Multi-scale Attention Network for Single Image Super-Resolution

Yan Wang† Yusen Li† Gang Wang Xiaoguang Liu
Nankai-Baidu Joint Lab, Nankai University

Code: https://github.com/icandle/MAN

Abstract

ConvNets can compete with transformers in high-level tasks by exploiting larger receptive fields. To unleash the potential of ConvNet in super-resolution, we propose a multi-scale attention network (MAN), by coupling classical multi-scale mechanism with emerging large kernel attention. In particular, we proposed multi-scale large kernel attention (MLKA) and gated spatial attention unit (GSAU). Through our MLKA, we modify large kernel attention with multi-scale and gate schemes to obtain the abundant attention map at various granularity levels, thereby aggregating global and local information and avoiding potential blocking artifacts. In GSAU, we integrate gate mechanism and spatial attention to remove the unnecessary linear layer and aggregate informative spatial context. To confirm the effectiveness of our designs, we evaluate MAN with multiple complexities by simply stacking different numbers of MLKA and GSAU. Experimental results illustrate that our MAN can perform on par with SwinIR and achieve varied trade-offs between state-of-the-art performance and computations.

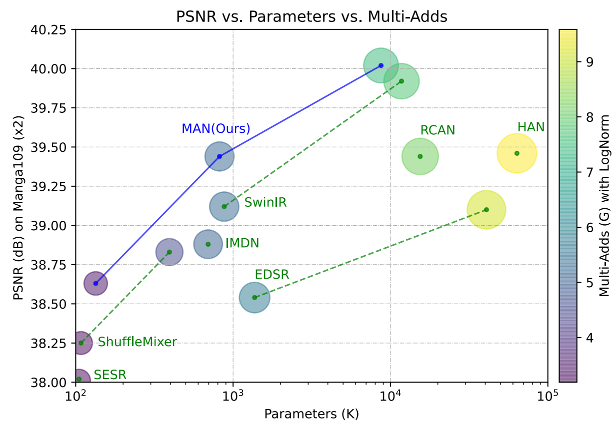

Figure 1. Trade-off between performance and model complexity on Manga109 [39] with ×2 SR scale. MANs can achieve higher PSNR with fewer parameters and computations.

1. Introduction

Image super-resolution (SR) is a widely concerned low-level computer vision task that focuses on rebuilding the missing high-frequency information from the low-quality input [16, 43, 46, 49]. However, it is ill-posed that one low-resolution (LR) image corresponds to countless potential high-resolution (HR) images, leading to difficulties in finding proper correlations between the LR and HR pixels. Due to the boom of deep neural networks, several CNN- and transformer-based SR models [11, 12, 30, 31, 58] have been developed that use prior and intra-image information to improve the reconstruction quality. Generally, they approach the issue from two perspectives.

The first and simplest way is to enlarge the model capacity by training the network with larger datasets and better strategies. Specifically, based on ImageNet [10], IPT [4] and HAT [6] conducted a sophisticated pre-training to excavate the capability of transformers in image processing. LSDIR [29] introduced a large-scale dataset to exploit model capacity fully. RCAN-it [32] leveraged reasonable training strategies to help RCAN [58] regain SOTA performance. Generally, these approaches are universal for neural models but increase burdensome training and data collection consumption.

Another effective way is to activate more intra-image information via better network design. One primary idea is to enlarge the perceptive fields, which means a deeper and wider network topology. Following this, [22, 31, 58] continuously expanded the network depth to hundreds of layers. Nevertheless, the improvement brought by this strategy is limited by over-training and high training costs. Recent models [20, 33, 51] have employed complex topology and attention mechanisms to capture more useful information, *e.g.*, multi-scale [25] design and non-local attention [40].

Recently, the transformer-based models [4, 13, 27, 30] have shown a remarkable representation ability of self-attention (SA), which gradually superseded ConvNets as the state-of-the-art model in both high- and low-level tasks. To

†Corresponding authors, email: {wangy, yusenli}@nbjl.nankai.edu.cn.

Table 1. SR Performance of ConvNets *vs.* Transformer backbones.

Backbone	#Params	#FLOPs	Set5	Set14	B100	U100
ConvNeXt-S [35]	833K	47.3G	32.22	28.62	27.60	26.07
VAN-S [15]	818K	46.0G	32.27	28.70	27.63	26.18
SwinIR-light [30]	896K	49.6G	32.30	28.73	27.65	26.30
MAN-light	840K	47.1G	32.33	28.76	27.67	26.31

fight back, many pure ConvNet (ConvNeXt [35], VAN [15]) sprouted and achieved comparable performance in high-level tasks. But do they perform well in low-level? In Tab. 1, we tested them and noticed VAN performs better than ConvNeXt but still falls behind transformers. To modernize the VAN to compete with transformers in the SR field, we reassess the design of VAN. Generally, VAN [15] explored kernel decomposition and proposed large kernel attention (LKA), where a large kernel can be replaced by stacked depth-wise, depth-wise dilation, and point-wise convolution layers. Despite LKA's efficiency in enabling vast receptive filed, we notice that the dilation convolution in LKA may cause blocking artifacts, which hurt the restored performance. Additionally, a fixed-size LKA is inflexible to fully exploit the pending features since surrounding and remote pixels play equal roles in reconstruction.

Motivated by these issues, we propose multi-scale large kernel attention (MLKA) that combines classical multi-scale mechanism and emerging LKA to build various-range correlations with relatively few computations. The multi-scale kernel can implicitly encode features from coarse to fine, which allows the model to mimic both CNNs and transformers. Moreover, to avoid potential block artifacts aroused by dilation, we adopt the gate mechanism to recalibrate the generated attention maps adaptively. To maximize the benefits of MLKA, we place it on the MetaFormer [53]-style (*Norm-TokenMixer-Norm-MLP*) structure rather than RCAN-style (*Conv-Act-Conv-TokenMixer*) to construct a multi-attention block (MAB). Although transformer-style MAB can deliver higher performance, the MLP feed-forward module is too heavy for large images. Inspired by recent work [5, 47], we propose a simplified gated spatial attention unit (GSAU) by applying spatial attention and gate mechanism to reduce calculations and include spatial information. Arming with the simple yet striking MLKA and GSAU, the MABs are stacked to build the multi-scale attention network (MAN) for the SR task. In Fig. 1, we present the superior performance of our MAN. To summarize, our contributions are as follows:

- We propose multi-scale large kernel attention (MLKA) for obtaining long-range dependencies at various granularity levels by combining large kernel with gate and multi-scale mechanisms, which significantly increases model representation capability.
- We integrate gate mechanisms and spatial attention to construct a simplified feed-forward network called GSAU which has better performance than a multi-layer perceptron (MLP) while reducing parameters and calculations.
- Through simply stacking the proposed modules, we present multi-scale attention network (MAN) family capable of achieving a trade-off between model complexity and performance in both lightweight and performance-oriented SR tasks.

2. Related Work

2.1. Single Image Super-Resolution

Numerous deep-learning models [12, 22, 48, 54] have been proposed for SISR since the pioneering work SRCNN [11] introduced a 3-layer convolutional neural network (CNN) to map the correlation between LR and HR images. Depending on the model complexity, we can classify these solutions as the classical (performance-oriented) SR and the lightweight SR.

For the classical SR task, models are delicately designed for better reconstruction quality. Specifically, VDSR [22] and EDSR [31] were proposed to exploit deeper information by residual learning and increasing depth and width. RCAN [58] then developed EDSR by introducing channel attention (CA) and residual in residual (RIR) to further excavate intermediate features. After RCAN, many works [8, 41, 51] added attention mechanisms to the EDSR structure to boost performance. Recently, vision transformers [4, 30] with self-attention (SA) were introduced to improve image restoration and refresh the SOTA performance.

For tiny and lightweight SR, the model size is constrained for mobile device deployment. The recursive learning was considered effective in decreasing the parameters in DRCN [21], DRRN [45], and LapSRN [24]. However, recursively using modules only reduces model size but maintains high computation costs. More recent works leverage productive operations, *e.g.*, channel splits and attention module, to exploit the hierarchical features. For example, IMDN [19] proposed information multi-distillation and contrast-aware channel attention.

2.2. Attention in Super-Resolution

The attention mechanism can be viewed as a discriminative selection process that focuses on informative regions and ignores the irrelevant noise of pending features. Many SR networks apply attention modules to exploit latent correlations among the immediate features. Following RCAN [58] that first adopted channel attention, SAN [8] leveraged second-order channel attention to adapt the channel-wise features through second-order statistics. Several works introduced spatial attention to enrich the feature maps, *e.g.*, enhanced spatial attention in RFANet [33], and spatial-channel attention in HAN [41]. Additional CNN-based works have utilized and refined non-local attention (NLA) to obtain long-range correlations [40, 51] and achieved an appreciable performance gain. Inspired by vision transformers [34, 47], self-attention has been employed in SR to capture long-term adaptability, *e.g.*, IPT [4] and SwinIR [30]. More recently, DAT [7] leveraged SA along both channel and spatial dimensions and enabled an effective information aggregation to achieve a prominent record. GRL [28] utilized varied SA to explicitly model image hierarchies from coarse to fine to improve the recovery quality.

3. Methodology

3.1. Network Architecture

As illustrated in Fig. 2, the proposed MAN is constituted of three components: the shallow feature extraction module (SF), the deep feature extraction module (DF) based on multiple multi-scale attention blocks (MAB), and the high-quality image reconstruction module. Given an input LR image $I_{LR} \in \mathbb{R}^{3\times H\times W}$, the SF module is first utilized to extract the primitive feature $F_p \in \mathbb{R}^{C\times H\times W}$ by a single 3×3 convolution function $f_{SF}(\cdot)$ as follows:

$$F_p = f_{SF}(I_{LR}). \tag{1}$$

The F_p is then sent to cascading MABs for further extraction, termed as $f_{DF}(\cdot)$, which can be formulated as:

$$F_r = f_{DF}(F_p), \tag{2}$$

where the F_r is the estimated high-frequency feature for final restoration. By adding the long residual feature, the final reconstruction component restores the HQ images $I_{SR} \in \mathbb{R}^{3\times H\times W}$ by:

$$I_{SR} = f_{RC}(F_p + F_r), \tag{3}$$

where $f_{RC}(\cdot)$ represents reconstruction module implemented by a 3×3 convolution and a pixel-shuffle layer for efficiency.

In terms of optimization, we utilize the widely used ℓ_1 loss for a fair comparison with state-of-the-art methods [31, 41, 58]. Specifically, supposing an input batch of N images, *i.e.* $\{I_i^{LR}, I_i^{HR}\}_{i=1}^N$, training MAN is to minimize the ℓ_1:

$$\ell_1(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|f_{MAN}(I_i^{LR}) - I_i^{HR}\right\|_1 \tag{4}$$

where $f_{MAN}(\cdot)$ is the proposed network and Θ denotes its trainable parameters.

3.2. Multi-scale Attention Block (MAB)

Inspired by recent breakthroughs in transformers, we reconsider the basic convolutional block for feature extraction in the SISR task. In contrast to many RCAN [58]-style blocks, the proposed block incorporates MetaFormer [53]-style functionality to achieve a promising extraction result. As shown in Fig. 3, MAB consists of two components: the multi-scale large kernel attention (MLKA) module and the gate spatial attention unit (GSAU).

Given input feature X, the whole process of MAB is:

$$\begin{aligned}
&N = LN(X),\\
&X = X + \lambda_1 f_3(MLKA(f_1(N)) \otimes f_2(N)),\\
&N = LN(X),\\
&X = X + \lambda_2 f_6(GSAU(f_4(N), f_5(N))),
\end{aligned} \tag{5}$$

where $LN(\cdot)$ and λ are layer normalization and learnable scaling factors, separately. $MLKA(\cdot)$ and $GSAU(\cdot)$ are proposed MLKA and GSAU modules introduced in the following sections. $\otimes$ and $f_i(\cdot)$ represent element-wise multiplication and i-th point-wise convolution that keeps the dimensions. To preserve instance details and accelerate convergence, we employ layer normalization rather than batch normalization or none normalization.

3.3. Multi-scale Large Kernel Attention (MLKA)

The attention mechanism can force networks to focus on crucial information and ignore irrelevant ones. Previous SR models adopt a series of attention mechanisms, including channel attention (CA) and self-attention (SA), to obtain more informative features. However, these methods fail to simultaneously uptake local information and long-range dependence, and they often consider the attention maps at a fixed reception field. Enlightened by the latest visual attention research [15], we propose multi-scale large kernel attention (MLKA) to resolve these problems by combining large kernel decomposition and multi-scale learning. Specifically, the MLKA consists of three main functions, large kernel attentions (LKA) for establishing interdependence, the multi-scale mechanism for obtaining heterogeneous-scale correlation, and gated aggregation for dynamic recalibration.

Large kernel attention. Given the input feature maps $X \in \mathbb{R}^{C\times H\times W}$, the LKA adaptively builds the long-range relationship by decomposing a $K \times K$ convolution into three components: a $(2d-1)\times(2d-1)$ depth-wise convolution $f_{DW}(\cdot)$, a $\lceil\frac{K}{d}\times\frac{K}{d}\rceil$ depth-wise d-dilation convolution $f_{DWD}(\cdot)$, and a point-wise convolution $f_{PW}(\cdot)$, which can be formulated as:

$$LKA(X) = f_{PW}(f_{DWD}(f_{DW}(X))). \tag{6}$$

Multi-scale mechanism. To learn the attention maps with omni-scale information, we enhance the fixed LKA with the group-wise multi-scale mechanism. Supposing the input feature maps $X \in \mathbb{R}^{C\times H\times W}$, the module first splits it into n-pieces $X_1, X_2, \ldots, X_n$ of $\lfloor\frac{C}{n}\rfloor \times H \times W$. For i-th group of features X_i, an LKA decomposed by $\{K_i, d_i\}$

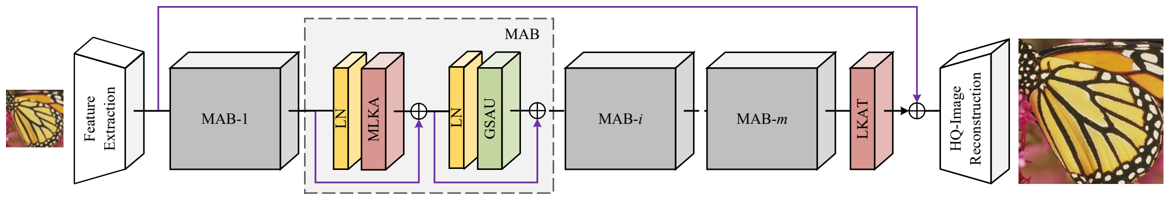

Figure 2. Overview of our multi-scale attention network (MAN).

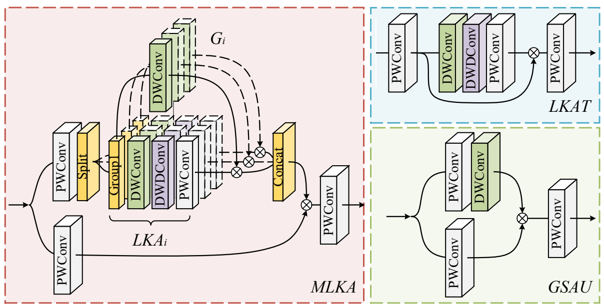

Figure 3. Details of proposed modules.

is utilized to generate a homogeneous scale attention map LKA_i. In detail, we leverage three groups of LKA: $\{7, 2\}$ implemented by 3-5-1, $\{21, 3\}$ by 5-7-1, and $\{35, 4\}$ by 7-9-1, where a-b-1 means cascading $a \times a$ depth-wise, $b \times b$ depth-wise-dilated, and point-wise convolutions.

Gated aggregation. Different from many high-level computer vision tasks, the SR task has a worse tolerance for dilation and partition. As shown in the Fig. 4, although the larger LKA captures wider responses of pixels, the blocking artifacts appear in the generated attention maps of larger LKA. For i-th group input X_i, to avoid the block effect, as well as to learn more local information, we leverage spatial gate to dynamically adapt $LKA_i(\cdot)$ into $MLKA_i(\cdot)$ by:

$$MLKA_i(X_i) = G_i(X_i) \otimes LKA_i(X_i), \tag{7}$$

where $G_i(\cdot)$ is the i-th gate generated by $a_i \times a_i$ depth-wise convolution, and $LKA_i(\cdot)$ is the LKA decomposed by a_i-b_i-1. In Fig. 4, we provide the visual results of the gated aggregation. It can be observed that the block effects are removed from the attention maps and the $MLKA_i$s are more reasonable. In particular, the $MLKA_i$ with larger receptive fields reacts more on long-range dependence while the smaller $MLKA_i$ tends to retain local texture.

Complexity analysis. To compare the complexities of MLKA, LKA, and SA, we present their theoretical floating point operation (FLOPs). Given input $X \in \mathbb{R}^{C \times H \times W}$, the computational cost of $M \times M$ window-based SA is $2M^2HWC$. Within LKA with fixed $\{K, d\}$, the budget of decomposition is $(\lceil \frac{K}{d} \rceil^2 + (2d-1)^2 + C)HWC$. In general, the window size M and kernel size K determine theoretical computational complexity by the quadratic increase. For the proposed MLKA with n groups of $\{K_i, d_i\}$, the total computation is denoted as $(\sum \frac{1}{n}(\lceil \frac{K_i}{d_i} \rceil^2 + 2(2d_i - 1)^2 + \frac{C}{n}) + C)HWC$. The blue terms are the additional calculations brought by gated aggregation and projection. Since the feature is separated into small groups (divided by n), we can control the computational cost while flexibly employing varied kernels to capture both local and global information.

3.4. Gated Spatial Attention Unit (GSAU)

In transformer blocks, a feed-forward network (FFN) is an essential part of enhancing feature representation. However, the commonly used MLP with wide intermediate channels is too heavy for SR networks, especially for large image inputs. Inspired by [5, 9, 17, 47], we integrate simple spatial attention (SSA) and gated linear unit (GLU) into the proposed GSAU to enable an adaptive gating mechanism and reduce the parameters and calculations.

To capture spatial information more efficiently, we adopt a single layer depth-wise convolution to weight the feature map. Given the dense-transformed X and Y, the key process of GSAU can be represented as:

$$GSAU(X, Y) = f_{DW}(X) \otimes Y, \tag{8}$$

where $f_{DW}(\cdot)$ and $\otimes$ indicate depth-wise convolution and element-wise multiplication, respectively. By applying a spatial gate, the GSAU can remove the nonlinear layer and capture local continuity under considerate complexity.

3.5. Large Kernel Attention Tail (LKAT)

In previous SR networks [8, 30, 31, 41, 58], the vanilla convolution layer is widely used as the tail of the deep extraction backbone. However, it has a flaw in establishing long-range connections, therefore limiting the representative capability of the finial reconstruction feature. In order to summarize more reasonable information from the stacked MABs, we introduce the 7-9-1 LKA in the tail module. Concretely, the LKA is wrapped by two 1×1 convolutions as depicted in Fig. 3.

Table 2. Ablation studies on components of MAN. The impact of LKAT, Multi-scale mechanism, and GSAU are shown upon MAN-tiny/light ($\times 2$). In detail, we replace LKAT with convolution layer, Multi-scale with LKA (5-7-1), and GSAU with MLP.

Method	LKAT	Multi-Scale	GSAU	#Params	#Mult-Adds	Set5 [2]		Set14 [55]		BSD100 [38]		Urban100 [18]	
						PSNR (Δ)	SSIM	PSNR (Δ)	SSIM	PSNR (Δ)	SSIM	PSNR (Δ)	SSIM
MAN-tiny				108K	24.2G	37.71	0.9594	33.24	0.9148	31.97	0.8973	31.08	0.9178
	✓			121K	27.2G	37.75 (↑0.04)	0.9595	33.27 (↑0.03)	0.9154	32.01 (↑0.04)	0.8979	31.25 (↑0.17)	0.9199
	✓	✓		143K	29.9G	37.77 (↑0.06)	0.9596	33.30 (↑0.06)	0.9153	32.01 (↑0.04)	0.8979	31.30 (↑0.22)	0.9202
	✓	✓	✓	134K	29.9G	37.79 (↑0.08)	0.9598	33.31 (↑0.07)	0.9155	32.02 (↑0.05)	0.8980	31.33 (↑0.25)	0.9206
MAN-light				737K	165.8G	38.01	0.9605	33.55	0.9179	32.23	0.9005	32.14	0.9287
	✓			756K	170.0G	38.05 (↑0.04)	0.9607	33.60 (↑0.05)	0.9182	32.25 (↑0.02)	0.9007	32.23 (↑0.09)	0.9297
	✓	✓		835K	187.6G	38.07 (↑0.06)	0.9607	33.62 (↑0.07)	0.9181	32.26 (↑0.03)	0.9009	32.42 (↑0.28)	0.9308
	✓	✓	✓	820K	184.0G	38.07 (↑0.06)	0.9608	33.69 (↑0.14)	0.9188	32.29 (↑0.06)	0.9012	32.43 (↑0.29)	0.9316

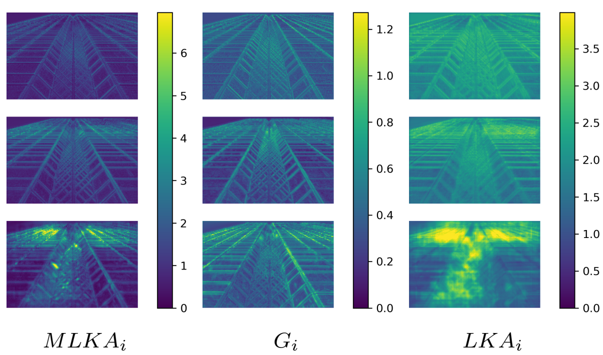

Figure 4. Visual activation maps of Eq. (8) in the 16-th layer of MAN-light. From top to bottom are the corresponding feature maps of 3-5-1, 5-7-1, and 7-9-1, respectively.

4. Experiments

4.1. Datasets and Metrics

Following latest works [30, 32, 37], we utilize DIV2K [1] and Flickr2K [31], which contain 800 and 2650 images, to train our models. For the testing phase, we evaluate our method on five commonly used datasets: Set5 [2], Set14 [55], BSD100 [38], Urban100 [18], and Manga109 [39]. In addition, two standard evaluation metrics, peak-signal-to-noise-ratio (PSNR) and the structural similarity index (SSIM) [50], are applied in Y channel of the YCbCr images to measure the quality of restoration.

4.2. Implementation Details

For more comprehensive evaluations of the proposed methods, we train three different versions of MAN: tiny, light, and base, to resolve the classic SR tasks under different complexities. Following [30], we stack 5/24/36 MABs and set the channel width to 48/60/180 in the corresponding tiny/light/base MAN. Three multi-scale decomposition modes are utilized in MLKA, listed as 3-5-1, 5-7-1, and 7-9-1. The 7×7 depth-wise convolution is used in the GSAU.

In the training stage, the training pairs are augmented by horizontal flips and random rotations of 90°, 180°, and 270°. The {patch size, batch size} is set to $\{48 \times 48, 32\}$ and $\{64 \times 64, 16\}$ in the training-from-scratch and fine-turning stage, respectively. The ℓ_1 loss is adopted to discriminate the pixel-wise restoration quality for fairness. All models are trained using the Adam optimizer [23] with β_1=0.9, β_2=0.99. The learning rate is initialized as 5×10^{-4} and scheduled by cosine annealing learning for 1600K iterations in training anew while setting as 1×10^{-4} for 800K in fine-turning. All experiments are conducted by Pytorch [42] framework on 4 Nvidia RTX 3090 GPUs.

Table 3. Ablation study on block structure.

Method	#Params	#FLOPs	Set5 [2]		BSD100 [38]	
			PSNR	SSIM	PSNR	SSIM
RCAN-style	924K	53.0G	32.16	0.8945	27.60	0.7371
Metaformer-style	840K	47.1G	32.33	0.8967	27.67	0.7396

4.3. Ablation Studies

In this section, we validate the effectiveness of the proposed components from coarse to fine. In detail, we first investigate the combination of all proposed modules and then examine each of them individually. *For fairness and simplicity, we adopt the same training for 200K iterations.*

Overall study on components of MAN. In Tab. 2, we present the results of deploying the proposed components on our tiny and light networks. In general, the best performances are achieved by employing all proposed modules. Specifically, 0.25 dB and 0.29 dB promoting on Urban100 [18] can be observed in MAN-tiny and MAN-light, while the parameters and calculations increase negligibly. Among these components, the LKAT module and multi-scale mechanism are more important to enhance quality. Without any of them, the PSNR will drop by 0.09 dB. The GSAU is an economical replacement for MLP. It reduces 15K parameters and 3.6G calculations while bringing significant improvements across all datasets.

Study on block structures. Within MAB, we choose the emerging metaformer style rather than the RCAN-style structure to deploy MLKA. To fully explore their effectiveness, we implement and compare two versions of MABs

Table 4. Ablation studies of multi-scale and decomposition type (LKA/MLKA). The results are tested on MAN-light for ×4 SR task. The LKA (5-7-1) from VAN [15] is used as the baseline for comparison.

Method	Decomposition			#Params	#FLOPs	Set5 [2]		Set14 [55]	
	3-5-1	5-7-1	7-9-1			PSNR (Δ)	SSIM (Δ)	PSNR (Δ)	SSIM (Δ)
LKA	✓			703K	39.4G	32.23 (↓0.04)	0.8956 (↓.0007)	28.70 (↓0.02)	0.7842 (↓.0004)
		✓		761K	42.7G	32.27	0.8963	28.72	0.7846
			✓	841K	47.4G	32.25 (↓0.02)	0.8958 (↓.0005)	28.71 (↓0.01)	0.7845 (↓.0001)
MLKA	✓	✓		803K	45.0G	32.32 (↑0.05)	0.8968 (↑.0005)	28.72 (↑0.00)	0.7848 (↑.0002)
		✓	✓	900K	50.6G	32.33 (↑0.06)	0.8968 (↑.0005)	28.74 (↑0.02)	0.7852 (↑.0006)
	✓	✓	✓	840K	47.1G	32.33 (↑0.06)	0.8967 (↑.0004)	28.76 (↑0.04)	0.7856 (↑.0010)

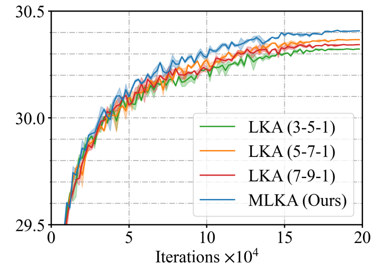

Table 5. Ablation study on varied FFNs.

Method	#Params	#FLOPs	Set5	Set14	B100	U100
MLP [13]	854K	48.0G	32.31	28.73	27.65	26.26
Simple-Gate [5]	768K	43.1G	32.28	28.74	27.66	26.28
CFF [47]	1140K	64.3G	32.35	28.76	27.67	26.34
GSAU	840K	47.1G	32.33	28.76	27.67	26.31

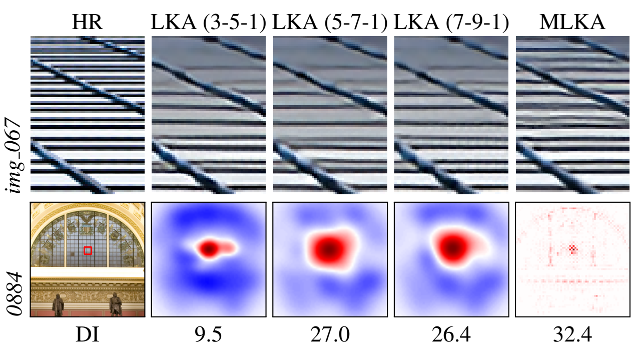

Figure 5. Comparisons between LKA and MLKA. Rows 1: visual comparisons. Row 2: Cols 2-4: The difference maps of the area of interest between LKA and MLKA. The red regions are noticed by almost both LKA and MLKA while the blue represent additional interest areas of MLKA. Col 5: LAM results of MLKA.

in the Tab. 3. The experimental results indicate that the transformer-style MAB surpasses the RCAN-style one by a large margin. On Set5 [2], the PSNR is increased from 32.15 dB to 32.33 dB by employing the transformer structure. The results show the transformer-style MAB is more efficient in balancing the performance and computations.

Study on MLKA. To justify our design of MLKA, we conduct ablation experiments on multi-scale and kernel decomposition. Specifically, we consider three LKA and three MLKA implementations in Tab. 4. We first investigate the effects of kernel size on LKA. When we increase the kernel size, the PSNR decreases after an initial increase, which is inconsistent with high-level tasks [15]. It is due to long-range correlation and local textural information is indispensable in image restoration tasks. Up until this point, we introduce MLKA to refine features at comprehensive scales. In Tab. 4, we also illustrate the training evaluation results of LKAs and proposed MLKA. The MLKA outperforms other LKAs throughout the training phase. For the visual comparison and local attribution map (LAM) [14] results shown in Fig. 5, MLKA brings higher DI and more activated pixels, thereby helping to recover more details on both images from Urban100. In addition, we briefly discuss MLKA of different combinations. These results suggest the MLKA with all three decomposition types can trade off parameters, computations, and performance.

Study on FFNs. To further confirm the efficiency of the proposed GSAU, we compare it with some other FFNs. In Tab. 5, we validate four advanced designs: MLP, Simple Gate, CFF, and our GSAU. The GSAU delivers comparable performance to the powerful CFF while occupying 73% of the parameters and calculations, showing effectiveness.

4.4. Comparisons with classical SR models

To validate the effectiveness of our MAN, we compare our normal model to several SOTA classical ConvNets [8, 37, 40, 41, 58, 59]. We also add SwinIR [30] for reference. In Tab. 6, the quantitative results show that our MAN exceeds other convolutional methods to a large extent. The maximum improvement on PSNR reaches 0.69 dB for ×2, 0.77 dB for ×3, and 0.81 dB for ×4. Moreover, we compare our MAN with SwinIR. For ×2, our MAN achieves competitive or even better performance than SwinIR. The PSNR value on Manga109 is boosted from 39.92 dB to 40.02 dB. For ×4, MAN is slightly behind SwinIR because the latter uses the ×2 model as the pre-trained model. More importantly, MAN is significantly smaller than existing methods.

In Fig. 6, we also visualize the qualitative results of several models on the Urban100 (×4) benchmark dataset. For *img_024*, compared with other models generating the distorted fence, our MAN rebuilds a clear structure from the blurred input. Similarly, in *img_073*, MAN is the only model that restores the windows of the building.

4.5. Comparisons with tiny/light SR models

To verify the efficiency and scalability of our MAN, we compare MAN-tiny and MAN-light to some state-of-the-art tiny [12, 26, 27, 44, 56] and lightweight [19, 30, 36, 52, 57]

Table 6. Quantitative comparison (average PSNR/SSIM) with state-of-the-art ConvNets for **classical image SR**. The best and second best performances are **highlighted** and underlined, respectively. "†" and '+" indicate using pre-training and self-ensemble strategy, respectively.

Method	Scale	#Params	#FLOPs	Set5 [2]		Set14 [55]		BSD100 [38]		Urban100 [18]		Manga109 [39]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN [58]	×2	15.4M	3.5T	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [8]	×2	15.9M	3.1T	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
HAN [41]	×2	63.6M	14.6T	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
IGNN [59]	×2	49.5M	-	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
NLSA [40]	×2	41.8M	9.6T	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
DFSA+ [37]	×2	-	-	38.38	0.9620	34.33	0.9232	32.50	0.9036	33.66	0.9412	39.98	0.9798
MAN	×2	8.7M	1.7T	38.42	0.9622	34.40	0.9242	32.53	0.9043	33.73	0.9422	40.02	0.9801
MAN+	×2	8.7M	-	**38.44**	**0.9623**	**34.49**	**0.9248**	**32.55**	**0.9045**	**33.86**	**0.9430**	**40.13**	**0.9804**
SwinIR† [30]	×2	11.8M	2.3T	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
RCAN [58]	×3	15.6M	1.6T	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [8]	×3	15.9M	1.6T	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
HAN [41]	×3	64.3M	6.5T	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
IGNN [59]	×3	49.5M	-	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
NLSA [40]	×3	44.7M	4.6T	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
DFSA+ [37]	×3	-	-	34.92	0.9312	30.83	0.8507	29.42	0.8128	29.44	0.8761	35.07	0.9525
MAN	×3	8.7M	0.8T	34.91	0.9312	30.88	0.8514	29.43	0.8138	29.52	0.8782	35.06	0.9526
MAN+	×3	8.7M	-	**34.97**	**0.9315**	**30.91**	**0.8522**	**29.47**	**0.8144**	**29.65**	**0.8799**	**35.21**	**0.9533**
SwinIR† [30]	×3	11.9M	1.0T	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
RCAN [58]	×4	15.6M	0.9T	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [8]	×4	15.9M	0.9T	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
HAN [41]	×4	64.2M	3.8T	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IGNN [59]	×4	49.5M	-	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
NLSA [40]	×4	44.2M	3.0T	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
DFSA+ [37]	×4	-	-	32.79	0.9019	29.06	0.7922	27.87	0.7458	27.17	0.8163	31.88	**0.9266**
MAN	×4	8.7M	0.4T	32.81	0.9024	29.07	0.7934	27.90	0.7477	27.26	0.8197	31.92	0.9230
MAN+	×4	8.7M	-	**32.87**	**0.9030**	**29.12**	**0.7941**	**27.93**	**0.7483**	**27.39**	**0.8223**	**32.13**	0.9248
SwinIR† [30]	×4	11.9M	0.6T	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260

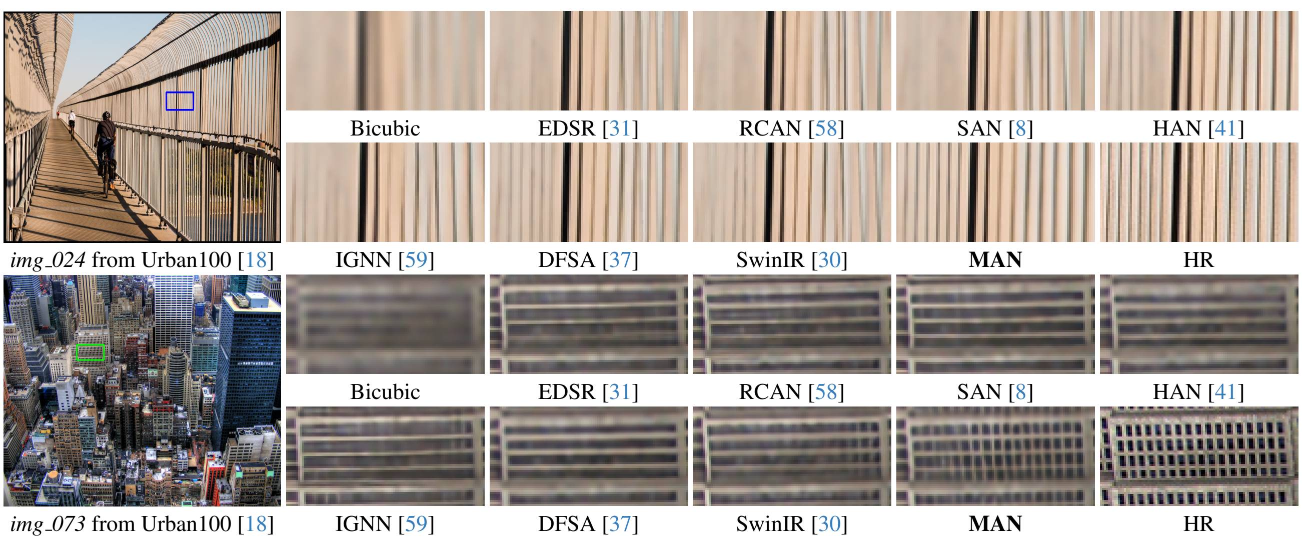

Figure 6. Visual comparison for classical SR models with an upscaling factor ×4.

SR models. Tab. 7 presents the numerical results that our MAN-tiny/light outperforms all other tiny/lightweight methods. Specifically, MAN-tiny exceeds second place by about 0.2 dB on Set5, Urban100, and Manga109, and around 0.07 dB on Set14 and BSD100. We also list EDSR-baseline [31] for reference. Our tiny model has less than 150K parameters but achieves a similar restoration quality with EDSR-baseline, which is 10× larger than ours. Similarly, our MAN-light surpasses both CNN-based and transformer-based SR models. In comparison with IMDN (CNN) and SwinIR-light/ELAN-light (Transformer), our model leads by 0.66 dB/0.23 dB on Urban100 (×4) bench-

Table 7. Quantitative comparison (average PSNR/SSIM) with state-of-the-art approaches for **tiny/light image SR** on benchmark datasets (×4). The best and second best performances are **highlighted** and underlined, respectively.

Method	Scale	#Params	#FLOPs	Set5 [2]		Set14 [55]		BSD100 [38]		Urban100 [18]		Manga109 [39]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FSRCNN [12]	×4	12K	4.6G	30.71	0.8657	27.59	0.7535	26.98	0.7150	24.62	0.7280	27.90	0.8517
LAPAR-C [26]	×4	115K	25.0G	31.72	0.8884	28.31	0.7740	27.40	0.7292	25.49	0.7651	29.50	0.8951
ECBSR-M10C32 [56]	×4	98K	5.7G	31.66	0.8911	28.15	0.7776	27.34	**0.7363**	25.41	0.7653	-	-
ShuffleMixer-tiny [44]	×4	113K	8.0G	31.88	0.8912	28.46	0.7779	27.45	0.7313	25.66	0.7690	29.96	0.9006
ETDS-L [3]	×4	170K	9.8G	31.69	0.8889	28.31	0.7751	27.37	0.7302	25.47	0.7643	-	-
MAN-tiny	×4	150K	8.4G	**32.07**	**0.8930**	**28.53**	**0.7801**	**27.51**	0.7345	**25.84**	**0.7786**	**30.18**	**0.9047**
EDSR-baseline [31]	×4	1518K	114G	32.09	0.8938	28.58	0.7813	27.57	0.7357	26.04	0.7849	30.35	0.9067
IMDN [19]	×4	715K	40.9G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
LatticeNet [36]	×4	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
DIPNet [52]	×4	543K	-	32.20	0.8950	28.58	0.7811	27.59	0.7364	26.16	0.7879	30.53	0.9087
SwinIR-light [30]	×4	897K	49.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN-light [57]	×4	601K	43.2G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
MAN-light	×4	840K	47.1G	**32.50**	**0.8988**	**28.87**	**0.7885**	**27.77**	**0.7429**	**26.70**	**0.8052**	**31.25**	**0.9170**
EDSR [31]	×4	43090K	2895G	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148

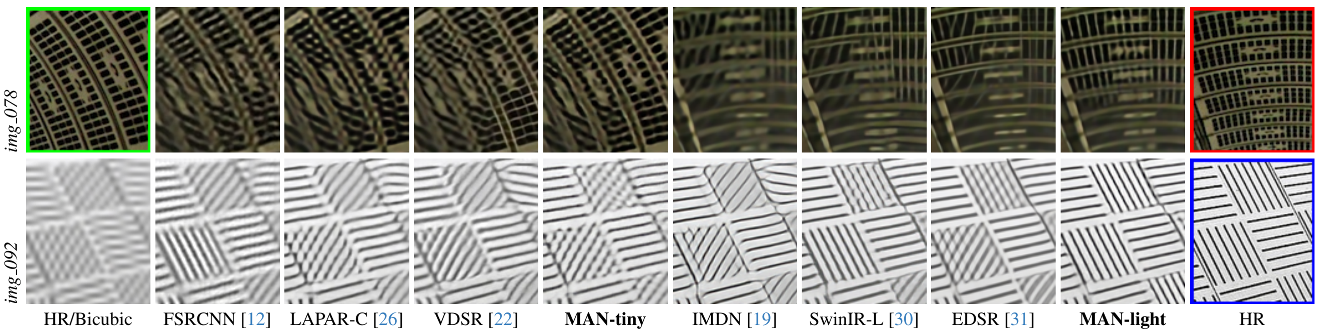

Figure 7. Visual comparison for tiny/lightweight SR models with an upscaling factor ×4.

Method	#Params	#FLOPs	Set5	B100	U100	M109
IPT [4]	115.5M	-	38.37	32.48	33.76	-
EDT-B [27]	11.5M	37.6G	38.45	32.52	33.80	39.93
HAT [6]	20.8M	103.7G†	38.63	32.62	34.45	40.26
DAT [7]	14.7M	245.4G†	38.58	32.61	34.37	40.33
MAN	8.7M	19.8G	38.42	32.53	33.73	40.02

Table 8. Quantitative comparison with sota transformers (×2). #FLOPs are calculated with 48×48/64×64† inputs.

mark. Moreover, our MAN-light is superior to traditional performance-oriented EDSR. In detail, the proposed model takes only 2% of the parameters and computations of EDSR while having high PSNR on all benchmarks.

In Fig. 7, we also exhibit the visual results of several tiny/lightweight models on Urban100 (×4). For *img_078*, the tiny and light models are tested with the patches framed by green and red boxes, respectively. Generally, MANs can restore the texture better and clearer than other methods.

4.6. Comparisons with SR Transformers

Although MANs achieve remarkable improvement compared to existing ConvNet-based models, more comparison with some transformer-based approaches is essential. In Tab. 8, we include the competitive IPT [4], EDT [27], HAT [6], and DAT [7] for discussion. MAN achieves similar quality as EDT-B with only 75% params and 52% FLOPs (48×48 input). The HAT and DAT are much larger models than EDT or MAN, which perform superior to both. In a word, MAN can perform on par with or even better than these transformer-based methods (SwinIR, EDT) with similar model sizes, showing ConvNet's vitality in low-level.

5. Conclusion

This paper proposes a multi-scale attention network (MAN) for super-resolution under multiple complexities. MAN adopts transformer-style blocks for better modeling representation. To effectively and flexibly establish long-range correlations among various regions, we develop multi-scale large kernel attention (MLKA) that combines large kernel decomposition and multi-scale mechanisms. Furthermore, we propose a simplified feed-forward network (GSAU) that integrates gate mechanisms and spatial attention to activate local information and reduce model complexity. Extensive experiments have demonstrated that our CNN-based MAN can achieve better performance than previous SOTA ConvNets and keep pace with transformer-based methods in a more efficient manner.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 1122–1131, Honolulu, USA, 2017. IEEE Computer Society. 5

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, pages 1–10, Surrey, UK, 2012. BMVA Press. 5, 6, 7, 8

[3] Jiahao Chao, Zhou Zhou, Hongfan Gao, Jiali Gong, Zhengfeng Yang, Zhenbing Zeng, and Lydia Dehbi. Equivalent transformation and dual stream network construction for mobile image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14102–14111, 2023. 8

[4] 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 *CVPR*, pages 12299–12310, virtual, 2021. Computer Vision Foundation / IEEE. 1, 2, 3, 8

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 2, 4, 6

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 1, 8

[7] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12312–12321, 2023. 3, 8

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, Long Beach, USA, 2019. Computer Vision Foundation / IEEE. 2, 4, 6, 7

[9] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, pages 933–941, Sydney, Australia, 2017. PMLR. 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, Miami, USA, 2009. IEEE Computer Society. 1

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2016. 1, 2

[12] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, pages 391–407, Amsterdam, The Netherlands, 2016. Springer. 1, 2, 6, 8

[13] 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 *ICLR*, Virtual Event, Austria, 2021. 1, 6

[14] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *CVPR*, pages 9199–9208, 2021. 6

[15] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *arXiv preprint arXiv:2202.09741*, 2022. 2, 3, 6

[16] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *CVPR*, pages 449–456, Colorado Springs, USA, 2011. 1

[17] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V Le. Transformer quality in linear time. *arXiv preprint arXiv:2202.10447*, 2022. 4

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, Boston, MA, USA, 2015. IEEE Computer Society. 5, 7, 8

[19] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, pages 2024–2032, Nice, France, 2019. ACM. 2, 6, 8

[20] Kui Jiang, Zhongyuan Wang, Peng Yi, and Junjun Jiang. Hierarchical dense recursive network for image super-resolution. *PR*, 107:107475, 2020. 1

[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, Las Vegas, USA, 2016. IEEE Computer Society. 2

[22] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, Las Vegas, NV, USA, 2016. IEEE Computer Society. 1, 2, 8

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, San Diego, USA, 2015. 5

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, pages 5835–5843, Honolulu, USA, 2017. IEEE Computer Society. 2

[25] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, pages 527–542, Munich, Germany, 2018. Springer. 1

[26] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. LAPAR: linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. In *NeurIPS*, virtual, 2020. 6, 8

[27] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 1, 6, 8

[28] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 3

[29] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 1

[30] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, Montreal, Canada, 2021. IEEE. 1, 2, 3, 4, 5, 6, 7, 8

[31] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 1132–1140. IEEE Computer Society, 2017. 1, 2, 3, 4, 5, 7, 8

[32] Zudi Lin, Prateek Garg, Atmadeep Banerjee, Salma Abdel Magid, Deqing Sun, Yulun Zhang, Luc Van Gool, Donglai Wei, and Hanspeter Pfister. Revisiting rcan: Improved training for image super-resolution. *arXiv preprint arXiv:2201.11279*, 2022. 1, 5

[33] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, pages 2356–2365, Seattle, USA, 2020. Computer Vision Foundation / IEEE. 1, 2

[34] 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 *ICCV*, pages 9992–10002, Montreal, Canada, 2021. IEEE. 3

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 2

[36] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *ECCV*, pages 272–289, Glasgow, UK, 2020. Springer. 6, 8

[37] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV*, pages 4268–4277, Montreal, Canada, 2021. IEEE. 5, 6, 7

[38] David R. Martin, Charless C. 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 *ICCV*, pages 416–425, Vancouver, Canada, 2001. IEEE Computer Society. 5, 7, 8

[39] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multim. Tools Appl.*, 76(20):21811–21838, 2017. 1, 5, 7, 8

[40] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, pages 3517–3526, virtual, 2021. Computer Vision Foundation / IEEE. 1, 3, 6, 7

[41] 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 *ECCV*, pages 191–207, Glasgow, UK, 2020. Springer. 2, 3, 4, 6, 7

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, Vancouver, Canada, 2019. 5

[43] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *CVPR*, pages 1–8, Anchorage, USA, 2008. 1

[44] Long Sun, Jinshan Pan, and Jinhui Tang. Shufflemixer: An efficient convnet for image super-resolution. *arXiv preprint arXiv:2205.15175*, 2022. 6, 8

[45] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, pages 2790–2798, Honolulu, HI, USA, 2017. IEEE Computer Society. 2

[46] Yu-Wing Tai, Shuaicheng Liu, Michael S Brown, and Stephen Lin. Super resolution using edge prior and single image detail synthesis. In *CVPR*, pages 2400–2407, San Francisco, USA, 2010. 1

[47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media*, 8(3):415–424, 2022. 2, 3, 4, 6

[48] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *CVPRW*, pages 777–785, 2022. 2

[49] Yan Wang, Shijie Zhao, Yi Liu, Junlin Li, and Li Zhang. Camixersr: Only details need more "attention". *arXiv preprint arXiv:2402.19289*, 2024. 1

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

[51] Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. Efficient non-local contrastive attention for image super-resolution. *arXiv preprint arXiv:2201.03794*, 2022. 1, 2, 3

[52] Lei Yu, Xinpeng Li, Youwei Li, Ting Jiang, Qi Wu, Haoqiang Fan, and Shuaicheng Liu. Dipnet: Efficiency distillation and iterative pruning for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1701, 2023. 6, 8

[53] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. *arXiv preprint arXiv:2111.11418*, 2021. 2, 3

[54] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yulun Zhang, and Radu Timofte. See more details: Efficient image super-resolution by experts mining. *arXiv preprint arXiv:2402.03412*, 2024. 2

[55] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces - 7th International Conference*, pages 711–730, Avignon, France, 2010. Springer. 5, 6, 7, 8

[56] Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *ACM MM*, pages 4034–4043, Virtual Event, China, 2021. ACM. 6, 8

[57] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. *arXiv preprint arXiv:2203.06697*, 2022. 6, 8

[58] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 294–310, Munich, Germany, 2018. Springer. 1, 2, 3, 4, 6, 7

[59] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *NeurIPS*, virtual, 2020. 6, 7