iiTransformer: A Unified Approach to Exploiting Local and Non-Local Information for Image Restoration

Soo Min Kang
sm2.kang@samsung.com
Youngchan Song
yc87.song@samsung.com
Hanul Shin
hsky.shin@samsung.com
Tammy Lee
tammy.lee@samsung.com

Samsung Research
Samsung Electronics Co., Ltd.
Seoul, Republic of Korea

Abstract

The goal of image restoration is to recover a high-quality image from its degraded input. While impressive results on various image restoration tasks have been achieved using CNNs, the convolution operation has limited its ability to utilize information outside of its receptive field. Transformers, which use the self-attention mechanism to model long-range dependencies of its input, have demonstrated promising results in various high-level vision tasks. In this paper, we propose intra-inter Transformer (iiTransformer) by explicitly modelling long-range dependencies at the pixel- and patch-levels since there are benefits to considering *both* local and non-local feature correlations. In addition, we provide a boundary artifact-free solution to support images with arbitrary sizes. We demonstrate the potential of iiTransformer as a general purpose backbone architecture through extensive experiments on various image restoration tasks.

1 Introduction

Image restoration aims to recover a high-quality (HQ) image from its degraded low-quality (LQ) image. It plays a fundamental role in computer vision as its result can largely influence subsequent high-level vision tasks in recognizing and/or understanding image data. While great advancements in CNNs have shown impressive results on various image restoration tasks (e.g., image denoising), the basic building blocks of these models, *convolutions*, have shown limited ability to exploit contextual information outside of its local receptive field. They often forego repetitive data available within the image itself due to its distance.

There are benefits to exploiting data at various sub-region levels of an image. That is, utilizing data at the local level enables use of essential context information contained in the local vicinity of a degraded pixel, while using information at the non-local[1] level takes advantage of the data recurrence property in natural images [16, 47]. These local and non-local

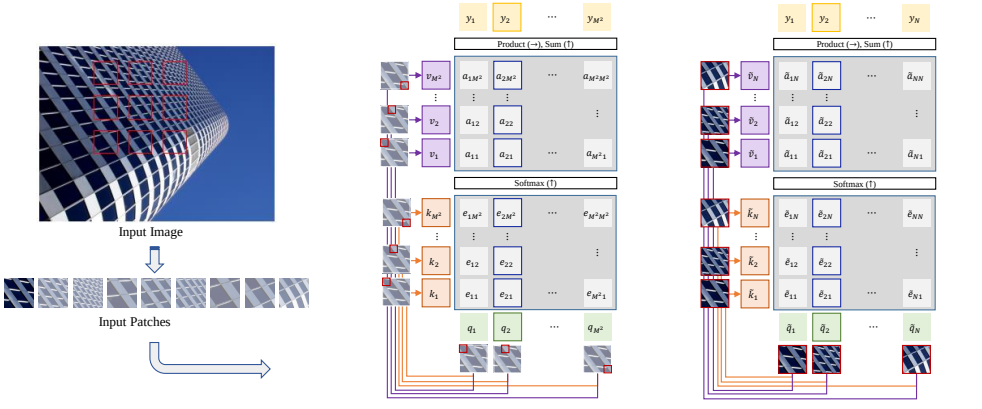


Figure 1: iiTransformer exploits local and non-local information within an image through intra and inter self-attention (SA) modules, resp. The *intra SA* module (centre) treats *pixels* as tokens (red boxes at different pixel locations from the same patch) for *local* similarity computation, while the *inter SA* module (right) treats *patches* as tokens (different patches in red) for *non-local* correlation analysis. As in a typical Transformer, each token is projected with appropriate projection matrices to obtain query (q_i), key (k_i), and value (v_i). Each query q_i is compared to all keys ($k_j \; \forall j$) to obtain a scaled dot product e_{ij}, which is normalized by the softmax function along j. The normalized similarity values, a_{ij}, are used as attention weights to compute the output y_i as the weighted sum of the values (i.e., $y_i = \sum_j a_{ij} v_j$).

relationships can be captured by considering the long-range dependencies at the pixel- or patch-level using the self-attention (SA) module of Transformers. Indeed, we refer to the SA module that treats *pixels* as tokens to compute local pixelwise correlations as *intra SA* and the SA module that treats *patches* as tokens to compute non-local patchwise correlation as *inter SA* (cf. Fig. 1). To the best of our knowledge, recent works based on Transformers for low-level vision explore either local or non-local information but not both. Furthermore, existing inter SA module-based vision Transformer models are limited to inferencing images whose resolution are consistent with its training. Consequently, patch boundary artifacts inevitably occur as shown in Fig. 2b. In this paper, we propose a unified approach that exploits both local and non-local information using intra and inter SA modules, resp. We provide a solution that enables an efficient switch between local and non-local SA mechanisms without introducing additional parameters. Also, we propose a boundary artifact-free solution to support images of arbitrary sizes when applying the inter SA mechanism (see Fig. 2c).

The main contributions of this paper are three-fold. First, we propose a unified approach based on Transformers that exploits both local and non-local information for image restoration. Second, we provide a boundary artifact-free solution for processing inference images whose resolution do not match its training image. Third, we provide extensive experiments on various image restoration tasks demonstrating state-of-the-art performance and effectiveness of the proposed approach.

[1] We avoid the use of the term 'global' to avoid the misconception that information within the *entire* image must be used, which limits its ability to operate on high-resolution images. Instead, our search range includes a small neighbourhood around the patch of interest.

(a) Noisy Input Image (LQ) (b) Presence of Boundary Artifacts (c) Absence of Boundary Artifacts (d) GT

Figure 2: To restore (a) a degraded low quality image, (b) existing inter SA module-based vision Transformers [7] require the resolution of inference image to match those of training (i.e., $(W \times H)_{test} = (W \times H)_{train}$). Thus, boundary artifacts along the $(W \times H)_{train}$ grid, marked by red pointers, is apparent. (c) In iiTransformer, the inter-patch correlations are masked for flexibility in inference size and outputs a boundary artifact-free high quality image without the need for pixel overlap processing. Effects are best viewed electronically.

2 Related Works

Image Restoration. With the availability of large-scale data, learning-based methods that learn a mapping function from low- to high-quality images have been prevalent for image restoration and have demonstrated significant performance improvement over several traditional methods. In particular, the work of DnCNN [42] for image denoising, ARCNN [11] for JPEG compression artifact removal, and SRCNN [12] for super-resolution (SR) marked a turning point for image restoration, where enormous efforts were made in developing novel architecture designs using CNNs, such as residual blocks [20, 39, 42], dense blocks [35, 45], hierarchical structures [22], channel attention [10, 31, 44], and non-local attention [10, 26, 46] mechanisms. While these explorations further enhanced the capabilities of learning-based methods, its primitive operation, *convolutions*, restricted the ability to capture long-range dependencies between pixels due to its limited receptive field. Moreover, fixing convolutional filter weights after training hindered its ability to adapt to different input contents.

Non-local. A flurry of work exploited the strong internal data recurring tendency in natural images [16, 34, 47] and non-local self-similarity based approaches flourished in solving various computer vision tasks. Non-local means [4] is a classical filtering algorithm that uses patch appearance similarity and its spatial distance to compute the weighted sum at the specified location. This non-local filtering idea later developed into CBM3D [9, 28] which applies block matching on a group of similar patches then uses 3D filtering; it continues to be a solid image denoising baseline even compared with deep neural networks. More recently, non-local neural block [38] was introduced to incorporate non-local operations into deep neural networks. Use of non-local neural blocks in image restoration based models, as in NLRN [26], SAN [10], and RNAN [46], have assisted the network to make better use of image structure cues by considering the patchwise long-range feature correlations.

Transformers. The great success of Transformers [37] in natural language processing, which models long-range dependencies in the data, has attracted much attention in the computer vision community. Starting from Vision Transformer (ViT) [13] for image classification, Transformers have been actively explored in various high-level vision tasks from recognition [13, 27], detection [5, 27], and segmentation [27], to list a few. IPT [7] is one of the first Transformer-based models in low-level vision; it dissects the image into patches to be used as tokens for patchwise attention in the Transformer, like ViT. Both ViT and IPT are limited to images of fixed resolution. Thus, directly applying patchwise attention-based

Transformer models to low-level vision tasks inevitably results in patch boundary artifacts for images with resolutions larger than the training image. To ameliorate this issue, SwinIR [24] and Uformer [40] apply pixelwise local attention. However, this design choice limits the receptive field of the model disregarding data recurring tendency in natural images.

Existing works that simultaneously consider local information using convolutions and non-local information via non-local blocks as in RNAN [46] or MLP blocks as in MAXIM [36] suffer from convolution's inability to adapt to different inputs due to fixed filter weights or necessitate a multi-stage framework for training stability. As self-attention is a core-component of Transformers, Transformer-based models are able to dynamically change its response based on the input as opposed to CNN-based models, demonstrating its superiority in recent image restoration works. In this work, we propose a novel Transformer-based network, *iiTransformer*, capable of achieving state-of-the-art results without bells and whistles (e.g., multi-stage framework or an advanced loss function) by utilizing local and non-local information using Transformers.

3 Methodology

As convolutional filter weights are fixed upon completion of training, the same weights are shared over the entire image space, not adapting to different inputs. Self-attention (SA), on the other hand, is calculated by computing the weighted sum of the features at other positions with respect to the query; thus, the response changes dynamically based on the input. Furthermore, SA is able to capture long-range dependencies between features by *directly* computing the interactions at any two positions rather than stacking multiple convolutional layers to enlarge its receptive field. Since self-attention is a core component of Transformers, the use of Transformers is a natural choice for obtaining flexible and expressive features pertaining to the input and for modelling long-range dependencies. Long-range dependencies can be applied at the pixel-level (i.e., locally) to capture essential context information within the local neighbourhood of the degraded pixel or at the patch-level (i.e., non-locally) to model patch recurrence in natural images. To this end, we propose iiTransformer – a unified approach that exploits both local and non-local information using intra- and inter multi-head self-attention modules, resp. In this section, we present the overall architecture of iiTransformer, followed by detailed description on the key components of the framework.

3.1 Overall Framework

Given a degraded image of low quality (LQ), the goal of image restoration is to recover a high quality (HQ) image. iiTransformer recovers the HQ image from its LQ counterpart through the following three modules: shallow feature extraction, deep feature extraction, and image reconstruction, as illustrated in Fig. 3a. The *shallow feature extraction* stage uses a 3×3 convolution to project the input from the image space to a higher dimensional space to extract a shallow feature. The *deep feature extraction* module uses local and non-local attention mechanisms via intra-inter Transformer blocks (iiTB), detailed in the next section, to extract high-dimensional features. Finally, the *reconstruction module* maps the feature vector from a high-dimensional feature space to the image space by applying convolutional layers on the final feature obtained from the deep feature extraction module. The reconstruction module differs based on the task; that is, a standard 3×3 convolution is used for tasks that do

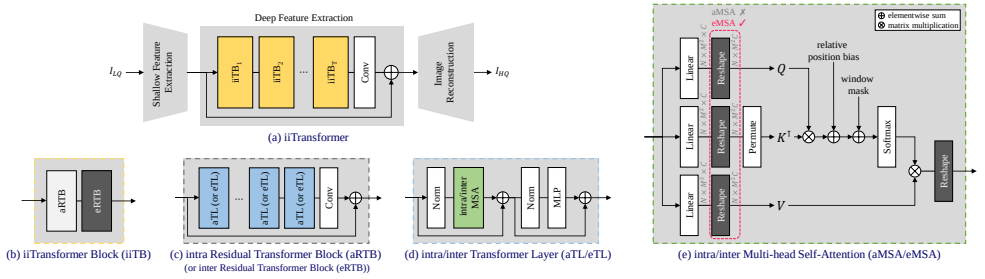


Figure 3: (a) Overview of the iiTransformer architecture. (b) Each iiTB in the deep feature extraction module is made of an aRTB and an eRTB. (c) aRTB (or eRTB) consist of a stack of aTLs (or eTLs). (d) aTL (or eTL) is a Transformer layer with an aMSA (or an eMSA) module. (e) The self-attention module computes pixel- or patch-wise self-attention mechanism by applying the reshape operation prior to and following the attention calculation.

not require upsampling (e.g., denoising and compression artifact removal) and sub-pixel convolution layers [33] for tasks that require upsampling (e.g., super-resolution).

3.2 Intra-Inter Transformer Block (iiTB) and its Constituents

iiTB and aRTB/eRTB. We extract deep features by processing the shallow feature through a sequence of T *intra-inter Transformer blocks* (**iiTB**s) followed by a convolution and a residual connection. iiTB is based on two main modules[2]: the intra residual Transformer block (**aRTB**) and the inter residual Transformer block (**eRTB**), which enable extraction of expressive features by attending to local and non-local information of the input. As shown in Fig. 3c, aRTB and eRTB share a similar structure, which are residual blocks with multiple intra Transformer layers (**aTLs**) and inter Transformer layers (**eTLs**) for aRTB and eRTB, resp., followed by a convolution prior to the skip connection. The spatially invariant convolution enhances the translational equivariance of residual transformer blocks (aRTB and eRTB) and the residual connection allows aggregation of features from different levels [24].

aTL and eTL. The intra and inter Transformer layers (aTL and eTL) share similar overall structure as the Transformer encoder in ViT [13] (Fig. 3d), where the major difference lies in the multi-head self-attention (MSA) module. Given an input feature, aTL extracts features based on its *local* information by treating *pixels* as tokens in the intra MSA module (aMSA) with attention focused *within* the patch. On the contrary, eTL extracts features based on *non-local* patchwise attention in the inter MSA module (eMSA) treating *patches* as tokens and computes similarities *across* patches. The key difference between aMSA and eMSA lie in the shape of the projected tokens used to compute the attention matrix (i.e., queries and keys). While both MSA modules accept flattened $M \times M$ patches as input features of shape $\frac{HW}{M^2} \times M^2 \times C$, the set of projected tokens is different in aMSA and eMSA. The set of projected tokens of aMSA is of shape $\frac{HW}{M^2} \times M^2 \times C$, while those of eMSA is reshaped to $\frac{HW}{M^2} \times M^2 C$ for employment of patchwise attention (see dark grey regions in Fig. 3e). This reshaping operation that acts as a switch between local and non-local attention is crucial for

[2]While we experimented with more sophisticated designs of fusing local and non-local information, we observed marginal benefits, if there were any. We direct curious readers to supplementary material for results to the various designs, including the use of U-Net structure within the iiTransformer. Consequently, we opt to bide with the simple design of sequentially processing the input feature of iiTB through aRTB followed by eRTB, as in Fig. 3b.

not introducing *additional* parameters. Hereon, we omit the use of the term *multi-head* in MSA when possible to describe the intra and inter self-attention mechanisms with brevity.

aMSA. Suppose $X \in \mathbb{R}^{N\times M^2\times C}$ represents a set of flattened non-overlapping $M\times M$ patches partitioned from a feature of shape $C\times H\times W$ such that $N=\frac{HW}{M^2}$ denotes the number of partitioned patches. As in a typical Transformer, the set of queries, keys, and values for X is computed by applying the projection matrices, $W^Q, W^K, W^V \in \mathbb{R}^{C\times C}$, resp., as follows:

$$Q = XW^Q,\ K = XW^K,\ V = XW^V, \qquad Q,\ K,\ V \in \mathbb{R}^{N\times M^2\times C}. \tag{1}$$

The local attention in the intra SA module is computed as

$$\texttt{aSA}(X) = \texttt{attention}(Q,\ K,\ V) = \texttt{softmax}\left(\frac{QK^{\mathsf{P}}}{\sqrt{C}} + B\right)V,\ \texttt{aSA}(X) \in \mathbb{R}^{N\times M^2\times C}, \tag{2}$$

where P denotes the permutation of the last two axes of a tensor and $B \in \mathbb{R}^{M^2\times M^2}$ is a pixel-wise relative position bias whose values are taken from a learnable positional bias parameter $B' \in \mathbb{R}^{(2M-1)\times(2M-1)}$, as in [24, 27].

eMSA. The non-local attention matrix required in the inter SA module applies the `reshape` operation on Q,K,V prior to and following the SA mechanism, as follows:

$$\tilde{Q} = \texttt{reshape}(Q),\ \tilde{K} = \texttt{reshape}(K),\ \tilde{V} = \texttt{reshape}(V), \qquad \tilde{Q},\ \tilde{K},\ \tilde{V} \in \mathbb{R}^{N\times M^2C} \tag{3}$$

$$\texttt{eSA}(\tilde{X}) = \texttt{attention}(\tilde{Q},\ \tilde{K},\ \tilde{V}) = \texttt{softmax}\left(\frac{\tilde{Q}\tilde{K}^{\mathsf{P}}}{\sqrt{C}} + \tilde{B}\right)\tilde{V},\ \texttt{eSA}(\tilde{X}) \in \mathbb{R}^{N\times M^2C} \tag{4}$$

$$\texttt{eSA}(X) = \texttt{reshape}(\texttt{eSA}(\tilde{X})), \qquad \texttt{eSA}(X) \in \mathbb{R}^{N\times M^2\times C}, \tag{5}$$

where $\tilde{B} \in \mathbb{R}^{N\times N}$ is a patchwise relative position bias whose values are taken from a learnable parameter $\tilde{B}' \in \mathbb{R}^{(2N-1)\times(2N-1)}$. While we considered scaling the dot product in (4) as a function of the reshaped query and key dimensions, $\frac{1}{\sqrt{M^2C}}$, to avoid the vanishing gradient problem, we observed degraded performance and used the same scaling factor as aSA in (2).

For both intra and inter self-attention modules, we employ the shifted window approach [27] (denoted `window mask` in Fig. 3e) by shifting the feature $\left(\lfloor\frac{M}{2}\rfloor, \lfloor\frac{M}{2}\rfloor\right)$ pixels before patch partitioning to ensure that the local and non-local relationships are modelled on a diverse set of patches across various layers.

Supporting Arbitrary Resolutions. As vision Transformers assume input images of fixed size (i.e., $(W\times H)_{train} = (W\times H)_{test}$)[3], the computation of non-local similarity on high-resolution images greatly impedes use in high-resolution tasks. Consequently, we add a mask to the *patchwise* relative position bias in eMSA to provide flexibility toward arbitrarily sized inference images. Since similar patches are likely to reside in clusters and very distant patches tend to recur less [47], we mask inter-patch correlations whose distances exceed the furthest possible distance that patches in a training image can possess, as follows:

$$\tilde{B}_{test}[d(P_i,P_j)] = \begin{cases} \tilde{B}_{train}[d(P_i,P_j)] & \text{if } d(P_i,P_j) \le d^{max}_{train}, \\ -\infty & \text{otherwise,} \end{cases} \tag{6}$$

where $d(P_i,P_j)$ denotes the distance between patches P_i and P_j, d^{max}_{train} is set to the distance between the furthest patches in a training image, and $\tilde{B}[k]$ is an element in $\tilde{B}$ indexed at k.

[3]We avoid applying a rescaling operation to match the resolution of the inference to the training image as the rescaling operation will result in a blurred output, which is an extremely undesirable outcome in image restoration.

This solution enforces values in the attention matrix corresponding to inter-patch relationship (i.e., computations in `softmax` of (4)) whose distances exceed d_{train}^{max} to yield to 0 such that the contribution of those patches are minimized during inter-patch feature extraction. This offers a border boundary artifact free-solution during inference (see Fig. 2c), since test images need not be cropped to match the training image size to support arbitrarily sized images.

3.3 Loss

In the task of image restoration, numerous loss functions in various combinations have been explored, such as L_1, L_2, perceptual [19], adversarial [17], frequency [8], and Charbonnier [4]. To test the effectiveness of the iiTransformer *architecture*, we optimize the network by using only the L_1 pixel reconstruction loss for all tasks, as follows:

$$\mathcal{L}(\hat{I}_{HQ},\, I_{HQ}) = \|\hat{I}_{HQ} - I_{HQ}\|_1, \tag{7}$$

where $\hat{I}_{HQ}$ is the estimated HQ output from the network (e.g., denoised, compression artifact removed, or super-resolved image) and I_{HQ} is the corresponding HQ groundtruth image.

4 Experiments

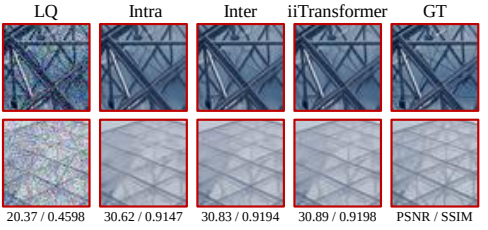
To test the effectiveness of iiTransformer on various image restoration tasks, we conduct experiments on synthetic and real image denoising, compression artifact reduction (CAR), and single image super-resolution (SISR). For image denoising, additive white Gaussian noise (AWGN) with a standard deviation of $\sigma \in \{15, 25, 50\}$ is applied to obtain synthetic noisy images and images from existing dataset are used for real noise (i.e., SIDD [1]). JPEG compression with quality factor $q \in \{10, 20, 30, 40\}$ is considered to obtain compressed images for CAR. Finally, low-resolution images are super-resolved to high-resolution by scale of $s \in \{2, 3, 4\}$ in SISR. For quantitative evaluation, we report PSNR and SSIM on benchmark datasets for each task. We provide details associated with each task in supplemental.

Training Details. We trained separate models for different image restoration tasks and degradation levels end-to-end (i.e., no fine-tuning was employed). We cropped random patches from the training set of DIV2K [2] for all restoration tasks, except real denoising where we employed SIDD [1]. The networks were trained using the Adam optimizer [21] with $(\beta_1, \beta_2) = (0.9, 0.99)$ and learning rate initialized to 2e-4 using the multi-step learning rate scheme. The models were trained on eight NVIDIA Tesla P100 GPUs for up to 500K iterations for all tasks.

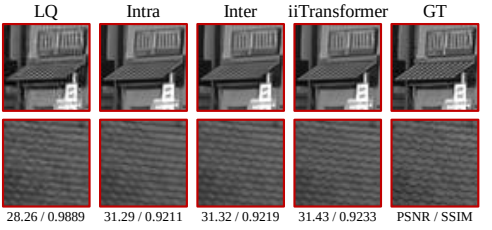
Implementation Details. In iiTransformer, we used $T = 2$ iiTBs with 8 aTLs and 8 eTLs in aRTB and eRTB, resp. MSA modules were employed with 6 heads and the high dimensional feature space was set to $C = 180$ channels. $M \times M = 8 \times 8$ patches were used.

4.1 Ablation Study

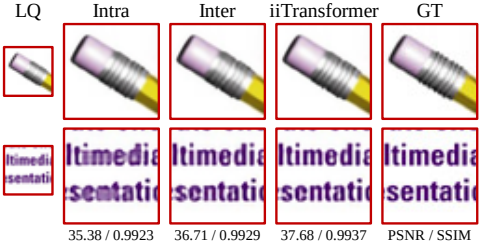
To study the benefits of modelling local and non-local relationships in image restoration, we conduct an ablation study by building a Transformer consisting of (i) only intra MSAs (called *intraTransformer*), (ii) only inter MSAs (called *interTransformer*), and (iii) both intra and inter MSAs (iiTransformer). We ensure the same set of parameters is executed for each studied Transformer by replacing (and not removing) the corresponding block in iiTB.



(a) AWGN Image Denoising for $\sigma = 25$ of regions cropped from *223061* [79]



(b) JPEG CAR for $q = 30$ of regions cropped from *buildings* [32]



(c) SISR at $s = 2$ of regions cropped from *ppt3* [41]

(d) Preferred Attention Map

Figure 4: Qualitative ablation study comparing Transformers using (i) only intra MSAs, (ii) only inter MSAs or (iii) both intra and inter MSAs (iiTransformer). The advantages of using intra MSA is evident in top and inter MSA in bottom rows of a-c.

We provide qualitative and quantitative results of the study on all tasks with select degradation levels in Fig. 4 and Tab. 1, resp. The advantages of utilizing local information can be observed in regions with small details (top rows of Fig. 4a-c), while repetitively occurring patterns, such as edges and corners benefit from computing non-local correlations (bottom rows of Fig. 4a-c). Specifically, the ends of the rootop that is highly degraded in LQ of Fig. 4b top row is recovered by the intraTransformer and iiTransformer, while the diagonal stripes appear blurred in the output of the interTransformer. The hexagonal shape in the bottom row of Fig. 4b is better restored in interTransformer and iiTransformer than intraTransformer. In Fig. 4d, we highlight regions that benefit from attending to local, non-local, and either information in blue, red, and black, resp. Indeed, it shows that textural areas, such as the details on a ferrule is best captured by the intra MSA module, as it is able to capture varying textural details in other neighbourhoods. The edges along the block letters are best captured via the inter MSA module, and homogeneous regions without any structure are captured similarly via the intra and inter MSA modules. The ability to adaptively attend to local and non-local regions depending on the degree of texturedness in iiTransformer is evidenced in its ability to preserve both highly detailed small structures and crisp edges.

Quantitative results on all benchmark datasets demonstrate the robustness of iiTransformer in natural images across various image restoration tasks. Indeed, it is worthy to note that the second best algorithm differs task-to-task (i.e., interTransformer is the second best model for image denoising and CAR, while intra Transformer is the second best for SISR). Consequently, iiTransformer is able to take advantage of both intra- and interTransformer independent of task at hand.

Dataset	intraTransformer	interTransformer	iiTransformer
Kodak24 [ ]	32.25 / 0.8833	32.31 / 0.8846	**32.35 / 0.8848**
BSDS68 [ ]	30.94 / 0.8793	30.98 / **0.8799**	**31.01** / 0.8798
McMaster18 [ ]	31.64 / 0.8515	31.67 / 0.8522	**31.73 / 0.8523**
Urban100 [ ]	31.42 / 0.8983	31.68 / 0.9021	**31.74 / 0.9021**

(a) AWGN Denoising for $\sigma = 25$

Dataset	intraTransformer	interTransformer	iiTransformer
Classic5 [ ]	33.54 / 0.8938	33.56 / 0.8939	**33.61 / 0.8945**
LIVE1 [ ]	32.73 / 0.9144	32.73 / 0.9144	**32.77 / 0.9149**

(b) JPEG CAR for $q = 30$

Dataset	intraTransformer	interTransformer	iiTransformer
Set5 [ ]	38.22 / 0.9610	38.12 / 0.9603	**38.25 / 0.9611**
Set14 [ ]	33.97 / **0.9208**	33.96 / 0.9201	**34.08** / 0.9207
BSDS100 [ ]	31.79 / 0.8949	31.77 / 0.8947	**31.81 / 0.8954**
Urban100 [ ]	33.07 / 0.9362	32.93 / 0.9356	**33.27 / 0.9378**
Manga109 [ ]	39.28 / 0.9775	39.11 / 0.9773	**39.36 / 0.9781**

(c) SISR for scale $\times 2$

Table 1: Ablation experiments of using intra MSAs only (intraTransformer), inter MSAs only (interTransformer), and both (iiTransformer). The best performing algorithm is in **bold**.

4.2 Comparison to State-of-the-Art

We compare iiTransformer with several image restoration methods, including a conventional method (CBM3D [9] for denoising, SA-DCT [14] for CAR, bicubic interpolation for SISR), a CNN-based method without attention (DnCNN [42] for denoising and CAR, SRResNet [23] for SISR), CNN-based method with attention (RNAN [46] for denoising and CAR, RCAN [44] and SAN [10] for SISR), and Transformer-based methods (IPT [7] and SwinIR [24] for all tasks). For fair architectural comparison, we retrained all competing methods, a separate model for each task on DIV2K [1], except real denoising which was trained on SIDD [1], and used default hyperparameters indicated by the authors with the same L_1 reconstruction loss as iiTransformer. No self-ensemble strategy [25] was used for inference.

We provide quantitative results in Tab. 2, which demonstrates iiTransformer surpassing all other methods for all degradation levels in image denoising and CAR by up to 0.34 dB and 0.11 dB on PSNR, resp. iiTransformer continues to perform better than other SISR methods in majority of the scales and datasets with a maximum gain of 0.16 dB on two tests and loss of 0.07 dB on one compared to state-of-the-art, SwinIR [24], which uses more MSAs and utilizes more parameters than the iiTransformer. We provide qualitative results in Fig. 5 to further confirm the benefits of using local and non-local correlations in unison, where the fine-details in the textures of the fabric (Fig. 5a), the chair back (Fig. 5b), and the edges along the letters (Fig. 5c) are best preserved by the proposed iiTransformer. More qualitative examples can be found in the supplemental material.

5 Conclusion

In this paper, we proposed iiTransformer, an effective framework based on Transformers that combines local and non-local attention mechanisms to extract features at various sub-region levels of the image. The local context surrounding the degraded pixel is captured using the intra self-attention modules and the internal data repetition property in natural images is captured by the inter self-attention module. To ensure the inter self-attention module can flexibly process images of various resolutions, we propose to mask the patchwise relative position bias to provide a boundary-artifact free solution. State-of-the-art results on benchmark datasets for various image restoration tasks demonstrate the potential of iiTransformer as a strong general-purpose image restoration backbone architecture.

References

[1] A. Abdelhamed, S. Lin, and M. S. Brown. A High-Quality Denoising Dataset for Smartphone Cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] E. Agustsson and R. Timofte. NTIRE 2017 Challenge on Single Image Super-

Dataset	Noise Level	CBM3D [9]	DnCNN [43]	RNAN [44]	IPT [6]	SwinIR [23]	iiTransformer
Kodak24 [12]	15	34.18 / 0.9143	33.85 / 0.9077	34.10 / 0.9140	34.81 / 0.9242	34.99 / 0.9264	**35.09 / 0.9275**
BSDS68 [29]		33.24 / 0.9187	33.03 / 0.9147	33.28 / 0.9203	33.81 / 0.9282	33.90 / 0.9296	**33.96 / 0.9302**
McMaster18 [43]		34.18 / 0.9143	33.58 / 0.8922	33.67 / 0.8928	34.51 / 0.9051	34.76 / 0.9082	**34.85 / 0.9092**
Urban100 [15]		33.52 / 0.9243	32.81 / 0.9152	32.87 / 0.9172	33.98 / 0.9291	34.29 / 0.9317	**34.53 / 0.9336**
Kodak24 [12]	25	31.37 / 0.8632	30.95 / 0.8486	31.67 / 0.8718	32.20 / 0.8826	32.25 / 0.8834	**32.35 / 0.8848**
BSDS68 [29]		30.17 / 0.8587	29.92 / 0.8485	30.56 / 0.8705	30.93 / 0.8789	30.94 / 0.8794	**31.01 / 0.8798**
McMaster18 [43]		30.64 / 0.8327	30.49 / 0.8261	31.01 / 0.8384	31.59 / 0.8506	31.64 / 0.8515	**31.73 / 0.8523**
Urban100 [15]		30.54 / 0.8855	29.55 / 0.8631	30.40 / 0.8822	31.40 / 0.8977	31.42 / 0.8984	**31.74 / 0.9021**
Kodak24 [12]	50	27.17 / 0.7563	24.51 / 0.7070	27.59 / 0.7670	27.93 / 0.7828	28.00 / 0.7858	**28.09 / 0.7882**
BSDS68 [29]		25.85 / 0.7295	23.48 / 0.6963	26.34 / 0.7477	26.59 / 0.7605	26.63 / 0.7627	**26.70 / 0.7642**
McMaster18 [43]		25.71 / 0.7104	22.47 / 0.6558	26.02 / 0.7157	26.39 / 0.7316	26.44 / 0.7341	**26.53 / 0.7366**
Urban100 [15]		25.65 / 0.7939	22.21 / 0.7135	25.96 / 0.7976	26.52 / 0.8183	26.57 / 0.8211	**26.91 / 0.8314**
SIDD [1]	Real	34.41 / 0.8504	32.14 / 0.7502	38.80 / 0.9099	39.09 / 0.9134	39.40 / 0.9159	**39.66 / 0.9189**
Params Size	-	-	2.13 MB	54.55 MB	257.57 MB	43.70 MB	37.49 MB

(a) Additive White Gaussian Noise (AWGN) and real denoising results on benchmark datasets.

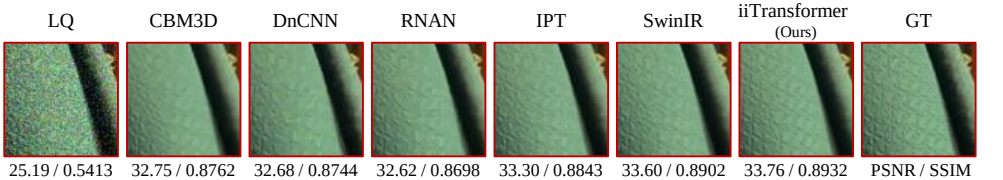
Dataset	Quality	SA-DCT [14]	DnCNN [43]	RNAN [44]	IPT [6]	SwinIR [23]	iiTransformer
Classic5 [14]	10	28.58 / 0.7867	29.16 / 0.7924	28.91 / 0.7912	29.79 / 0.8130	30.03 / 0.8188	**30.06 / 0.8196**
LIVE1 [14]		27.60 / 0.7936	28.36 / 0.8067	27.97 / 0.8000	28.71 / 0.8185	28.88 / 0.8226	**28.91 / 0.8232**
Classic5 [14]	20	28.16 / 0.7897	31.58 / 0.8584	30.14 / 0.8334	31.87 / 0.8647	32.24 / 0.8707	**32.36 / 0.8722**
LIVE1 [14]		27.22 / 0.7911	30.78 / 0.8775	29.32 / 0.8505	30.96 / 0.8821	31.28 / 0.8871	**31.33 / 0.8878**
Classic5 [14]	30	28.88 / 0.7909	32.87 / 0.8853	31.48 / 0.8663	33.36 / 0.8915	33.51 / 0.8934	**33.61 / 0.8945**
LIVE1 [14]		27.90 / 0.7958	32.12 / 0.9071	30.66 / 0.8851	32.54 / 0.9122	32.71 / 0.9141	**32.77 / 0.9149**
Classic5 [14]	40	28.97 / 0.7958	33.84 / 0.9007	32.43 / 0.8849	34.23 / 0.9051	34.35 / 0.9066	**34.44 / 0.9074**
LIVE1 [14]		27.91 / 0.8026	33.14 / 0.9230	31.64 / 0.9042	33.56 / 0.9276	33.72 / 0.9290	**33.78 / 0.9296**
Params Size	-	-	2.13 MB	54.55 MB	257.56 MB	43.70 MB	37.49 MB

(b) JPEG compression artifact removal (CAR) results on benchmark datasets.

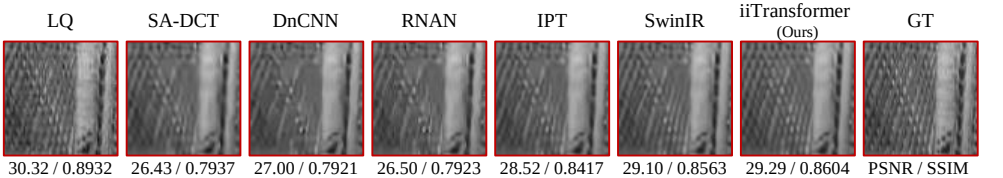
Dataset	Scale	Bicubic	SRResNet [22]	RCAN [44]	SAN [4]	IPT [6]	SwinIR [23]	iiTransformer
Set5 [3]	x2	33.65 / 0.9295	37.94 / 0.9597	38.19 / 0.9605	38.12 / 0.9603	37.80 / 0.9591	**38.25 / 0.9611**	**38.25 / 0.9611**
Set14 [41]		30.31 / 0.8687	33.63 / 0.9176	33.94 / 0.9197	33.86 / 0.9196	33.39 / 0.9157	33.92 / **0.9213**	**34.08** / 0.9207
BSDS100 [29]		29.57 / 0.8442	31.70 / 0.8936	31.78 / 0.8950	31.73 / 0.8944	31.71 / 0.8936	31.78 / 0.8950	**31.81 / 0.8954**
Urban100 [15]		26.87 / 0.8392	32.16 / 0.9280	32.99 / 0.9358	32.78 / 0.9340	31.86 / 0.9252	33.23 / 0.9371	**33.27 / 0.9378**
Manga109 [30]		30.76 / 0.9334	38.43 / 0.9759	39.15 / 0.9774	38.94 / 0.9769	38.25 / 0.9752	39.28 / 0.9774	**39.36 / 0.9781**
Set5 [3]	x3	30.37 / 0.8685	34.32 / 0.9265	34.66 / 0.9291	34.61 / 0.9286	34.14 / 0.9250	**34.82 / 0.9302**	34.75 / 0.9298
Set14 [41]		27.63 / 0.7743	30.29 / 0.8410	30.48 / 0.8452	30.51 / 0.8452	30.24 / 0.8393	**30.73 / 0.8488**	30.70 / 0.8487
BSDS100 [29]		27.24 / 0.7419	28.84 / 0.8015	28.90 / 0.8038	28.91 / 0.8037	28.84 / 0.8013	28.94 / 0.8048	**28.96 / 0.8055**
Urban100 [15]		24.45 / 0.7332	28.15 / 0.8516	28.75 / 0.8643	28.67 / 0.8625	27.89 / 0.8459	28.96 / 0.8682	**29.12 / 0.8708**
Manga109 [30]		26.91 / 0.8537	33.38 / 0.9429	34.04 / 0.9468	33.81 / 0.9458	33.16 / 0.9407	34.42 / 0.9492	**34.43 / 0.9493**
Set5 [3]	x4	28.42 / 0.8111	32.19 / 0.8949	32.33 / 0.8979	32.41 / 0.8977	31.96 / 0.8911	**32.68 / 0.9012**	32.58 / 0.9003
Set14 [41]		26.08 / 0.7029	28.56 / 0.7804	28.73 / 0.7849	28.66 / 0.7839	28.45 / 0.7769	28.85 / 0.7877	**28.90 / 0.7887**
BSDS100 [29]		25.99 / 0.6708	27.33 / 0.7311	27.40 / 0.7333	27.36 / 0.7325	27.33 / 0.7304	**27.41 / 0.7342**	**27.41** / 0.7339
Urban100 [15]		23.14 / 0.6560	26.08 / 0.7843	26.56 / 0.7993	26.46 / 0.7969	25.88 / 0.7772	26.73 / **0.8050**	**26.75** / 0.8049
Manga109 [30]		24.86 / 0.7841	30.34 / 0.9064	30.91 / 0.9127	30.79 / 0.9115	29.96 / 0.8999	**31.26 / 0.9174**	31.20 / 0.9168
Params Size	×4	-	5.79 MB	59.48 MB	58.98 MB	257.70 MB	43.70 MB	37.49 MB

(c) Single Image Super-resolution (SISR) results on benchmark datasets.

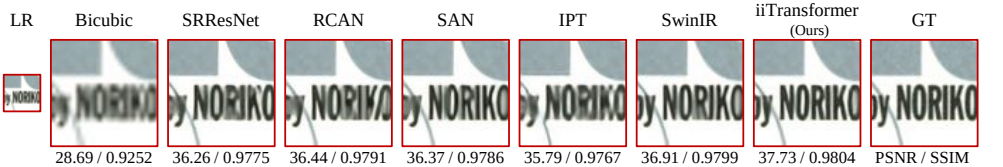
Table 2: Quantitative comparison of iiTransformer to other methods. The best performing algorithm for each dataset, degradation level, and restoration task is marked in **bold**.



(a) AWGN Image denoising with $\sigma = 15$ of region cropped from 2 in the McMaster18 dataset [43].



(b) JPEG CAR with $q = 10$ of region cropped from *barbara* in the Classic5 dataset [14].



(c) SISR x3 of region cropped from *MitsutenaideDaisy* in the Manga109 dataset [30].

Figure 5: Qualitative comparison of various restoration methods. All compared results were retrained using the same dataset and data augmentation techniques for fair comparison.

Resolution: Dataset and Study. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[3] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi Morel. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *British Machine Vision Conference (BMVC)*, 2012.

[4] A. Buades, B. Coll, and J. M. Morel. A Non-Local Algorithm for Image Denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

[6] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two Deterministic Half-Quadratic Regularization Algorithms for Computed Imaging. In *International Conference on Image Processing (ICIP)*, 1994.

[7] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[9] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Transactions on Image Processing (TIP)*, 2007.

[10] T. Dai, J. Cai, Y. Zhang, S. T. Xia, and L. Zhang. Second-order Attention Network for Single Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] C. Dong, Y. Deng, C. C. Loy, and X. Tang. Compression Artifacts Reduction by a Deep Convolutional Network. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[12] C. Dong, C. C. Loy, K. He, and X. Tang. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2015.

[13] 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 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[14] A. Foi, V. Katkovnik, and K. Egiazarian. Pointwise Shape-Adaptive DCT for High-Quality Denoising and Deblocking of Grayscale and Color Images. *IEEE Transactions on Image Processing (TIP)*, 2007.

[15] R. Frazen. Kodak Lossless True Color Image Suite, 1999. URL http://r0k.us/graphics/kodak/.

[16] D. Glasner, S. Bagon, and M. Irani. Super-Resolution from a Single Image. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[18] J. Huang, A. Singh, and N. Ahuja. Single Image Super-resolution from Transformed Self-Exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] J. Johnson, A. Alahi, and F. F. Li. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[20] J. Kim, J. K. Lee, and K. M. Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] D. P. Kingma and J. L. Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[22] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. SwinIR: Image Restoration using Swin Transformer. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[26] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang. Non-local Recurrent Network for Image Restoration. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[27] 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. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[28] Y. Makinen, L. Azzari, and A. Foi. Collaborative Filtering of Correlated Noise: Exact Transform-Domain Variance for Improved Shrinkage and Patch Matching. *IEEE Transactions on Image Processing (TIP)*, 2020.

[29] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *IEEE International Conference on Computer Vision (ICCV)*, 2001.

[30] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa. Sketch-based Manga Retrieval using Manga109 Dataset. In *Multimedia Tools and Applications*, 2017.

[31] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen. Single Image Super-Resolution via a Holistic Attention Network. In *European Conference on Computer Vision (ECCV)*, 2020.

[32] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE Image Quality Assessment Database Release 2, 2004. URL https://live.ece.utexas.edu/research/quality/subjective.htm.

[33] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] A. Shocher, N. Cohen, and M. Irani. "Zero-Shot" Super-Resolution using Deep Internal Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Y. Tai, J. Yang, X. Liu, and C. Xu. MemNet: A Persistent Memory Network for Image Restoration. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[36] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li. MAXIM: Multi-Axis MLP for Image Processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All You Need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[38] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, 2018.

[40] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. Uformer: A General U-Shaped Transformer for Image Restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[41] R. Zeyde, M. Elad, and M. Protter. On Single Image Scale-Up Using Sparse-Representations. In *International Conference on Curves and Surfaces*, 2010.

[42] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing (TIP)*, 2017.

[43] L. Zhang, X. Wu, A. Buades, and X. Li. Color Demosaicking by Local Directional Interpolation and Non-local Adaptive Thresholding. *Journal of Electronic Imaging*, 2011.

[44] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *European Conference on Computer Vision (ECCV)*, 2018.

[45] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual Dense Network for Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu. Residual Non-local Attention Networks for Image Restoration. In *International Conference on Learning Representations (ICLR)*, 2019.

[47] M. Zontak and M. Irani. Internal Statistics of a Single Natural Image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.