Reciprocal Attention Mixing Transformer for Lightweight Image Restoration

Haram Choi[1*] Cheolwoong Na[2] Jihyeon Oh[2] Seungjae Lee[2] Jinseop Kim[2] Subeen Choe[2]
Jeongmin Lee[3] Taehoon Kim[4] Jihoon Yang[2†]
[1]RippleAI [2]Machine Learning Research Lab., Sogang University [3]LG Innotek [4]LG AI Research

Abstract

Although many recent works have made advancements in the image restoration (IR) field, they often suffer from an excessive number of parameters. Another issue is that most Transformer-based IR methods focus only on either local or global features, leading to limited receptive fields or deficient parameter issues. To address these problems, we propose a lightweight network, Reciprocal Attention Mixing Transformer (RAMiT). It employs our proposed dimensional reciprocal attention mixing Transformer (D-RAMiT) blocks, which compute bi-dimensional self-attentions in parallel with different numbers of multi-heads. The bi-dimensional attentions help each other to complement their counterpart's drawbacks and are then mixed. Additionally, we introduce a hierarchical reciprocal attention mixing (H-RAMi) layer that compensating for pixel-level information losses and utilizes semantic information while maintaining an efficient hierarchical structure. Furthermore, we revisit and modify MobileNet V2 to attach efficient convolutions to our proposed components. The experimental results demonstrate that RAMiT achieves state-of-the-art performance on multiple lightweight IR tasks, including super-resolution, low-light enhancement, deraining, color denoising, and grayscale denoising. Codes are available at https://github.com/rami0205/RAMiT.

1. Introduction

Lightweight image restoration (IR) or enhancement techniques are essential for addressing inherent flaws in images captured in the wild, especially those taken by devices with low computational power. These techniques aim to reconstruct high-quality images from their distorted low-quality counterparts. However, many lightweight IR tasks with the popular vision Transformer [14] based methods remain relatively unexplored. Although many recent Transformer [52] networks have improved the IR domain [7, 9, 56, 65, 69], they are infeasible for real-world applications due to their large number of parameters. Furthermore, even the state-of-the-art lightweight IR networks consume intensive computational costs [5, 10, 33, 38, 73]. Another problem is that some IR models mainly focus on expanding the receptive field with respect to locality [9, 10, 33, 56, 73], which is insufficient to capture the global dependency in an image. This is critical because the IR networks need to refer to repeated patterns and textures distributed throughout the image [18, 38]. Meanwhile, others have tried to enlarge the receptive field globally [5, 65, 69] but have overlooked important local (spatial) information, which is conventionally essential for recovery tasks [9, 10, 21, 56]. Fig. 1 visualizes a few examples in which a successful IR depends on the ability to consider both local and global features in a given distorted low-quality image, emphasizing how significant the problem is.

To address these problems, we propose a lightweight IR network called **RAMiT** (**R**eciprocal **A**ttention **Mi**xing **T**ransformer). As shown in Fig. 2a, RAMiT consists of a shallow module, three hierarchical and the last stages composed of $\mathbb{K}_a$ D-RAMiT blocks before and after an upsampling bottleneck layer, an H-RAMi, and a final reconstruction module. Our proposed **D-RAMiT** (**D**imensional **R**eciprocal **A**ttention **Mi**xing **T**ransformer) blocks include a novel bi-dimensional self-attention (SA) mixing module. This operates spatial and channel SA mechanisms [33, 65] in parallel with different multi-heads, and mixes them. To overcome the drawbacks of each SA, we allow the results from the previous block to help the respective counterparts' SA procedures. Consequently, RAMiT can capture both local and global dependencies. Additionally, we propose an efficient component, **H-RAMi** (**H**ierarchical **R**eciprocal **A**ttention **Mi**xer) that mixes the multi-scale attentions resulting from four hierarchical stages. This component complements pixel-level information losses caused by downsampled features, and enhances semantic-level representations. It enables RAMiT to re-think where and how much attention to pay in the given input feature maps. For a mixture of each reciprocal (*i.e.,* dimensional and hierarchical) attention result, we modify MobileNet V2 [48]. Utilizing this **MobiVari** (**Mobi**leNet **Vari**ant) layer, we can ef-

*This work has been done during Master's course in Sogang University.
†Corresponding author.

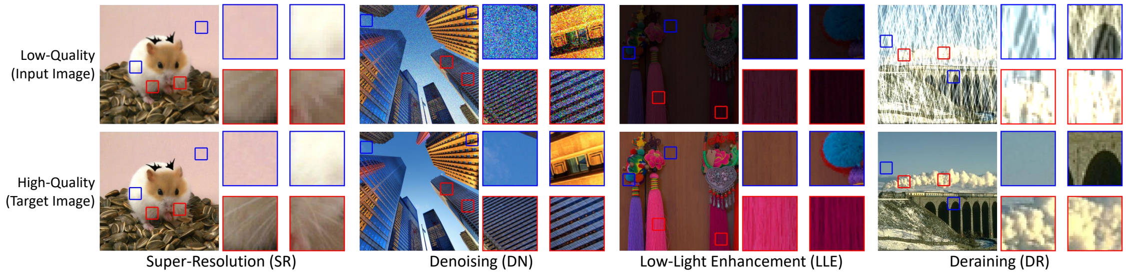


Figure 1. The importance of locality and global dependency in image restoration tasks. (**Blue boxes**) Local features are informative enough to recover most parts, meaning that the contribution of locally adjacent pixels is crucial. (**Red boxes**) Some areas seem more challenging due to high levels of distortion (blurring, noise, darkness, or obstruction). They require global dependency, which can often be detected in repeated patterns or textures distributed throughout the entire image.

ficiently and effectively attach the convolutions to the network.

The experimental results demonstrate that the various lightweight IR works are improved by our RAMiT. As a result, we establish state-of-the-art performance on five different lightweight IR tasks, including super-resolution, low-light enhancement, deraining, color denoising, and grayscale denoising, showing applicability of RAMiT to general low-level vision tasks. Notably, RAMiT achieves these results with fewer operations or parameters than the other networks.

The summaries of our main contributions are as follows:

(1) We propose a dimensional reciprocal attention mixing Transformer (**D-RAMiT**) block. The spatial and channel self-attentions with the different numbers of multi-heads operate in parallel using and are fused. Therefore, the network can capture both local and global context, which is critical for image restoration tasks.

(2) A hierarchical reciprocal attention mixing (**H-RAMi**) layer is introduced. It compensates for pixel-level information losses caused by downsampled features of hierarchical structure, and utilizes semantic-level information, while maintaining an efficient hierarchical structure.

(3) Our RAMiT achieves **state-of-the-art** results on five different lightweight image restoration tasks. It is noteworthy that RAMiT requires fewer parameters or operations compared to existing methods.

2. Related Work

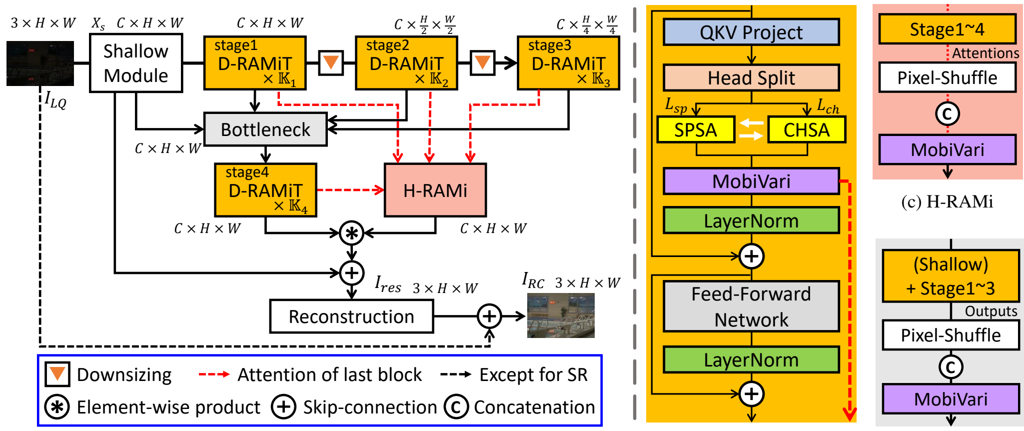
Window Self-Attention. After Vision Transformer (ViT) [14] appeared, Swin Transformer [36] proposed window self-attention (WSA) to solve the excessive time complexity of ViT. Self-attention is computed with the tokens in a non-overlapping local window. However, since the receptive field of WSA was limited within a small window, some following high-level vision studies tried to overcome this issue. GGViT [63], CrossFormer [54], and MaxViT [51] utilized dilated windows to capture the dependency in non-local regions. Focal Transformer [59] gradually widened surrounding regions (*key, value*) of a local window (*query*). CSwin[13] extended square windows to cross-shaped rectangle windows. VSA [72] dynamically varied the window size, breaking the local constraint. DaViT [12] alternately placed spatial WSA and channel self-attention blocks to consider both local and global dependencies in an image.

WSA for Image Restoration. The image restoration (IR) tasks aim to recover a high-quality image from a degraded low-quality counterpart. SwinIR [33] firstly adapted window self-attention (WSA) in this domain and achieved outstanding results. Thereafter, many studies employed WSA and overcame the limited receptive field. Uformer [56] proposed locally-enhanced feed-forward network to refer to neighbor pixels. ELAN [73] split channels of input feature maps into different sized windows, efficiently enlarging the local receptive field. Following [51, 54, 63], ART [69] exploited the dilated window attention. NGswin [10] introduced an N-Gram method helping WSA to consider surrounding pattern and texture. Moreover, Restormer [65] and NAFNet [8] utilized channel-attention rather than spatial WSA for maximizing the capability of attention mechanism in capturing global dependency. Related to the approaches above, we aim to address the weakness of the plain WSA

3. Methodology

3.1. Overall Architecture of RAMiT

As shown in Fig. 2a, given a low-quality image $I_{LQ} \in \mathbb{R}^{3(\text{or}1)\times H\times W}$, a 3×3 convolutional **shallow module** produces $X_s \in \mathbb{R}^{C\times H\times W}$, where H and W are height and width of I_{LQ}, and C is channel. X_s passes through hierarchical **encoder stages** consisting of $\mathbb{K}_a$ **D-RAMiT** (**D**imensional **R**eciprocal **A**ttention **Mi**xing **T**ransformer, Sec. 3.2 and Fig. 2b) blocks, where a indicates the stage number. D-RAMiT calculates self-attention (SA) in bi-dimensions (spatiality and channel) with the different num-



(a) Overall Architecture of RAMiT (b) D-RAMiT Block (c) H-RAMi (d) Bottleneck

Figure 2. Overall architecture of RAMiT. **(a)** The size indicates dimension of output from each component. The operation of $I_{LQ} + I_{res}$ is omitted for super-resolution tasks. I_{RC} equals to $I_{res} \in \mathbb{R}^{3\times rH \times rW}$ (r: an upscale factor). **(b)** The different multi-heads (L_{sp}, L_{ch}) are assigned to each self-attention (SA) module. Being multiplied to *value* of each counterpart, both SAs help each other (white arrows, optional depending on tasks). The bi-dimensional attentions are mixed by our MobileNet variant, MobiVari[1]. **(c)** H-RAMi mixes the hierarchical attentions resulting from the last blocks of each stage. Before MobiVari enhances and mixes the attentions, this module upsamples and concatenates multi-scale attentions. **(d)** Our bottleneck adopts the SCDP bottleneck of NGswin [10].

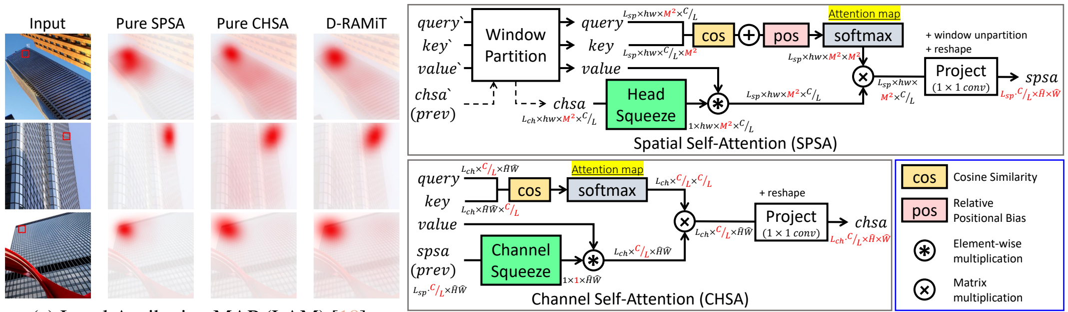
bers of multi-heads. After projecting *query, key, value* and splitting current feature map into L heads, L_{sp} and $L_{ch}(= L - L_{sp})$ heads are assigned to spatial and channel self-attention modules, respectively. For both SAs, we employ scaled-cosine attention and post-normalization [37]. The reciprocally computed attentions are mixed by **MobiVari**[1] (**Mobi**leNet **Vari**ants). Afterwards, the output passes through layer-norm (LN) [4] with skip connection [20], feed-forward network, and LN. At the end of the first and second stages, we downsample the feature maps by half, but maintain the channels. While the **downsizing layers** follow the patch-merging practice of Swin Transformers [36], we replace a plain linear projection of these layers with our MobiVari.

When the stage3 ends, X_s and multi-scale outputs from stage 1, 2, 3 are fed into a **bottleneck layer** (Fig. 2d), which is the same as SCDP bottleneck from NGswin [10] except that depth- and point-wise convolution switches over to our MobiVari. The bottleneck taking multi-scale features can compensate for information loss caused by the downsizing layers. Using a bottleneck output, the **stage4** composed of $\mathbb{K}_4$ D-RAMiT blocks operates in the same way as the other stages. Then, the merged attention results outputted by the last Transformer blocks of all the stages are conveyed to an **H-RAMi** layer (**H**ierarchical **R**eciprocal **A**ttention **Mi**xer, Sec. 3.3 and Fig. 2c). H-RAMi upsamples them into $H \times W$ using a pixel-shuffler [49] and aggregates them, which is merged by MobiVari. This layer is simple but robust to pixel-level information losses as is our bottleneck. The remixed hierarchical attention is element-wise multiplied to the stage4 output. A global skip-connection adds the result with X_s [29], which is then fed into the **reconstruction module** to produce a residual image I_{res}. The reconstruction module follows the common practice [2, 10, 33], but places two MobiVari layers before the original version to boost the performances (detailed in Appendix Sec. A.1). Finally, $I_{res}+I_{LQ}$ makes a reconstructed image I_{RC} (ignored for super-resolution, *i.e.*, $I_{res} = I_{RC}$).

[1] MobiVari modifies the activation function and residual connections and the expansion convolution of the original MobileNet V2 [48]. We detail the MobiVari structure in Appendix Sec. A.1.

3.2. Dimensional Reciprocal Attention Mixing Transformer Block

Motivation. To improve low-level vision tasks like image restoration (IR), it is crucial to refer to repeated patterns and textures distributed through an entire image (*i.e.*, global or non-local context) [18, 38], as already presented in Fig. 1. Nevertheless, while many approaches for high-level vision tasks, such as classification, have enriched non-locality [12, 51, 54, 63], most lightweight IR methods lack the capability to capture global dependency. They maximize only “locality” by adding correlation of adjacent neighbors to a local window [10], or splitting the channels into three groups and the corresponding sizes of local windows within which the self-attention is computed [73]. Meanwhile, channel-attention mechanism is theoretically capable of equipping global dependency by involving all pixels along the channel dimension [8, 25, 65, 74]. Fig. 3a visualizes the actual receptive field of different self-attention methods using the Local Attribution Map [18]. The channel self-attention (<u>**CHSA**</u>) views nearly global areas but performs



(a) Local Attribution MAP (LAM) [18].

(b) Pipelines of our dimensional reciprocal self-attentions.

Figure 3. **(a)** The depth of the red areas indicates the extent to which the regions contribute to recovering a red box of an input. D-RAMiT utilizes both local and global dependencies, meaningfully expanding the receptive field compared to the pure SPSA (see Appendix Sec. A.3). **(b)** Our bi-dimensional self-attention schemes help each other to further boost image restoration performances.

poorly (Tab 4a), because lightweight CHSA focuses on the unnecessary parts with deficient trainable parameters [11]. On the other hand, the spatial self-attention (**SPSA**[2]) suffers from the limited receptive field despite intensive computational costs (Tab 4a), which suggests the potential for further improvement. Hence, our goal is to incorporate local and global context rather than merely enlarging "local" receptive field.

Proposed Method. We propose a bi-dimensional reciprocal self-attention, which is implemented by operating both SPSA and CHSA in parallel (Fig. 2b). Our proposed method can capture both local and global range dependency, thereby improving the IR performances. As illustrated in Fig. 3b, our SPSA and CHSA pipelines adapt the local window self-attention of SwinIR [33] and the transposed attention of Restormer [65], respectively. We assign the different numbers of multi-heads L_{sp} and L_{ch} ($L_{sp} + L_{ch} = L$) to SPSA and CHSA to compute reciprocal attention *Attn*, as follows:

$$Attn = \text{MobiVari}(\texttt{Concat}[SPSA, CHSA]) \tag{1}$$

Each self-attention and the corresponding heads are obtained by Eq. 2 and Eq. 3, respectively:

$$\begin{aligned} SPSA &= \mathcal{P}_{sp}(\texttt{Concat}[head_1^{sp}, ..., head_{L_{sp}}^{sp}]), \\ CHSA &= \mathcal{P}_{ch}(\texttt{Concat}[head_{L_{sp}+1}^{ch}, ..., head_{L}^{ch}]) \end{aligned} \tag{2}$$

$$\begin{aligned} head_i^{sp} &= Softmax(cos(Q_i^{sp}, (K_i^{sp})^T)/\tau + B)V_i^{sp}, \\ head_i^{ch} &= Softmax(cos(Q_i^{ch}, (K_i^{ch})^T)/\tau)V_i^{ch} \end{aligned} \tag{3}$$

$Q_i^{sp}, K_i^{sp}, V_i^{sp}$ and $Q_i^{ch}, K_i^{ch}, V_i^{ch}$ are *query*, *key*, *value* for SPSA and CHSA, respectively; *cos* calculates cosine similarity [37]; $B \in \mathbb{R}^{M^2 \times M^2}$ is the relative positional bias [36]; τ is a trainable scalar that is set larger than 0.01 [37]; $\mathcal{P}_{sp}$, $\mathcal{P}_{ch}$ denotes the reshape and projection

[2] In this paper, SPSA indicates the local window-based self-attention proposed by Swin Transformer [36].

$$\Omega(SPSA) = 4\hat{H}\hat{W}C^2 + 2M^2\hat{H}\hat{W}C$$
$$\Omega(CHSA) = 4\hat{H}\hat{W}C^2 + 2\hat{H}\hat{W}C^2/L$$

Task	Pure CHSA	Pure SPSA	**D-RAMiT** (proposed)
SR ×2	153.4G / 957K	173.4G / 975K	**163.4G / 940K**
SR ×4	39.6G / 978K	44.6G / 996K	**42.1G / 961K**
Denoising	583.2G / 952K	659.9G / 970K	**620.8G / 935K**

*SR: Super-Resolution
*Both methods have the same number of layers and channels.

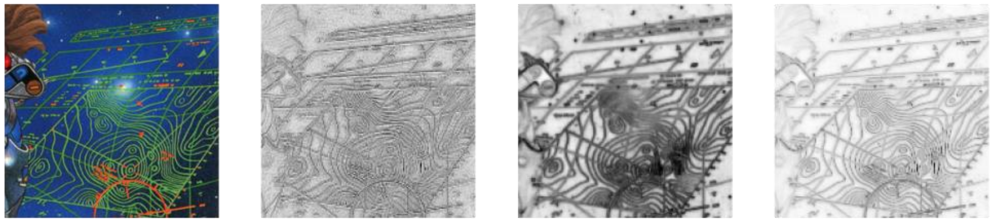
Table 1. **(Eq.)** Time complexity. **(Tab.)** Mult-Adds / #Parameters.

layer. Similar to our work, DaViT [12] has sequentially placed the same numbers of SPSA and CHSA blocks. However, it can consider global context only after attending to spatial dimension (see Appendix Sec. A.2). In contrast, D-RAMiT processes both SAs in parallel, allocating more heads to SPSA (*e.g.*, L_{sp}:L_{ch}=75%:25%). Then, our MobiVari mixes local and global attentions as well as enhances locality by 3×3 depth-wise convolution [56, 65]. The subsequent process follows Fig. 2b.

Reciprocal Helper. Our bi-dimensional modules help each other to compensate for each others' weaknesses, thereby further boosting lightweight IR performances. When operating SPSA of ℓ-*th* block, *value* is element-wise multiplied with the CHSA output of $(\ell - 1)$-*th* block, before multiplying attention map[3] and *value*. The inverse process applies to CHSA as well. It is noteworthy that intensities of information on each SA differ. Each single channel from the previous CHSA has various global representations. Thus, we squeeze (average-pool) it at head dimension before product. On the other hand, averaging channels of SPSA can preserve valuable local properties. As a result, we squeeze feature of the previous SPSA at channel dimension. The first D-RAMiT block of each stage excludes this step due to absence of the previous features with a same resolution. We verify the effects of this approach in Tab. 4b.

Efficiency. The pure SPSA module employed by other IR networks [10, 33, 73] have quadratic time complexity to a local window size. On the other hand, the time com-

[3] Following [73], we remove the attention mask to avoid inefficiency when a cyclic shift [36] is operated.



(a) High-Quality Image (ground-truth) (b) Stage4 Output (o_4) (c) H-RAMi output (o_h) (d) $o_4 * o_h$

Figure 4. Impacts of H-RAMi. **(a)** A ground-truth high-quality image. **(b)**, **(c)** The feature maps after stage 4 and H-RAMi. **(d)** Element-wise product of (b) and (c) (Remind Fig. 2a). (b), (c), (d) are obtained by max-pooling along channel and standardization. More are in Appendix Sec. A.5.

plexity of a CHSA module is usually lower than that of an SPSA, as channels per head (C/L) is mostly not larger than a local window area (M^2) in the equations of Tab. 1. Our proposed D-RAMiT, thus, is more efficient than the pure SPSA. Moreover, D-RAMiT significantly compensates the limited capability of the pure CHSA (see Tab. 4a). Mult-Adds is evaluated on a 1280×720 high-resolution image.

3.3. Hierarchical Reciprocal Attention Mixer

Motivation. There are many evidences that a hierarchical network is usually less effective for IR tasks [10, 11, 23, 76]. This is because the goal of IR is to predict pixel values one by one (*i.e.*, dense prediction) inferring recovery patterns when given the distribution of other pixels [18]. However, downsizing feature maps significantly loses important pixel-level information, which prevents many IR researchers from employing hierarchical structures [2, 5, 39, 45, 73, 74]. Nevertheless, a hierarchical architecture has several advantages. First, reducing the feature map size can lower time complexity. For example, non-hierarchical SwinIR-light [33] requires intensive computations (See Tab. 2). Furthermore, a hierarchical structure can learn semantic-level feature representation as well as pixel-level [19, 53]. To complement the demerits and leverage the merits, we propose the Hierarchical Reciprocal Attention Mixing layer.

Proposed Method. As presented in Fig. 2c, our H-RAMi layer is simple but effective. Inspired by SCDP bottleneck [10], we apply the same strategy to "multi-scale attentions" from the hierarchical encoder stages instead of the final outputs. H-RAMi takes the attentions merged by MobiVari before layer-norm [4] (a red dashed arrow next to a violet rectangle of Fig. 2b) of the last D-RAMiT blocks in the hierarchical stage $1, 2, 3, 4$. After we upsample the resolutions of the mixed bi-dimensional attentions (inputs) into $H \times W$, they are concatenated and mixed by our MobiVari. Therefore, our H-RAMi can take advantage of both multi-scale and bi-dimensional attentions, re-considering where and how much attention to pay semantically and globally. Fig. 4 illustrates the impacts of H-RAMi. The output of stage 4 at (b) produces relatively unclear edges for fine-grained areas. This vulnerability stems from less abundant pixel-level information than non-hierarchical networks [5, 33, 73]. However, H-RAMi reconstructs attentive areas and produces clearer borders at (c) by taking both pixel- and semantic-level information. As a result, the re-attended feature map at (d) contains more apparent and obvious boundaries, which enhances the image restoration accuracy (Tab. 4a).

4. Experiments

4.1. Experimental Setup

Training. We randomly cropped low-quality (LQ) images into various sizes of patches according to each task. The training data was augmented by the random horizontal flip and rotation (90°, 180°, 270°) as done in the recent works [5, 10, 33, 73]. We minimized L_1 pixel-loss between I_{RC} and a ground truth high-quality image I_{HQ}: $\mathcal{L} = \|I_{HQ} - I_{RC}\|_1$ with Adam [30] optimizer. For image super-resolution (SR), 800 high and low resolution image pairs from DIV2K [1] dataset were used. The low-resolution images were acquired by the MATLAB bicubic kernel from corresponding high-resolution images. The color and grayscale image denoising (DN) models were trained on DFBW, a merged dataset of 800 DIV2K, 2,650 Flickr2K [50], 400 BSD500 [3], and 4,744 WED [40] images, following [11, 33, 65, 69]. The random Gaussian noise level σ ranging $[0, 50]$ was used to get noisy LQ images. For low-light image enhancement (LLE), 1,785 dark and bright image pairs were utilized (485 LOL [57] + 1,300 VE-LOL [35]), which were either captured or synthesized. Next, we trained our deraining (DR) model on 13,711 synthesized rainy and clean image pairs of Rain13K [64] collected from [17, 31, 60, 67, 68]. Other details are in Appendix Sec. B.

Evaluation. For SR, we evaluated the performances on the five benchmark datasets, composed of Set5 [6], Set14 [66], BSD100 [42], Urban100 [26], and Manga109 [43]. We calculated PSNR (dB) and SSIM [55]

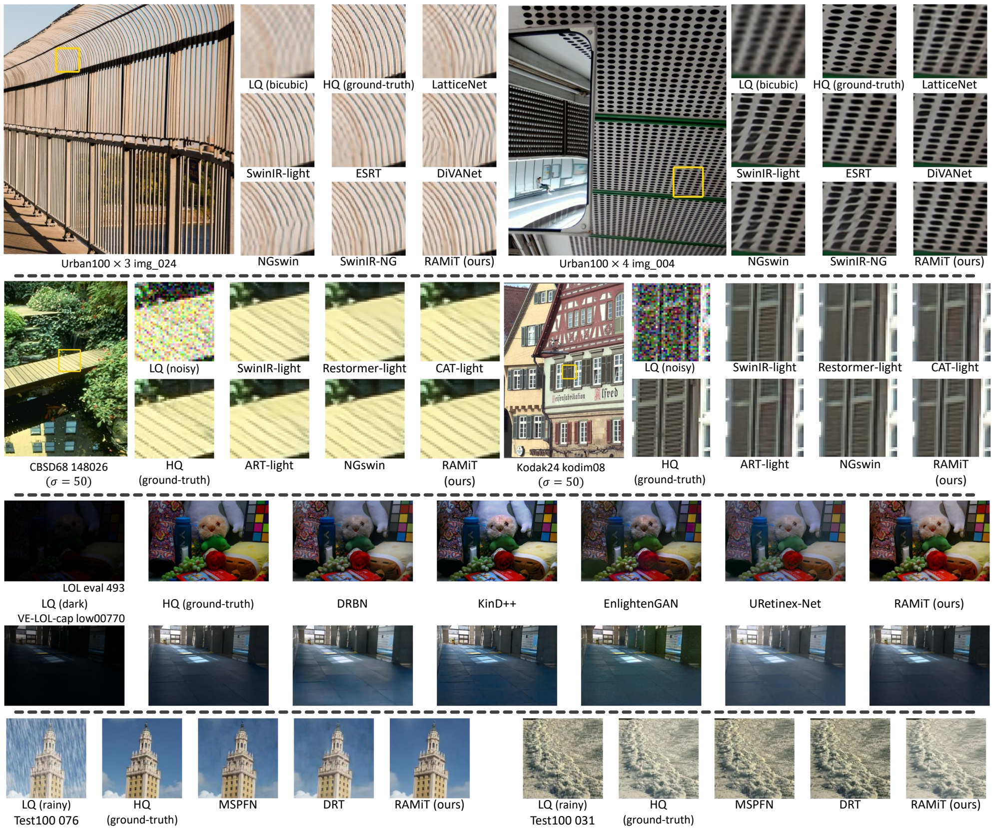


Figure 5. Visual comparisons of multiple lightweight image restoration tasks. LQ: Low-Quality input. HQ: High-Quality target. **(1st row)** Super-Resolution. **(2nd row)** Denoising. **(3rd row)** Low-Light Enhancement. **(4th row)** Deraining. More results are provided in Appendix Sec. C.

scores on the Y channel of the YCbCr space. The same metrics were calculated for testing DR, which involves Test100 [68] and Rain100H [60] datasets. To test DN performances, Gaussian noise with different levels σ of $\{15, 25, 50\}$ is added. We reported PSNR and SSIM on the RGB channel of CBSD68 [42], Kodak24 [16], McMaster [71], and Urban100 for color DN and on Y channel of Set12 [70], BSD68 [42], and Urban100 for grayscale DN. The same metrics for color DN were employed to evaluate the LLE performances on 15 LOL [57] and 100 VE-LOL-cap [35] test images.

4.2. Qualitative Comparisons

Fig. 5 presents the visual comparisons with other models, which were selected based on existing state-of-the-art studies for each task. The illustration demonstrates that our proposed dimensional and hierarchical attention mixing methods were able to recover more accurate textures and patterns than other methods. Our combination of "local and global" and "pixel- and semantic-level" features made our proposed approach effective. More results are in Appendix Sec. C.

4.3. Quantitative Comparisons

Image Super-Resolution (SR). In Tab. 2, we compared our RAMiT with other state-of-the-art lightweight SR methods, including CARN (ECCV18) [2], LatticeNet (ECCV20) [39], SwinIR-light (ICCVW21) [33], FMEN (CVPRW22) [15], ESRT (CVPRW22) [38], ELAN-light (ECCV22) [73], DiVANet (PR23) [5], NGswin (CVPR23) [10], and SwinIR-NG (CVPR23) [10]. We also reported the number of operations (Mult-Adds) of each model. Our RAMiT gained PSNR up to 0.12dB while consuming only $59.6 \sim 67.7\%$ of the operations used by SwinIR-NG. Especially, RAMiT offers the best trade-off between efficiency and performance on $\times 2$ and $\times 4$ tasks among the compared approaches. For a concern of the number of parameters, see Appendix Sec. A.6.

Low-Light Image Enhancement (LLE). RAMiT substantially surpassed the state-of-the-art LLE methods, including DRBN (CVPR20) [61], KinD++ (IJCV21) [75], EnlightenGAN (TIP21) [28], and URetinex-Net (CVPR22) [58], as recorded in Tab. 3a. Our method restored much more accurate brightness and objects from the extremely dark image than other models. While they

Method	Mult-Adds	#Params	Scale	Set5 [6]		Set14 [66]		BSD100 [42]		Urban100 [26]		Manga109 [43]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN [2]	222.8G	1,592K	×2	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
LatticeNet [39]	169.5G	756K	×2	38.06	0.9607	33.70	0.9187	32.20	0.8999	32.25	0.9288	38.94	0.9774
SwinIR-light [33]	243.7G	910K	×2	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
FMEN [15]	172.0G	748K	×2	38.10	0.9609	33.75	0.9192	32.26	0.9007	32.41	0.9311	38.95	0.9778
ESRT [38]	191.4G	677K	×2	38.03	0.9600	33.75	0.9184	32.25	0.9001	32.58	0.9318	39.12	0.9774
ELAN-light [73]	168.4G	582K	×2	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.12	0.9783
DiVANet [5]	189.0G	902K	×2	38.16	0.9612	33.80	0.9195	32.29	0.9012	32.60	0.9325	39.08	0.9775
NGswin [10]	140.4G	998K	×2	38.05	0.9610	33.79	0.9199	32.27	0.9008	32.53	0.9324	38.97	0.9777
SwinIR-NG [10]	274.1G	1,181K	×2	38.17	0.9612	33.94	0.9205	32.31	0.9013	32.78	0.9340	39.20	0.9781
RAMiT (ours)	**163.4G**	**940K**	×2	**38.16**	**0.9612**	**34.00**	**0.9213**	**32.33**	**0.9015**	**32.81**	**0.9346**	**39.32**	**0.9783**
CARN [2]	118.8G	1,592K	×3	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
LatticeNet [39]	76.3G	765K	×3	34.40	0.9272	30.32	0.8416	29.10	0.8049	28.19	0.8513	33.63	0.9442
SwinIR-light [33]	109.5G	918K	×3	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
FMEN [15]	77.2G	757K	×3	34.45	0.9275	30.40	0.8435	29.17	0.8063	28.33	0.8562	33.86	0.9462
ESRT [38]	96.4G	770K	×3	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
ELAN-light [73]	75.7G	590K	×3	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
DiVANet [5]	89.0G	949K	×3	34.60	0.9285	30.47	0.8447	29.19	0.8073	28.58	0.8603	33.94	0.9468
NGswin [10]	66.6G	1,007K	×3	34.52	0.9282	30.53	0.8456	29.19	0.8078	28.52	0.8603	33.89	0.9470
SwinIR-NG [10]	114.1G	1,190K	×3	34.64	0.9293	30.58	0.8471	29.24	0.8090	28.75	0.8639	34.22	0.9488
RAMiT (ours)	**77.3G**	**949K**	×3	**34.63**	**0.9290**	**30.60**	**0.8467**	**29.25**	**0.8093**	**28.76**	**0.8646**	**34.30**	**0.9490**
CARN [2]	90.9G	1,592K	×4	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
LatticeNet [39]	43.6G	777K	×4	32.18	0.8943	28.61	0.7812	27.57	0.7355	26.14	0.7844	30.54	0.9075
SwinIR-light [33]	61.7G	930K	×4	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
FMEN [15]	44.2G	769K	×4	32.24	0.8955	28.70	0.7839	27.63	0.7379	26.28	0.7908	30.70	0.9107
ESRT [38]	67.7G	751K	×4	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
ELAN-light [73]	43.2G	601K	×4	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
DiVANet [5]	57.0G	939K	×4	32.41	0.8973	28.70	0.7844	27.65	0.7391	26.42	0.7958	30.73	0.9119
NGswin [10]	36.4G	1,019K	×4	32.33	0.8963	28.78	0.7859	27.66	0.7396	26.45	0.7963	30.80	0.9128
SwinIR-NG [10]	63.0G	1,201K	×4	32.44	0.8980	28.83	0.7870	27.73	0.7418	26.61	0.8010	31.09	0.9161
RAMiT (ours)	**42.1G**	**961K**	×4	**32.56**	**0.8992**	**28.83**	**0.7873**	**27.71**	**0.7418**	**26.60**	**0.8017**	**31.17**	**0.9170**

Table 2. Comparison of lightweight super-resolution results. Mult-Adds is evaluated on a 1280 × 720 high-resolution image. The best and second best results are in red and blue.

(a) Low-Light Image Enhancement (LLE).

Method	#Params	LOL [57]		VE-LOL-cap [35]	
		PSNR	SSIM	PSNR	SSIM
DRBN [61]	558K	18.80	0.8304	20.11	0.8545
KinD++ [75]	8,275K	21.80	0.8338	22.21	0.8430
EnlightenGAN [28]	8,640K	17.48	0.6507	18.64	0.6754
URetinex-Net [58]	361K	21.33	0.8348	21.22	0.8593
RAMiT (ours)	**935K**	**24.14**	**0.8423**	**28.73**	**0.8886**

(b) Image Deraining (DR).

Method	#Params	Test100 [68]		Rain100H [60]	
		PSNR	SSIM	PSNR	SSIM
UMRL [62]	984K	24.41	0.8290	26.01	0.8320
MSPFN [27]	13,350K	27.50	0.8760	28.66	0.8600
DRT [34]	1,180K	27.02	0.8470	29.47	0.8460
TAO-Net [32]	755K	28.59	0.8870	28.96	0.8640
RAMiT (ours)	**935K**	**30.44**	**0.9012**	**29.69**	**0.8775**

Table 3. Comparison of lightweight low-light image enhancement and image deraining results.

adhered to the conventional approaches, such as Retinex algorithms [46] and convolutional neural networks, our advanced Transformer easily defeated them by up to 6.52dB of the PSNR score.

Image Deraining (DR). Tab. 3b shows that RAMiT could more sufficiently remove rains than the state-of-the-art DR methods: UMRL (CVPR19) [62], MSPFN (CVPR20) [27], DRT (CVPRW22) [34], and TAO-Net (SPLetters22) [32]. We gained PSNR scores up to 1.73dB with the second smallest architecture. In particular, MSPFN network fell behind RAMiT in performance despite having 3.89 times more parameters than RAMiT.

Color Image Denoising (CDN). In Tab. 5a, we referred to the lightweight denoising Transformer baselines introduced by [11], such as SwinIR-light (ICCVW21) [33], Restormer-light (CVPR22) [65], CAT-light (NeurIPS22) [9], ART-light (ICLR23) [69], and NGswin (CVPR23) [10]. It is notable that SwinIR, CAT, and NGswin aimed to boost locality of a window-based spatial self-attention, while Restoremer and ART pursued an improved ability in capturing non-local dependency in an image. However, RAMiT surpassed them on every noise level and dataset through both local and global context.

Grayscale Image Denoising (GDN). As shown in Tab. 5b, our RAMiT was good at removing noise from the grayscale images as well. RAMiT reconstructed more similar images to ground-truth for human-perception in that our SSIM scores were the highest. Moreover, RAMiT gained PSNR scores on all noise levels up to 0.23dB.

4.4. Ablation Study

D-RAMiT. Tab. 4a ({i} *vs.* {iii} *vs.* {v}) compares our D-RAMiT with a pure SPSA and CHSA on SR ×2, ×4, CDN, LLE, and DR. The proposed D-RAMiT overcame the limited capacity of CHSA and the narrow receptive field of SPSA. Our method achieved better results on multiple tasks with fewer computations and parameters than SPSA. This effectiveness is also observed without the H-RAMi layer, another proposed method ({ii} *vs.* {iv}). Moreover, as shown in Tab. 4b, the reciprocal helper contributed

(a) D-RAMiT & H-RAMi (Mult-Adds / #Params / Average PSNR).

Transformer	H-RAMi	SR ×2	SR ×4	CDN $\sigma = 50$	LLE	DR	
Pure CHSA	*w/*	153.4G / 957K / 34.994	39.6G / 978K / 29.074	583.2G / 952K / 28.848	583.2G / 952K / 23.985	583.2G / 952K / 28.810	{i}
Pure SPSA	*w/o*	168.6G / 955K / 35.218	43.4G / 976K / 29.302	641.2G / 950K / 29.010	641.2G / 950K / 25.095	641.2G / 950K / 29.175	{ii}
Pure SPSA	*w/*	173.4G / 975K / 35.276	44.6G / 996K / 29.342	659.9G / 970K / 29.128	659.9G / 970K / 25.140	659.9G / 970K / 29.190	{iii}
D-RAMiT	*w/o*	158.5G / 920K / 35.310	40.9G / 940K / 29.338	602.1G / 914K / 29.205	602.9G / 914K / 26.365	602.1G / 914K / 29.940	{iv}
D-RAMiT	*w/*	163.4G / 940K / **35.324**	42.1G / 961K / **29.374**	620.8G / 935K / **29.275**	621.6G / 935K / **26.435**	620.8G / 935K / **30.065**	{v}

(b) Reciprocal Helper (*w/o* / *w/*).

Task	Mult-Adds (G)	PSNR
SR ×2	163.2 / 163.4	35.308 / **35.324**
SR ×3	77.16 / 77.26	31.482 / **31.508**
SR ×4	42.08 / 42.13	29.308 / **29.374**
LLE	620.8 / 621.6	25.915 / **26.435**

(c) MobiVari Activation Function.

Activation	SR ×2	CDN $\sigma = 50$	LLE
ReLU6 [24, 48]	35.304	29.220	25.530
ReLU [44]	35.322	29.270	26.160
GELU [22]	35.320	29.268	26.230
Swish$_{\beta=1}$ [47]	35.306	29.250	26.160
LeakyReLU [41]	**35.324**	**29.275**	**26.435**

Table 4. Ablation studies on our proposed methods. The reported PSNR scores represent the average values on the benchmark test datasets of each image restoration task provided in Tabs. 2, 3, 5. Mult-Adds is calculated on a 1280 × 720 high-quality image.

(a) Color Image Denoising (CDN).

Method	#Params	σ	CBSD68 [42]		Kodak24 [16]		McMaster [71]		Urban100 [26]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR-light [33]	905K	15	34.16	0.9323	35.18	0.9269	35.23	0.9295	34.59	0.9478
Restormer-light [65]	1,054K	15	33.99	0.9311	34.86	0.9244	34.69	0.9229	34.00	0.9439
CAT-light [9]	1,042K	15	34.01	0.9304	34.90	0.9237	34.83	0.9247	34.12	0.9443
ART-light [69]	1,084K	15	34.08	0.9315	35.00	0.9251	35.10	0.9282	34.44	0.9467
NGswin [10]	993K	15	34.12	0.9324	35.12	0.9268	35.17	0.9294	34.53	0.9476
RAMiT (ours)	**935K**	15	**34.23**	**0.9332**	**35.22**	**0.9276**	**35.31**	**0.9309**	**34.68**	**0.9488**
SwinIR-light [33]	905K	25	31.50	0.8883	32.69	0.8868	32.90	0.8977	32.23	0.9222
Restormer-light [65]	1,054K	25	31.33	0.8865	32.38	0.8833	32.44	0.8905	31.60	0.9161
CAT-light [9]	1,042K	25	31.37	0.8855	32.43	0.8822	32.58	0.8928	31.75	0.9167
ART-light [69]	1,084K	25	31.40	0.8864	32.49	0.8833	32.74	0.8956	32.03	0.9195
NGswin [10]	993K	25	31.44	0.8884	32.61	0.8865	32.82	0.8978	32.13	0.9215
RAMiT (ours)	**935K**	25	**31.59**	**0.8902**	**32.76**	**0.8887**	**33.02**	**0.9008**	**32.36**	**0.9244**
SwinIR-light [33]	905K	50	28.22	0.8006	29.54	0.8089	29.71	0.8339	28.89	0.8658
Restormer-light [65]	1,054K	50	28.04	0.7974	29.19	0.8034	29.31	0.8256	28.30	0.8559
CAT-light [9]	1,042K	50	28.11	0.7960	29.29	0.8024	29.48	0.8296	28.46	0.8573
ART-light [69]	1,084K	50	28.08	0.7950	29.27	0.8000	29.48	0.8279	28.62	0.8584
NGswin [10]	993K	50	28.13	0.8011	29.42	0.8087	29.59	0.8339	28.75	0.8644
RAMiT (ours)	**935K**	50	**28.37**	**0.8058**	**29.67**	**0.8143**	**29.91**	**0.8422**	**29.15**	**0.8729**

(b) Grayscale Image Denoising (GDN).

Method	#Params	σ	Set12 [70]		BSD68 [42]		Urban100 [26]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR-light [33]	903K	15	33.04	0.9052	31.78	0.8926	33.04	0.9317
Restormer-light [65]	1,053K	15	32.93	0.9039	31.76	0.8922	32.81	0.9306
CAT-light [9]	1,041K	15	32.91	0.9021	31.89	0.8913	31.80	0.8901
ART-light [69]	1,082K	15	32.93	0.9023	31.73	0.8911	32.89	0.9299
NGswin [10]	991K	15	33.04	0.9055	31.78	0.8927	32.99	0.9314
RAMiT (ours)	**932K**	15	**33.14**	**0.9070**	**31.82**	**0.8939**	**33.19**	**0.9346**
SwinIR-light [33]	903K	25	30.67	0.8669	29.32	0.8325	30.52	0.8963
Restormer-light [65]	1,053K	25	30.60	0.8659	29.32	0.8322	30.32	0.8952
CAT-light [9]	1,041K	25	30.60	0.8641	29.47	0.8330	29.32	0.8393
ART-light [69]	1,082K	25	30.52	0.8620	29.25	0.8285	30.30	0.8919
NGswin [10]	991K	25	30.65	0.8671	29.33	0.8324	30.46	0.8961
RAMiT (ours)	**932K**	25	**30.79**	**0.8694**	**29.37**	**0.8346**	**30.71**	**0.9013**
SwinIR-light [33]	903K	50	27.50	0.7966	26.35	0.7299	27.01	0.8190
Restormer-light [65]	1,053K	50	27.48	0.7960	26.38	0.7285	26.92	0.8190
CAT-light [9]	1,041K	50	27.49	0.7935	26.52	0.7333	26.06	0.7456
ART-light [69]	1,082K	50	27.26	0.7856	26.25	0.7194	26.68	0.8065
NGswin [10]	991K	50	27.42	0.7961	26.38	0.7298	26.96	0.8192
RAMiT (ours)	**932K**	50	**27.65**	**0.8013**	**26.46**	**0.7333**	**27.32**	**0.8306**

Table 5. Comparison of lightweight blind image denoising results. We refer to the baselines in [11].

to the improvement. This approach consumed only minor amounts of Mult-Adds and no extra parameters. Therefore, it was proven that our dimensional reciprocal self-attention mixing Transformers could be suitable for general IR tasks.

H-RAMi. Tab. 4a ({ii} *vs.* {iii}, {iv} *vs.* {v}) revealed that H-RAMi constituted another critical component, not only for our D-RAMiT but also for a pure SPSA. Regardless of tasks, this layer enabled the models to remain robust even when a hierarchical network caused information losses. We assumed that since a noisy image contained more distorted boundaries, the impacts of H-RAMi that could recover more accurate object boundaries (Sec. 3.3) were particularly significant in denoising tasks. Additionally, the results highlighted the remarkable efficiency in that H-RAMi required marginal additional operations and parameters, which accounted for a maximum of only 3.01% and 2.25% of the total costs, respectively.

MobiVari. In Tab. 4c, we investigated different non-linear activation functions for our MobiVari. LeakyReLU [41] resulted in the best stable performances across multiple tasks and was selected as the default option. Such stability of LeakyReLU can be attributed to its ability to better preserve relatively large absolute negative values compared to other activation functions. These values, which are occasionally generated by intermediate layers, may have a subtle influence on a feature map, ultimately leading to a significant difference in the final output of a network.

5. Conclusion

This paper proposed the Reciprocal Attention Mixing Transformers (RAMiT). To incorporate local and global context in an image, our Dimensional Reciprocal Attention Mixing Transformer (D-RAMiT) blocks computed bi-dimensional self-attentions in parallel and mixed them. The reciprocal helper was useful for this mechanism. Moreover, the Hierarchical Reciprocal Attention Mixing (H-RAMi) layer was also introduced, where the information losses caused by downsampling were complemented. For mixing attentions and other convolutional layers, we revisited and modified the MobileNet. As a result, our RAMiT achieved state-of-the-art performances on multiple lightweight image restoration tasks, including super-resolution, low-light enhancement, deraininig, color denoising, and grayscale denoising. In closing, we hope this work can be further developed and extended to other low-level tasks.

Acknowledgements. This paper was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No.2022-0-00956) and Korea Health Industry Development Institute (KHIDI) grant (No. H122C1983) funded by the Korea government (MSIT). Special thanks to Prof. Dr. Dasaem Jeong for valuable advises, and Namo Bang, Jeongkyun Park, Juntae Kim, and Gyusik Choi of the Sogang Synergy Journal Club for precious comments.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 5

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 252–268, 2018. 3, 5, 6, 7

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 5

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3, 5

[5] Parichehr Behjati, Pau Rodriguez, Carles Fernández, Isabelle Hupont, Armin Mehri, and Jordi Gonzàlez. Single image super-resolution based on directional variance attention network. *Pattern Recognition*, 133:108997, 2023. 1, 5, 6, 7

[6] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *BMVA press*, 2012. 5, 7

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 2, 3

[9] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *Advances in Neural Information Processing Systems*, 2022. 1, 7, 8

[10] Haram Choi, Jeongmin Lee, and Jihoon Yang. N-gram in swin transformers for efficient lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2071–2081, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[11] Haram Choi, Cheolwoong Na, Jinseop Kim, and Jihoon Yang. Exploration of lightweight single image denoising with transformers and truly fair training. In *Proceedings of the 2023 International Conference on Multimedia Retrieval*, 2023. 4, 5, 7, 8

[12] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 74–92. Springer, 2022. 2, 3, 4

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2

[15] Zongcai Du, Ding Liu, Jie Liu, Jie Tang, Gangshan Wu, and Lean Fu. Fast and memory-efficient network towards efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 853–862, 2022. 6, 7

[16] Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 4(2), 1999. 6, 8

[17] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017. 5

[18] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 1, 3, 4, 5

[19] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[22] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 8

[23] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4003–4012, 2020. 5

[24] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 8

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[26] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 5, 7, 8

[27] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020. 7

[28] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021. 6, 7

[29] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 3

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[31] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2744, 2016. 5

[32] Yufeng Li, Zhentao Fan, Jiyang Lu, and Xiang Chen. Tao-net: Task-adaptive operation network for image restoration and enhancement. *IEEE Signal Processing Letters*, 29:2198–2202, 2022. 7

[33] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[34] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 589–598, 2022. 7

[35] Jiaying Liu, Dejia Xu, Wenhan Yang, Minhao Fan, and Haofeng Huang. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 129:1153–1184, 2021. 5, 6, 7

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 4

[37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 3, 4

[38] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 457–466, 2022. 1, 3, 6, 7

[39] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *European Conference on Computer Vision*, pages 272–289. Springer, 2020. 5, 6, 7

[40] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26 (2):1004–1016, 2016. 5

[41] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, Georgia, USA, 2013. 8

[42] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 5, 6, 7, 8

[43] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 5, 7

[44] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 8

[45] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020. 5

[46] Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Retinex processing for automatic image enhancement. *Journal of Electronic imaging*, 13(1):100–110, 2004. 7

[47] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 8

[48] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1, 3, 8

[49] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3

[50] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 5

[51] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision–ECCV*

2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 459–479. Springer, 2022. 2, 3

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[53] Tao Wang, Guangpin Tao, Wanglong Lu, Kaihao Zhang, Wenhan Luo, Xiaoqin Zhang, and Tong Lu. Restoring vision in hazy weather with hierarchical contrastive learning. *arXiv preprint arXiv:2212.11473*, 2022. 5

[54] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *International Conference on Learning Representations, ICLR*, 2022. 2, 3

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[56] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 1, 2, 4

[57] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018. 5, 6, 7

[58] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2022. 6, 7

[59] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *Advances in Neural Information Processing Systems*, 2021. 2

[60] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017. 5, 6, 7

[61] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3063–3072, 2020. 6, 7

[62] Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8405–8414, 2019. 7

[63] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. *Advances in Neural Information Processing Systems*, 34: 12992–13003, 2021. 2, 3

[64] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 5

[65] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1, 2, 3, 4, 5, 7, 8

[66] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 5, 7

[67] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018. 5

[68] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019. 5, 6, 7

[69] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 5, 7, 8

[70] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 6, 8

[71] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging*, 20(2): 023016, 2011. 6, 8

[72] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vsa: learning varied-size window attention in vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 466–483. Springer, 2022. 2

[73] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, 2022. 1, 2, 3, 4, 5, 6, 7

[74] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 3, 5

[75] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021. 6, 7

[76] Yuzhi Zhao, Lai-Man Po, Qiong Yan, Wei Liu, and Tingyu Lin. Hierarchical regression network for spectral reconstruction from rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 422–423, 2020. 5