complexity with respect to input feature resolution. It is therefore only used in relatively low-resolution feature maps of particular layers [16], [17], [18].

More recently, self-attention mechanisms have arisen as a new trend in the field of computer vision, with Vision Transformer (ViT) [19] drawing attention by achieving a pleasant trade-off between accuracy and computational complexity in the image classification task. In ViT, the global attention mechanism, which can be viewed as a non-local operation, models self-similarity among non-overlapping split patches, as shown in Fig. 1 (a). Naturally, they also suffer from quadratic complexity with respect to the input feature resolution, which makes it nearly infeasible to apply the transformer to dense prediction tasks. To overcome this limitation, different from ViT has a columnar structure network in which the feature resolution remains unchanged, a hierarchical architecture was proposed by [20] and [21] to exploit multi-scale feature maps that are suitable for dense prediction tasks. Although they capture global self-similarity, they are far less capable of exploring the locality than CNNs, which is essential to image restoration.

In this context, numerous methods have been proposed to introduce the inductive bias of locality into transformer architectures [22], [23], [24], [25], [26]. Among them, *local* attention is considered in recent works [22], [23], [27], [28], [29] at the cost of restricting the receptive field in the transformer as shown in Fig. 1 (b). These approaches propose the local self-attention module, achieving a linear complexity to the input feature resolution. Since they constrain the self-attention computation only within a local patch, a shifting approach [22], [23], [29] is additionally applied to exchange information across non-overlapping patches. However, similar to CNNs, it considers only neighboring patches and thus still has an insufficient receptive field.

The aforementioned attention approaches take into account only patches of the same scale, but in various vision tasks, cross-scale aggregation is non-negligible due to the scales of similar objects (or patterns) that may often vary within an image. A number of studiesd [24], [30] have explored the cross-scale aggregation via interaction among concentric embeddings covering different receptive fields. However, since each patch typically has a single representative scale, considering all interactions between multi-scale embeddings is rather inefficient. This problem becomes more intensified in dense prediction tasks requiring more parameters and memory consumption.

To tackle these limitations, we propose a transformer-based image restoration network, cross-scale k-NN image Transformer (CS-KiT), where the locality, non-locality, and cross-scale aggregation are efficiently and comprehensively considered. First, to achieve cross-scale aggregation, we propose a scale-aware patch embedding (SPE) that estimates a representative scale of each patch through combinations of convolutions of different kernel sizes and forms mixed-scale patches. Consequently, the cross-scale aggregation can be implicitly accomplished by applying the proposed self-attention method to the patches with mixed-scale, where different ranges of receptive fields are covered through the SPE. Second, to capture locality while explicitly establishing non-local connectivity, we propose a novel attention mechanism, called k-NN local attention (KLA), which takes into account the local attention of k-nearest neighbor (NN) patches. To compensate for the lack of the long-range dependency inherent in local attention, the proposed method considers k matched patches that generate non-local connectivity between patches of different positions.

To be specific, the KLA first groups a set of similar patches for each base patch with k-NN search, and then aggregates k matched patches through local attention in a pixel-level, as shown in Fig. 1 (c). This enables our method to apply local attention over an entire image while maintaining a linear complexity with respect to the feature resolution. Additionally, the inductive bias of locality contributes to the enhanced capability of local feature extraction. For efficiency, KLA leverages an approximated k-NN search via locality sensitive hashing (LSH) [31] which assigns the same hash value to similar patches with a high probability, and then aggregates similar patches with the same hash values. For an efficient batch-wise parallel computation, patches are sorted according to a hash value and split into chunks of size k so that similar patches are placed in the same chunk. However, in practice, the bucket size, indicating the number of patches (with the same hash value) belonging to the bucket, varies and is often not divisible with the chunk size k. Thus, each chunk may contain isolated patches with different hash values. To deal with this issue, we propose a chunk shift under the assumption that the relation of patches would be similar in an adjacent block. By shifting and sharing patch indexes of the current block in the successive block in each stage, as shown in Fig. 2, chunk

shift allows an adjacent chunk to attend to a query patch, dealing with the isolated patches problem in an efficient way.

We validate the proposed method on various image restoration tasks on image denoising, deblurring, and deraining benchmarks, demonstrating superior performance to state-of-the-art restoration approaches. A preliminary version of this paper has appeared as a full paper in [32]. Compared with our previous work, we newly add (1) scale-aware patch embedding for cross-scale aggregation; (2) efficient handling of the isolated patches problem using chunk shift, and (3) extensive analysis of the proposed method.

II. RELATED WORK

A. IMAGE RESTORATION

1) CNN-BASED IMAGE RESTORATION

Image restoration aims at recovering a clean image from an input image degraded by various factors such as noise, motion blur, rain streak, or low-light conditions. With the rapid progress of CNN, image restoration techniques using deep learning achieved evident performance improvement compared to traditional approaches thanks to their representation power. By introducing residual learning that models degradation as residuals image, [7], [33], [34] surpass the methods directly estimate clean image [35], [36]. With further developments, a dense connection between layers within the same block [37], [38] is employed to form a deeper network. As another steam different from regression network, adversarial learning [39], [40] Although these approaches have obtained enormous success in image restoration, convolution operation inherently suffers from a lack of non-locality essential in image restoration. Consequently, this fact led to the necessity of incorporating non-local into the network.

2) NON-LOCAL IMAGE RESTORATION

In image restoration, the non-local operation has been widely used. In classical approaches [14], [41], a set of pixels grouped by self-similarity contributes to an output filtered response. Recently, some efforts [16], [17], [18] tried to integrate the non-local operation into CNNs for image restoration tasks by establishing the long-range dependency with global self-attention following the success of non-local neural networks [15]. However, its heavy computational cost limits the spatial resolution of feature maps. Rather than employing full connections within the input feature map, sparse connections were adopted in [42], [43], [44], [45], [46], and [47] to cut down on computational costs. N^3Net [42] and GCDN [43] find k-nearest neighbors that are close in the embedding space in a learnable manner, and aggregate them for efficient computation. According on the content of the images, DAGL [44] dynamically selects the number of neighbors for each query. IGNN [45] and CPNet [46] find k-NN patches among cross-scale feature maps by considering both sparseness and cross-scale patch recurrency. Nevertheless, the quadratic complexity for k-NN search of the aforementioned methods significantly slows down the overall procedure. NLSN [47] reduces the complexity of k-NN search process to be asymptotic linear by performing non-local sparse attention with locality sensitive hashing (LSH). Local information, however, cannot be captured in their attention module because the NLSN [47] approximated the full connection of the global attention at the pixel level.

3) MULTI-SCALE AGGREGATION FOR IMAGE RESTORATION

Multi-scale information is regarded as an indispensable property in image restoration tasks as it is beneficial to encompass various scales of objects, patterns, or degradations. As a naive way to deal with multi-scale information, exploiting feature maps of multiple scales by building hierarchical architecture [48], [49], [50] or multi-scale parallel branches [43], [51] has been explored. In another way, a progressive prediction [52], [53], [54] that gradually scales up the model capacity and difficulty of the problem (*e.g.* low resolution to high resolution) significantly improved the image restoration performance. However, these attempts have a certain limitation in dealing with cross-scale patch recurrence as they only build intra-scale relationships within single attention. To build cross-scale interactions, emerging approaches [45], [55], [56] focus on the interchange of mutual information across different scales. CSNLN [55] fuses multi-branch projections of cross-scale non-local attention, in-scale non-local attention, and identity mapping. TTSR [56] proposes a cross-scale feature integration module to transfer reference HR textures to the low-resolution image. IGNN [45] aggregates k-NN high-resolution counterparts corresponding to the low-resolution patch.

B. ATTENTION MECHANISM

Inspired by the human vision system, attention mechanisms have been introduced to allow the network to focus on saliency and benefit various tasks such as recognition [57], [58], [59], [60], object detection [61], [62], and semantic segmentation [63], [64]. In CNN-era, RAM [65], a pioneer of visual attention, exploits visual attention on the recurrent neural network to classify images. Afterward, STN [66] proposes a spatial transformer network that predicts an affine transformation to select the most relevant regions. SENet [57] proposes a squeeze-and-excitation block that squeezes the spatial resolution and captures the channel-wise relationship. CBAM [58] exploits both the spatial relationship between the feature map of the pooled channel and the channel-wise relationship in the spatially reduced feature map. Non-local neural networks [15] combine non-local property and spatial self-attention. By capturing long-range dependencies, they show superior performance on various visual tasks including video classification. As a non-local neural network has quadratic complexity with respect to an input resolution, CCNet [63] proposes axial-wise attention (horizontal and vertical directions) to achieve a less computational cost of non-local attention. From the finding that the self-similarities

modeled by non-local attention are almost the same for different query positions, GCNet proposes a more simplified non-local operation combined with SENet [57].

Recently, in [19], the Transformer architecture [67], originally proposed for natural language processing, was applied to the image classification task. This method, known as Vision Transformer (ViT), excels at capturing long-range dependencies by applying *global* attention to image patches, but it is not appropriate for dense predictions due to the quadratic complexity with respect to an input spatial resolution. References [20], [21], [68], [69] adopted the hierarchical architecture, where feature resolutions are gradually reduced for enabling the dense prediction, in contrast to ViT, which maintains a fixed spatial resolution across the whole architecture. PVT [20] constructs pyramid feature maps with the spatial reduction attention (SPA) layer. In order to recover fine-grained predictions, IPT [68] and DPT [21] propose an encoder-decoder architecture. However, these approaches based on global attention lack the ability to look into the locality, which is essential for image restoration.

Lately, Swin Transformer [22] has boosted object detection and segmentation performance with low complexity by leveraging local attention and a shifting strategy for patch connection. As the attention weights are produced between neighbor elements by the local attention, computations have linear complexity according to the spatial resolution and the inductive bias of locality is incorporated into the attention. Uformer [23] and SwinIR [29] clearly show remarkable performance in image restoration tasks by utilizing local attention. However, by taking neighboring patches into account, the shifting approach has a limited receptive field, thereby losing the non-local connectivity in the process. The proposed method, in contrast, makes non-local connectivity by carrying out local attention using k-NN patches. This enables us to capture locality in the attention module and impose the non-local connectivity with a linear complexity with respect to spatial resolution.

III. PROBLEM STATEMENT

It is well known that non-local self-similarity performs well in the task of image restoration. This necessitates the capability to capture a long-range dependency since analogous patterns are dispersed across the whole image. The ViT [19] applies the attention mechanism of an original Transformer [67] directly to image patch sequences. 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 $X_p \in \mathbb{R}^{N\times r^2C_{in}}$, where HW is the spatial resolution of the input feature map, C_{in} is the channel of the 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:

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

The learnable projection functions $\phi, \theta : \mathbb{R}^{N\times r^2C_{in}} \rightarrow \mathbb{R}^{N\times r^2C}$, and $\psi : \mathbb{R}^{N\times r^2C_{in}} \rightarrow \mathbb{R}^{N\times r^2C_{out}}$ 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 a 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, $\mathcal{O}(r^2N^2C)$, makes it hard to take advantage of global attention for dense prediction tasks.

The local attention mechanism [22], [23], [27], [28], [29] reduces the complexity by computing attention within a local patch. An input feature map X is split into non-overlapping patches, satisfying $X = \{x_i \in \mathbb{R}^{r^2\times C} \mid i = 1, \ldots, N\}$. The local attention is computed within each patch individually

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

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^2C, and are shared for all patches. The local attention achieves the linear complexity $\mathcal{O}(r^4NC)$ to the input feature resolution. However, as (2) is applied to each patch independently, no information is exchanged between patches. In order to enforce patch connectivity between neighbor patches with an enlarged receptive field, a shifting approach [22], [23], [29] is sequentially applied.

Also, both attention approaches overlook the scales of objects or patterns that may vary in real images, considering the interactions among features of a single scale only. Recently, some works considered cross-scale attention by introducing multi-scale information in feature embedding [30], linear projection [24], [70], or multi-path structure [71]. However, as an object or a pattern usually has a single representative scale, figuring correlations between multiple scales out calls for superfluous computations. Due to the computational burden coming from reconstructing high-resolution output in dense prediction tasks, a more efficient and effective attention layer design is required.

We overcome those limitations by leveraging k-NN patches of mixed-scale in the computation of the local attention. A novel non-local image restoration method, called cross-scale k-NN image transformer (CS-KiT), successfully equips locality, non-locality, and cross-scale aggregation.

IV. PROPOESD METHOD

A. OVERALL PIPELINE

Fig. 2 depicts the overall framework of the proposed method for image restoration. To get a restored clean image, We first apply three convolutions to a degraded input image I_d and then pass it through three stages of both the encoder and decoder network. Each stage is comprised of the patch partition, cross-scale k-NN Transformer Block (CS-KTB), and interpolation layer. The patch partition operation splits the

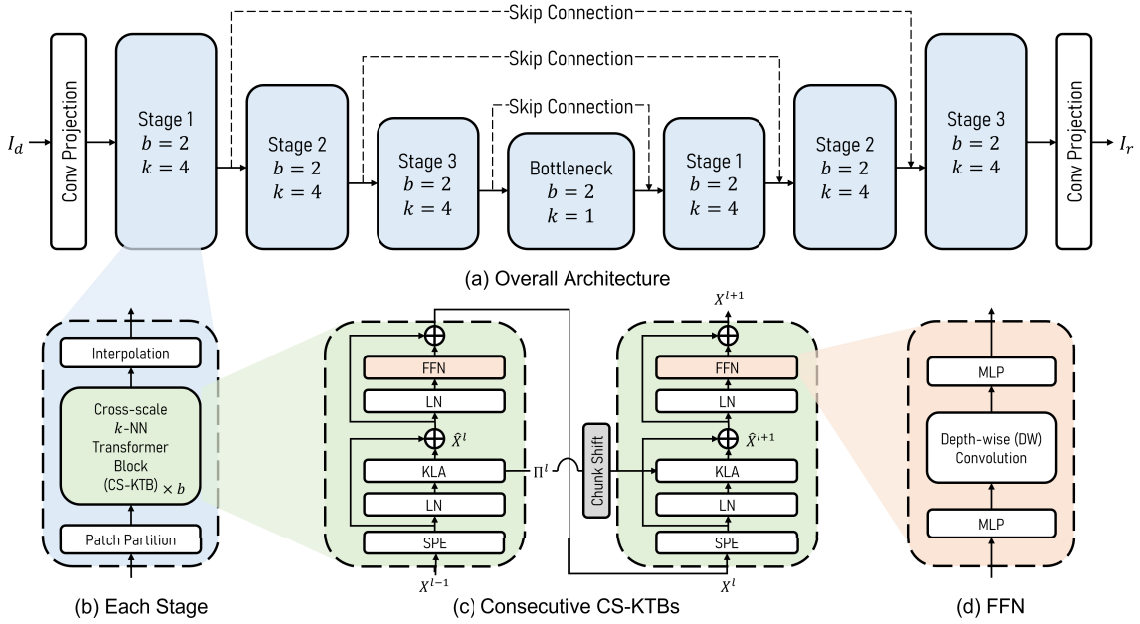


FIGURE 2. Overall architecture of the cross-scale k-NN Image Transformer (CS-KiT): (a) U-shaped hierarchical architecture is adopted for image restoration. (b) Each stage has two cross-scale k-NN Transformer Blocks (CS-KTBs), i.e., $b = 2$, and an interpolation layer. For a skip-connection, the output feature of i-th stage in the encoder network is concatenated with $(4 - i)$-th stage in the decoder network. (c) The CS-KTB consists of scale-aware patch embedding (SPE), layer normalization (LN), and k-NN local attention (KLA). Note that (c) shows the CS-KTBs when $b = 2$, and $(l + 1)$-th CS-KTB takes shifted patch indexes Π^l_{shift}. By leveraging the shifted index of the previous block, the KLA of the current block deals with the isolated patch problem described in Section IV-C. (d) feed-forward network (FFN) consists of depth-wise convolution (DW) and multi-layer perceptron (MLP).

input feature map X into non-overlapping patches with the patch size r, satisfying $X = \{x_i \in \mathbb{R}^{r^2 \times C} \mid i = 1, \ldots, N\}$.

In the CS-KTB, scale-aware patch embedding (SPE) is first conducted to introduce cross-scale aggregation into attention. SPE projects each patch into scale-specific spaces through convolutions of different kernel sizes, separately. We then estimate the scale score α with a learnable scale prediction function in each scale-specific branch. A representative scale of the patch is then defined as the weighted sum of scale-specific patches leveraging α as weights. The mixed-scale patches projected on different representative scales are normalized and fed into k-NN local attention (KLA).

KLA first clusters similar patches with locality sensitive hashing (LSH) [31] which is an approximated k-NN search algorithm. Assigned hash values by LSH stand for similarity; highly correlated patches have the same hash value with a high probability. The hash buckets, consisting of patches with the same hash value, mostly have non-uniform distribution, making it hard to aggregate patches within hash buckets in parallel. Therefore, we sort patches according to hash values and partition them into chunk sizes of k. Accordingly, each chunk may contain isolated patches that have a different hash value from most patches in a single chunk, degenerating non-local connectivity with its k-nearest neighbors. To deal with this problem, KiT [32] allows the previous chunks to contribute to the current chunk containing the query patch, but enlarging the attending chunk for attention causes an extra computational burden. As a more efficient way, we propose a chunk shift with the assumption that the k-NN relation is similar in an adjacent block. In chunk shift, the patch indexes of the current block are shared in the successive block and shifted so that isolated patches can be moved to the next chunk. By allowing the previous chunk to attend to the current chunk via chunk shift, the isolated patch issue is effectively resolved with no extra computation. Subsequently, KLA treats each chunk as a grouped local patch by rearranging and conducts local attention on each chunk. Note that, since each chunk includes patches of different positions and different scales, performing simple local attention on it achieves non-local aggregation while maintaining its lightweight computational complexity. At the end of CS-KTB, aggregated features are resized with an interpolation layer to build hierarchical architecture. When the input feature map passes through all stages of the network, the last three convolutions are conducted to predict residuals between the clean image and the degraded image.

The proposed network has a hierarchical U-shaped design for considering patterns at various scales. The aggregated features are processed through a layer of interpolation (downsampling for the encoder and upsampling for the decoder). For the purpose of restoring fine details, input feature maps and corresponding encoder features are concatenated in each stage of the decoder. Three convolutions are performed at the

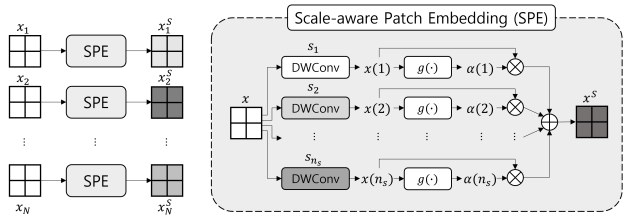


FIGURE 3. The proposed scale-aware patch embedding (SPE): Each patch is projected into spaces of different scales $S = \{s_1, \ldots, s_{n_s}\}$ by depth-wise (DW) convolutions with different kernel sizes. In each scale space, we predict the scale score of a patch, α, with learnable scale prediction function $g(\cdot)$. Finally, scale-specific patches are merged by leveraging α as weight.

end of the network to predict a restored image from the output feature map.

B. CS-KTB: CROSS-SCALE k-NN TRANSFORMER BLOCK

In l-th cross-scale k-NN Transformer Block (CS-KTB) ($l = 1, \ldots, b$), scale-aware patch embedding (SPE) is added in the front. After SPE projects patches to mixed-scale space, layer normalization is applied, followed by KLA. The intermediate feature X aggregated by KLA is passed through the rest steps, applying layer normalization and feed-forward network. Formally, CS-KTB is represented as:

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

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

where $\text{FFN}(X) = \text{MLP}(\text{DW}(\text{MLP}(X)))$. In l-th block of each stage ($l = 1, \ldots, b$), 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 the same as the CS-KTB except that no interpolation layer is employed and k is set to 1.

1) SPE: SCALE-AWARE PATCH EMBEDDING

The goal of scale-aware patch embedding (SPE) is to generate patches of mixed-scale as each patch embraces different scale information. Fig. 3 describes a detailed flow of scale-aware patch embedding. To build mixed-scale patches, we conduct depth-wise convolution operations with different kernel sizes $S = \{s_1, \ldots, s_{n_s}\}$, where n_s is the number of scale spaces that equals to the number of kernels. Embedded scale-specific patch is defined as $\{x_{i,j} \mid i = 1, \ldots, N \text{ and } j = 1, \ldots, n_s\}$, which are responses of different scales S. For cross-scale aggregation, naively aggregating all patches of various scales is computationally heavy. In a more efficient way, we assume that a single patch has a representative scale, similar to image descriptors such as SIFT [72]. With pre-defined discrete scales, we generate a mixed-scale patch with a weighted combination. Specifically, we define a learnable scale prediction function g: $\mathbb{R}^{r^2 \times C \to \mathbb{R}^1}$ which estimates the scale score of an input patch. The estimated scale scores from $g(\cdot)$ are used as weights for merging scale-specific patches:

$$\alpha_i = \delta([g(x_{i,1}), g(x_{i,2}), \ldots, g(x_{i,n_s})]), \tag{5}$$

where δ is the softmax operation for normalizing each weight. Finally, scale-aware patch embedding $X^S = \{x_1^S, \ldots, x_N^S\}$ can be obtained by

$$x_i^S = \sum_{j=1}^{n_s} \alpha_{i,j} \cdot x_{i,j}. \tag{6}$$

2) k-NEAREST NEIGHBOR SEARCH

To figure k-NN patches out, a brute-force approach computes the pair-wise distances between all patches. As this pair-wise distance involves the quadratic complexity to an input length, extensive works on k-NN search have been developed to reduce its cost [31], [73], [74]. We adopt locality sensitive hashing (LSH) [31] for k-NN search due to its linear computational complexity. LSH is an approximated k-NN search algorithm that hashes similar elements into the same buckets by using random rotation matrices. Here, the number of buckets is much smaller than the whole number of elements to ensure each bucket contains multiple elements.

In CS-KTB, in order to build buckets, the LSH projects divided patches into a unit hyper-sphere. Assuming there are m hash buckets, a hash value $L(x^S)$ is assigned by multiplying random rotation matrix $R \in \mathbb{R}^{N \times m/2}$ to a mixed-scale patch x^S as:

$$L(x^S) = arg\max([x^S R; -x^S R]), \tag{7}$$

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. But, similar patches may fall in different hash buckets in times as LSH relies on a random rotation matrix. Multi-round LSH, in which LSH is applied with multiple different rotation matrices h times, is employed to cope with this problem.

C. KLA: k-NN LOCAL ATTENTION

k-NN local attention (KLA) aims at achieving locality, non-locality, and cross-scale aggregation at the same time. As shwon in Fig. 4, KLA first rearranges patches so that similar patches are located around each other. As patches embrace mixed-scale information through SPE, applying local attention to the rearranged mixed-scale patches implicitly induces cross-scale aggregation. In addition, we partition patch sequences with a chunk size of k for efficient parallel computation in aggregation. However, as hash buckets have non-uniform distribution in practice, the patches with the same hash value may fall into adjacent buckets, resulting in isolated patches as shown in Fig. 5. To handle the problem of isolated patches, we propose a chunk shift that reuses sorted patch indexes of the current block in the successive block and shifts them.

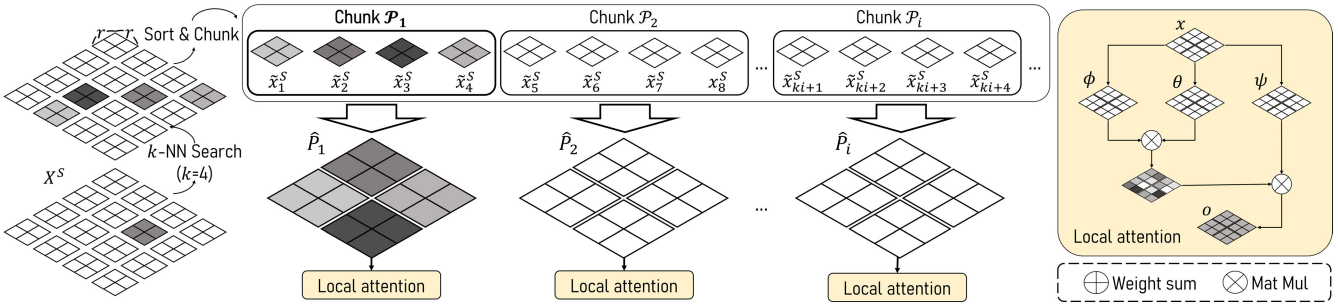


FIGURE 4. The proposed k-NN local attention (KLA): As an approximated k-NN search, LSH finds similarity of patches by assigning hash value to each patch with random rotation matrix R. After LSH assigns hash values to patches, patches are sorted by the hash values and partitioned with a chunk size of k. Then, k patches in the same chunk are spatially rearranged to form a single patch. Finally, KLA conducts local attention on each rearranged patch for cross-scale non-local aggregation.

1) LOCAL ATTENTION WITH SPATIAL REARRANGE

To restrict the contribution of patches to a query patch to those with the same hash value, we first sort patches based on hash values, and then divide the sorted patches into chunks each involving k patches (equivalent to the number of NN patches) for batching purpose, so that only patches in the same chunk are considered in the local attention. We denote $\pi : n \to n$ be a permutation that sorts the patches in ascending order of hash values:

$$\pi(p) < \pi(q) \Rightarrow L(x_p^S) \leq L(x_q^S). \tag{8}$$

For the sake of a simplicity, we define $\tilde{x}^S$ as a sorted mixed-scale patch where $\tilde{x}_p^S$ is equal to $x_{\pi(p)}^S$. Then, i-th chunk $\mathcal{P}_i$ for $i = 1, \ldots, N/k$ contains k patches,

$$\mathcal{P}_i = \{\tilde{x}_{ki+1}^S, \tilde{x}_{ki+2}^S, \tilde{x}_{ki+3}^S, \ldots, \tilde{x}_{ki+k}^S\}. \tag{9}$$

KLA performs local attention by regarding each chunk as a grouped local patch. In each chunk, there are k patches with size $\mathbb{R}^{k\times r^2\times C}$. We make each chunk to 2-dimensional patch $\hat{\mathcal{P}}_i = \Phi(\mathcal{P}_i)$, where $\Phi : \mathbb{R}^{k\times r^2\times C} \to \mathbb{R}^{kr^2\times C}$ is the spatial rearrange function.

As rearranged patches $\hat{\mathcal{P}}$ are mixed-scale and spatially grouped, aggregating cross-scale and cross-position patches can be done with simple local attention:

$$o_i = softmax(\frac{\phi(\hat{\mathcal{P}}_i)\theta(\hat{\mathcal{P}}_i)^{\mathsf{T}}}{\sqrt{C}})\psi(\hat{\mathcal{P}}_i), \tag{10}$$

where $o_i \in \mathbb{R}^{kr^2\times C}$ represents the i-th output patch of KLA, and ϕ, θ, and $\psi : \mathbb{R}^{kr^2\times C} \to \mathbb{R}^{kr^2\times C}$ are learnable projection functions.

2) CHUNK SHIFT

Chunk patch sequences with a size of k enable an efficient batch-wise parallel computation. However, 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, incurring isolated patches as shown in Fig. 5. The isolated patches that have a different hash value from major patches in the chunk may weaken the non-local connectivity of the KLA. We propose a chunk shift, an efficient way to deal with the isolated patch problem with no increase in computation. We denote Π^l as indexes of patches sorted by hash values in a l-th block. Shifted indexes Π^l_{shift} are defined by cycling shift Π^l with a size of $k/2$,

$$\Pi^l_{shift} = (\Pi^l + \frac{k}{2}) \bmod N. \tag{11}$$

We form consecutive CS-KTBs such that shifted indexes are utilized in the successive block ($\Pi^{l+1} = \Pi^l_{shift}$) under the assumption that the feature distribution of adjacent blocks would be similar. As shown in Fig. 2 (c), k-NN local attention (KLA) and KLA with shifted chunk are alternately conducted in the network. This configuration enables us to deal with the isolated patches problem and save computation of k-NN search.

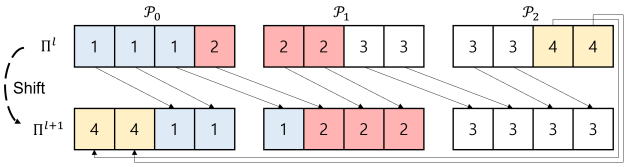


FIGURE 5. The proposed chunk shift: as the distribution of hash values is non-uniform, the number of patches in each bucket estimated by LSH is often different from a chunk size k. In l-th block, sorted index Π^l may contain isolated patches that have different hash values from patches in the chunk. By shifting chunks and reusing them in the next $(l+1)$-th block, we can tackle this problem efficiently with no computation increase. A number in each patch represents a hash value assigned.

D. TRAINING LOSS

Following existing image restoration approaches [7], [17], 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 [75] $\mathcal{L}_{char}$ and an edge loss $\mathcal{L}_{edge}$ for optimizing the network,

$$\begin{aligned} \mathcal{L}_{char} &= \sqrt{\|I - (I_d + I_r)\|^2 + \epsilon^2}, \\ \mathcal{L}_{edge} &= \sqrt{\|\triangle I - \triangle(I_d + I_r)\|^2 + \epsilon^2}, \\ \mathcal{L} &= \mathcal{L}_{char} + \lambda\mathcal{L}_{edge}. \end{aligned} \tag{12}$$

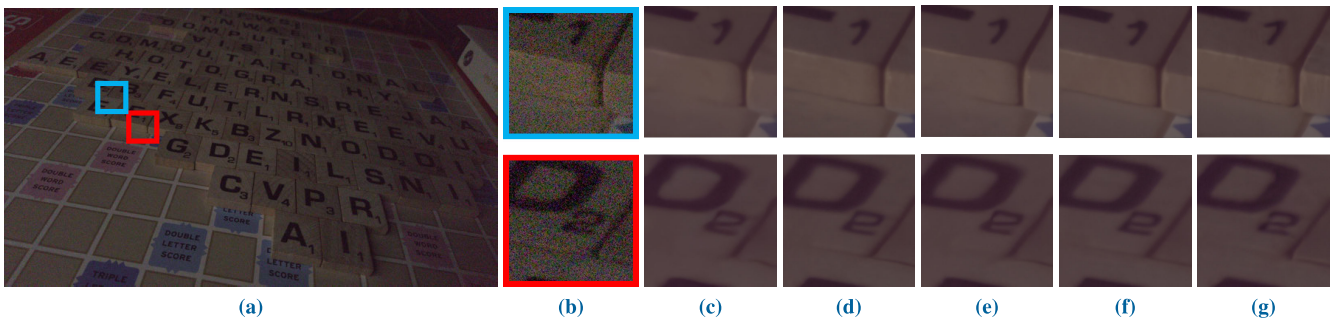


FIGURE 6. Visual comparisons on the SIDD [76] dataset: (a) Noisy input, (b) Cropped image, (c) CycleISP [77], (d) MPRNet [52], (e) Uformer [23], (f) KiT, and (g) CS-KiT.

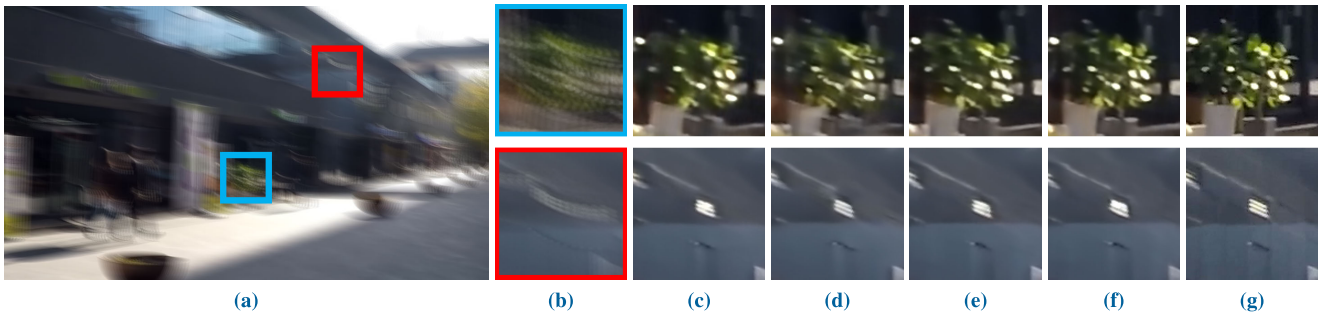


FIGURE 7. Visual comparisons on the GoPro [49] dataset: (a) Blurry image, (b) Cropped image [51], (c) CycleISP [77], (d) MPRNet [52], (e) Uformer [23], (f) KiT, and (g) CS-KiT.

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.

E. EXPERIMENTS

We used the AdamW optimizer to train the entire network using batches of 64 images cropped to 128×128 for 800k iterations. The learning rate was initially set at 2×10^{-4}, and until 10^{-6}, the linear warm-up strategy and cosine annealing were applied to decrease the learning rate. 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 was set to 1 since there are only a few patches (e.g. the number of patches is 8×8 when HW is 256×256). The number of scale spaces n_s was set to 3 and the kernel size of each depth-wise convolution was 3, 5, and 7, respectively. The number of CS-KTB in each stage b was set to 2 in all stages. 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, deblurring 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 [52], [78].

1) IMAGE DENOISING

We trained the CS-KiT with the SIDD [76] dataset consisting of 320 high-resolution real noisy images. Table 1 shows the quantitative results of image denoising methods on the SIDD [76] and DND [79] dataset. The evaluation results include the classical denoisng method [14], CNN-based methods [7], [50], [51], [52], [77], [80], [81], [82], [85], [89], self-attention based methods [44], transformer-based

TABLE 1. Denoising results on the SIDD [76] and DND [79] dataset. The bold and underlined numbers indicate the best and the second best results, respectively. The higher, the better for the PSNR and SSIM.

Method	SIDD		DND	
	PSNR	SSIM	PSNR	SSIM
BM3D [14]	25.65	0.685	34.51	0.851
DnCNN [7]	23.66	0.583	32.43	0.790
MLP [80]	24.71	0.641	34.23	0.833
CBDNet [50]	30.78	0.801	38.06	0.942
RIDNet [51]	38.71	0.951	39.26	0.953
AINDNet [81]	38.95	0.952	39.37	0.951
VDN [82]	39.28	0.956	39.38	0.952
SADNet [83]	39.46	0.957	39.59	0.952
DANet [84]	39.47	0.957	39.58	0.955
CycleISP [77]	39.52	0.957	39.56	<u>0.956</u>
MPRNet [52]	39.71	0.958	39.80	0.954
MIRNet [85]	39.72	0.958	39.88	<u>0.956</u>
DAGL [44]	-	-	39.83	**0.957**
Uformer [23]	39.77	0.959	<u>39.96</u>	<u>0.956</u>
KiT [32]	<u>39.80</u>	<u>0.959</u>	<u>39.96</u>	<u>0.956</u>
CS-KiT	**39.87**	**0.960**	**39.99**	<u>0.956</u>

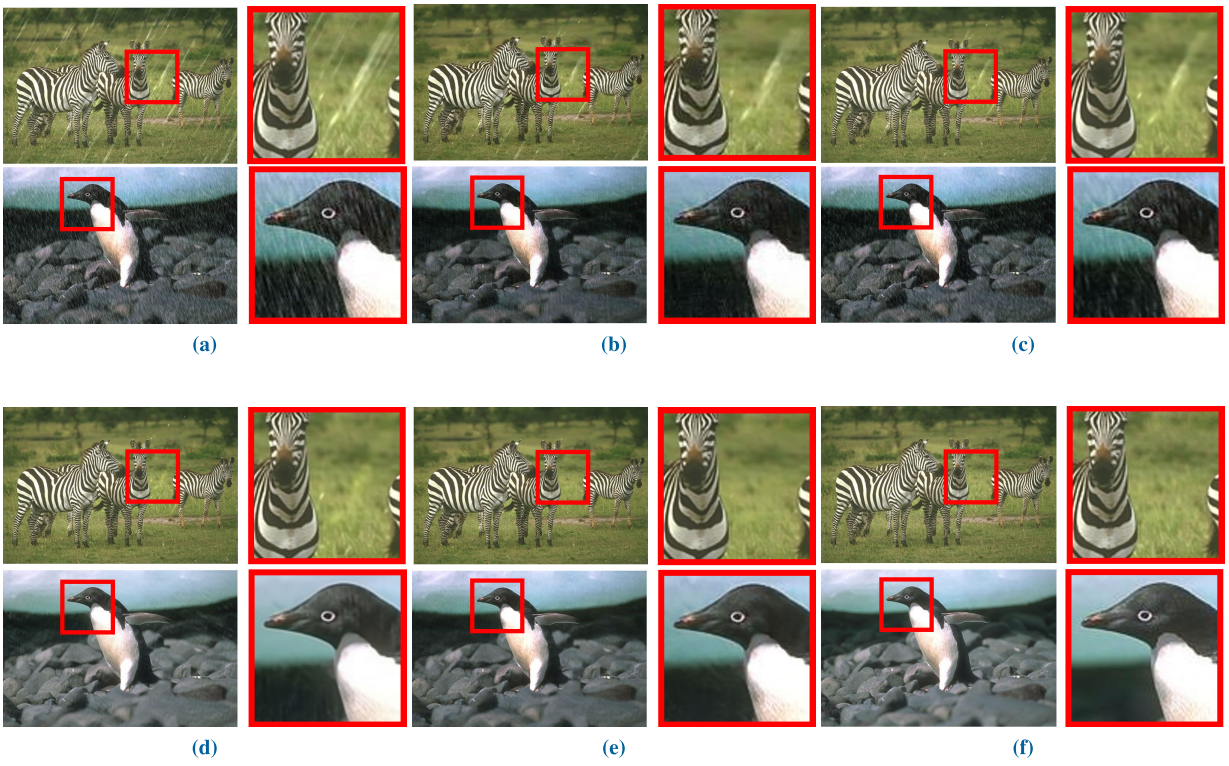


FIGURE 8. Visual comparisons on the Test100 [86] and Rain100L [87] dataset: (a) Rainy input, (b) RESCAN [88], (c) MPRNet [52], (d) KiT, (e) CS-KiT, and (f) ground truth.

methods [23], and our previous work [32]. As DND [79] dataset does not provide ground-truth labels, the results were obtained from the official benchmark. CS-KiT performs favorably against existing image denoising approaches. Specifically, compared to the previous winning method, Uformer, CS-KiT gets 0.1 dB gains in the SIDD dataset and 0.03dB in the DND dataset. Fig. 6 shows a visual comparison of the proposed method with previous algorithms. While most of the other methods failed to restore the exact number from noise, the results of CS-KiT correctly recovered the shape of the number, proving effectiveness in noise removal.

2) IMAGE DEBLURRING

We compared the state-of-the-art methods in image deblurring on GoPro [49] and HIDE [90] datasets. 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 [49] dataset were used, and 1,111 images of the GoPro [49] and 2,025 images of the HIDE [90] datasets were evaluated. Table 2 shows that the proposed method achieves state-of-the-art results for both GoPro and HIDE dataset. Compared to the previous best method, MIMO-UNet [95], while our previous work achieved a slight improvement of 0.02 dB in PSNR, CS-KiT widen this gap to 0.19dB, showing remarkable performance. This pattern also appears in visual comparison as shown in Fig. 7. Although the results of Uformer [23] and KiT restored the shape of objects, they are still blurry and missing high-frequency details, whereas CS-KiT restores more sharp textures than other competing methods.

TABLE 2. Deblurring results on the GoPro [49] and HIDE [90] dataset. The network was trained on GoPro dataset.

Method	GoPro		HIDE	
	PSNR	SSIM	PSNR	SSIM
DeepDeblur [49]	29.23	0.916	25.73	0.874
SRN [91]	30.26	0.934	28.36	0.915
PSS-NSC [92]	30.92	0.942	29.11	0.913
DMPHN [53]	31.20	0.945	29.09	0.924
SAPHN [93]	32.02	0.953	29.98	0.930
MT-RNN [94]	31.15	0.945	29.15	0.918
MPRNet [52]	32.66	0.959	30.96	0.939
MIMO-UNet [95]	32.68	0.959	-	-
KiT [32]	32.70	0.959	30.98	0.942
CS-KiT	**32.87**	**0.961**	**31.02**	**0.945**

3) IMAGE DERAINING

We applied CS-KiT to image deraining on five representative datasets, Test100 [86], Rain100H [87], Rain100L [87], Test2800 [96], and Test1200 [97]. Following the experimental setup of [78], 13,712 clean-rain image pairs sampled from multiple datasets [86], [87], [96], [97], [102] were used to train the network for image deraining. Fig. 8 depicts

TABLE 3. Deraining results of CS-KiT. The widely used five datasets [86], [87], [96], [97] are used for evaluation.

Method	Test100 [86]		Rain100H [87]		Rain100L [87]		Test2800 [96]		Test1200 [97]		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DerainNet [98]	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 [99]	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 [97]	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 [100]	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 [88]	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 [54]	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 [78]	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 [52]	30.27	0.897	30.41	0.890	36.40	0.965	33.64	0.938	32.91	0.916	32.73	0.921
SPAIR [101]	30.35	0.909	**30.95**	0.892	36.93	0.969	33.34	0.936	**33.04**	**0.922**	32.91	0.926
KiT [32]	30.26	0.908	30.47	0.897	36.65	0.969	33.85	0.941	32.81	0.918	32.81	0.927
CS-KiT	**30.41**	**0.911**	30.91	**0.900**	**37.17**	**0.971**	**33.93**	**0.942**	33.01	0.921	**33.05**	**0.929**

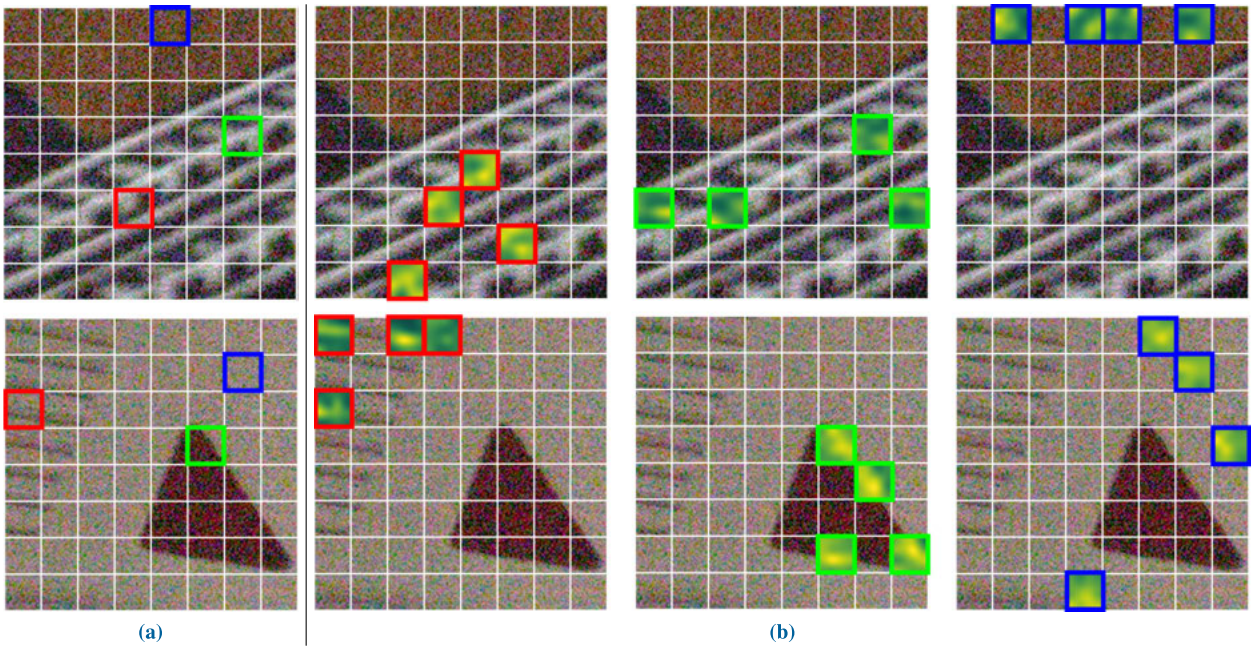


FIGURE 9. Visualization of the k-NN local attention: (a) input image and (b) k-NN patches. k-NN patches are discovered by LSH with $k = 4$. Patches belonging to the same chunk are marked with boxes of the same color. We also provide a visualization of learned attention produced by dot products between the center pixel of the query patch and k-NN patches.

visual results of rain streak removal compared to the previous works, PreNet [54], RESCAN [88], and MPRNet [52]. Under strong rain streak conditions, KiT successfully restores a clean image and outperforms the state-of-the-art methods. Moreover, CS-KiT is even better than KiT noticeably, clearly removing the rain streak. Table 3 shows that CS-KiT is superior not only qualitatively but also quantitatively. While our previous work, KiT, is competitive with SPAIR [101], CS-KiT surpasses SPAIR in both PSNR and SSIM.

F. ABLATION STUDY

We conducted the ablation study to analyze the effectiveness of our method in various aspects. All experiments were conducted on SIDD [76] for the image denoising task.

1) VISUALIZATION OF THE k-NN PATCHES

Our method aims to preserve fine details while achieving non-local connectivity efficiently, achieved by aggregating patches of different scales with similar characteristics. To visually validate this, we further visualize the patches belonging to the same chunk in Fig. 9. The leftmost images are divided into non-overlapping patches, where the patches marked with color boxes represent query patches for visualization. Similar k patches are clustered with a chunk in the right figures as the KLA utilizes LSH for k-NN search. The same color boxes serve to denote the k patches belonging to the same chunk. The patches with blue boxes have non-textured regions, while the patches with red and green boxes have similar patterns. This shows that the LSH effectively

TABLE 4. Ablation study of the chunk size k and hash rounds h. The PSNR is measured on the SIDD dataset for image denoising.

PSNR		h		
		1	4	16
k	1	38.96	39.04	39.12
	2	39.67	39.69	39.83
	4	39.85	39.87	39.87
	8	39.86	39.87	39.87
	16	39.87	39.88	39.88

finds visually analogous patches. We also provide a visualization of learned attention produced by dot products between the center pixel of the query patch and k-NN patches.

2) THE NUMBER OF k AND h

In CS-KiT, we also validated two hyper-parameters: the chunk size k and the number of hash rounds h. k determines the maximum number of patches used for performing the local attention and h is used to reduce the probability that similar patches fall into different hash buckets. These two scalable hyper-parameters make the trade-off between the computational complexity and the network capacity. Table 4 shows the denoising performance of the proposed method according to the two hyper-parameters. Similar to KiT, 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.

3) COMPUTATIONAL COST

We provide the performance comparison with state-of-the-art image restoration methods with respect to accuracy and computational cost. Fig. 10 depicts the graphs that illustrate the performance and computational cost of state-of-the-art methods. Other approaches are denoted by a circle symbol in green, our previous work is denoted by a circle symbol in red, and the proposed CS-KiT is denoted by a triangle symbol in red. The x-axis and y-axis of the graphs, respectively, indicate the performance evaluated by the PSNR and computational cost measured by Multiply-Accumulates (MACs). The MACs of all graphs are measured when an input resolution is 256×256. The proposed method outperforms Uformer [23] despite having a competitive computational cost in the image denoising on the SIDD dataset [76]. In the image deraining and deblurring, the KiT shows a slightly better performance yet with much less computational cost and the CS-KiT further improves performance.

4) CHUNK SHIFT

When the number of patches is indivisible with the size of chunks, each chunk may contain isolated patches having a different hash value from other patches of the same chunk as shown in Fig. 5. We deal with this problem by 1) *sharing* patch indexes in the successive block and 2) *shifting* shared patch indexes to bridge a connection between adjacent chunks. Table 5 shows the denoising results on the SSID dataset according to chunk shifting and sharing. When chunk indexes are only shared in the successive block and not shifted, no performance drop was aroused. It implies that k-NN relations of adjacent chunks are almost similar to each other. In addition, applying both shifting and sharing chunk indexes in the successive block resulted in a slight increase in performance, while the computational cost was slightly reduced due to the omission of the LSH in the successive block. By leveraging the chunk shift to successive block, connectivity between similar patches belonging to different buckets was established without extra computations.

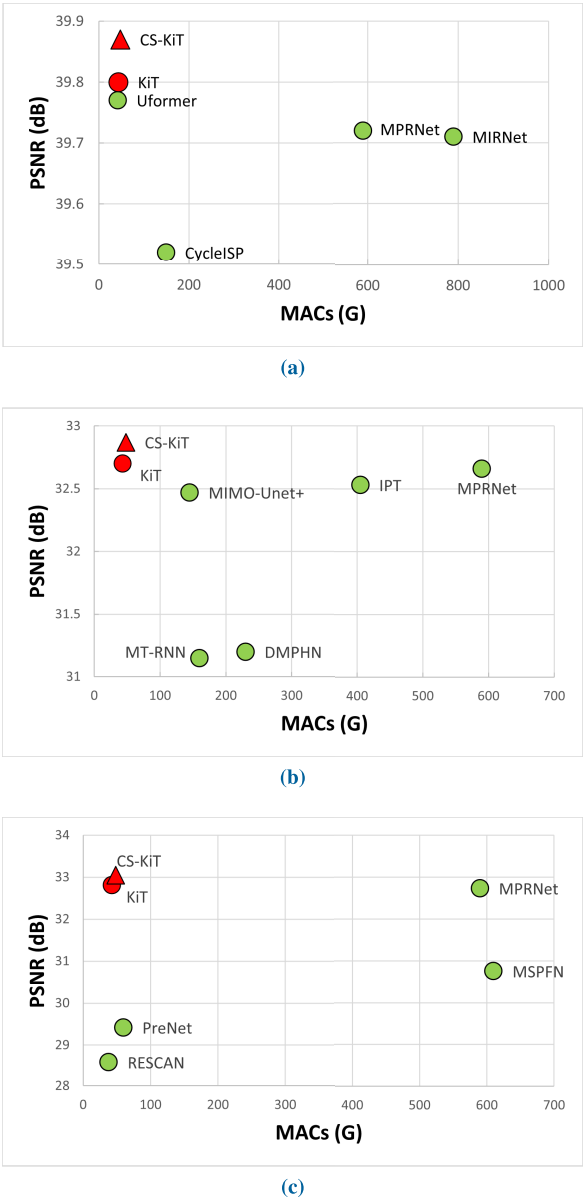


FIGURE 10. Performance *vs.* computational cost of state-of-the-art methods for the image restoration tasks: (a) image denoising, (b) image deblurring, and (c) image deraining.

TABLE 5. Effectiveness of the chunk shift.

Method	Chunk indexes		SIDD	
	Shift	Share	PSNR	SSIM
CS-KiT			39.84	0.960
		✓	39.84	0.960
	✓	✓	39.87	0.960

TABLE 6. Comparison between results of soft and hard scale prediction.

Method	Scale prediction	SIDD	
		PSNR	SSIM
KiT [32]	None	39.80	0.959
CS-KiT	gumbel softmax (hard)	39.83	0.959
	softmax (soft)	39.87	0.960

5) SCALE PREDICTION

Scale-aware patch embedding assumes that each patch has a representative scale in continuous space. Thus, softmax was adopted to estimate continuous scale when merging scale-specific scores. We compared the performance between soft scale estimation (softmax) and hard scale estimation (Gumbel softmax [103]) in Table 6. Soft prediction achieves better performance than hard prediction, which supports our assumption that the representative scale estimator should yield a continuous value. Moreover, even the cross-scale aggregation with the Gumbel softmax surpasses our previous work [32] which aggregates the patches of the same scale only, implying that the cross-scale aggregation is essential in image restoration.

V. CONCLUSION

We presented a transformer-based image restoration network, a cross-scale k-NN image transformer (CS-KiT), that meets essential conditions: locality, non-locality, and cross-scale aggregation through a novel attention mechanism, k-NN local attention (KLA). The core idea of KLA is to group similar patches in the whole image and conduct local attention to spatially grouped patches. To handle the quadratic computational complexity of brute-force k-NN search, we adopt locality sensitive hashing (LSH) which is approximated linear k-NN method. In addition, scale-aware patch embedding projects each patch to different scales to form mixed-scale patches. By feeding mixed-scale patches into a transformer block, cross-scale aggregation is carried out while conducting self-attention. Chunk shift handles the problem of isolated patches that occur when the patch sequence is indivisible by the size of the chunks. By sharing and shifting patch indexes in the successive block, the KLA enhances non-locality while saving k-NN computations. We demonstrated that the proposed CS-KiT achieved superior performance to the state-of-the-art methods on various image restoration benchmarks, in terms of quantitative/qualitative performance.

A. FUTURE WORKS

Due to the lack of inductive bias, transformer-based approaches typically require more training data than CNN counterparts. In visual recognition tasks, a self-supervised pre-training with a large-scale dataset (ImageNet) shows significant improvements compared to training from scratch. As not many datasets exist in image restoration, a few works of pre-training strategy have been investigated to solve the data-hungry issue. This has been partly addressed by pre-training the networks with synthetic degradation such as Gaussian noise or rain streak, but two limitations still remain unresolved. First, The domain gap between synthetic and real degradation makes it less effective when transferring the pre-trained network to the downstream task. Second, most works pre-train separate networks for different tasks. On account of this complex pre-training stage, the pre-training step has not been widely used in image restoration. In future work, we will continue to investigate the pre-training strategy leveraging real-world degradation datasets and the unified image restoration model across different degradation factors.

REFERENCES

[1] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3642–3649.
[2] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[4] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.
[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2016.
[7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 184–199.
[9] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2802–2810.
[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[12] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.
[13] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 60–65.
[14] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[15] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[16] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," 2019, *arXiv:1903.10082*.

[17] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," 2018, *arXiv:1806.02919*.

[18] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[20] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021, *arXiv:2102.12122*.

[21] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[23] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," 2021, *arXiv:2106.03106*.

[24] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," 2021, *arXiv:2107.00641*.

[25] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," 2021, *arXiv:2103.15808*.

[26] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "LocalViT: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*.

[27] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12894–12904.

[28] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision outlooker for visual recognition," 2021, *arXiv:2106.13112*.

[29] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[30] W. Wang, L. Yao, L. Chen, B. Lin, D. Cai, X. He, and W. Liu, "CrossFormer: A versatile vision transformer hinging on cross-scale attention," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–15.

[31] A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, and L. Schmidt, "Practical and optimal LSH for angular distance," 2015, *arXiv:1509.02897*.

[32] H. Lee, H. Choi, K. Sohn, and D. Min, "KNN local attention for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2139–2149.

[33] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3929–3938.

[34] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[35] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[36] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, Aug. 2017.

[37] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4539–4547.

[38] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[39] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3155–3164.

[40] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[41] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2272–2279.

[42] T. Plötz and S. Roth, "Neural nearest neighbors networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1087–1098.

[43] D. Valsesia, G. Fracastoro, and E. Magli, "Deep graph-convolutional image denoising," *IEEE Trans. Image Process.*, vol. 29, pp. 8226–8237, 2020.

[44] C. Mou, J. Zhang, and Z. Wu, "Dynamic attentive graph learning for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4328–4337.

[45] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3499–3509.

[46] Y. Li, X. Fu, and Z.-J. Zha, "Cross-patch graph convolutional network for image denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4651–4660.

[47] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3517–3526.

[48] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8878–8887.

[49] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.

[50] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1712–1722.

[51] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3155–3164.

[52] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14821–14831.

[53] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5978–5986.

[54] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3937–3946.

[55] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5690–5699.

[56] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5791–5800.

[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[58] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

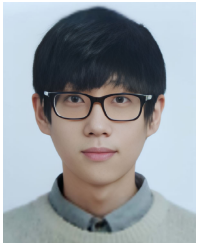
[59] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 17, 2022, doi: 10.1109/TPAMI.2022.3222871.

[60] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5281–5292, Aug. 2022.

[61] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

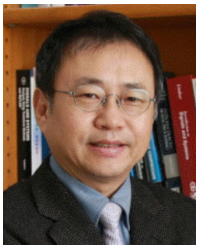
[62] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[63] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[64] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13065–13074.
[65] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[66] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[68] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.
[69] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*.
[70] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10853–10862.
[71] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7287–7296.
[72] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.
[73] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2010.
[74] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.
[75] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2, Sep. 1994, pp. 168–172.
[76] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1692–1700.
[77] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "CycleISP: Real image restoration via improved data synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2696–2705.
[78] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, and J. Jiang, "Multi-scale progressive fusion network for single image deraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8346–8355.
[79] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1586–1595.
[80] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2392–2399.
[81] Y. Kim, J. W. Soh, G. Y. Park, and N. I. Cho, "Transfer learning from synthetic to real-noise denoising with adaptive instance normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3482–3492.
[82] Z. Yue, H. Yong, Q. Zhao, L. Zhang, and D. Meng, "Variational denoising network: Toward blind noise modeling and removal," 2019, *arXiv:1908.11314*.
[83] M. Chang, Q. Li, H. Feng, and Z. Xu, "Spatial-adaptive network for single image denoising," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 171–187.
[84] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 41–58.
[85] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2020, pp. 492–511.
[86] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3943–3956, Nov. 2020.
[87] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1357–1366.
[88] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 254–269.
[89] S. Cheng, Y. Wang, H. Huang, D. Liu, H. Fan, and S. Liu, "NBNet: Noise basis learning for image denoising with subspace projection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4896–4906.
[90] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5572–5581.
[91] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8174–8182.
[92] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3848–3856.
[93] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3606–3615.
[94] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 327–343.
[95] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4641–4650.
[96] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3855–3863.
[97] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 695–704.
[98] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.
[99] W. Wei, D. Meng, Q. Zhao, Z. Xu, and Y. Wu, "Semi-supervised transfer learning for image rain removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3877–3886.
[100] R. Yasarla and V. M. Patel, "Uncertainty guided multi-scale residual learning-using a cycle spinning CNN for single image de-raining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8405–8414.
[101] K. Purohit, M. Suin, A. N. Rajagopalan, and V. N. Boddeti, "Spatially-adaptive image restoration using distortion-guided networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2309–2319.
[102] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, "Rain streak removal using layer priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2736–2744.
[103] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," 2016, *arXiv:1611.01144*.



HUNSANG LEE (Student Member, IEEE) received the B.S. and M.S. degrees from the Department of Computer Science and Engineering, Chungnam University, Daejeon, South Korea, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea. His current research interests include computer vision and deep learning.

HYESONG CHOI (Student Member, IEEE) received the B.S. degree from the Department of Computer Science and Engineering, Ewha Womans University, Seoul, South Korea, in 2019, where she is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering. Her current research interests include computer vision and deep learning.



KWANGHOON SOHN (Senior Member, IEEE) received the B.E. degree in electronic engineering from Yonsei University, Seoul, South Korea, in 1983, the M.S.E.E. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 1992. He was a Senior Member of the Research Engineer with the Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, South Korea, from 1992 to 1993, and a Postdoctoral Fellow with the MRI Center, Medical School of Georgetown University, Washington, DC, USA, in 1994. He was a Visiting Professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently an Underwood Distinguished Professor with the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3D image processing and computer vision.

DONGBO MIN (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2003, 2005, and 2009, respectively. From 2009 to 2010, he was a Postdoctoral Researcher with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. From 2010 to 2015, he was with the Advanced Digital Sciences Center, Singapore. From 2015 to 2018, he was an Assistant Professor with the Department of Computer Science and Engineering, Chungnam National University, Daejeon, South Korea. Since 2018, he has been with the Department of Computer Science and Engineering, Ewha Womans University, Seoul. His current research interests include computer vision, deep learning, and video processing.