Omni Aggregation Networks for Lightweight Image Super-Resolution

Hang Wang[2*] Xuanhong Chen[1*] Bingbing Ni[1,2†] Yutian Liu[1] Jinfan Liu[1]
[1]Shanghai Jiao Tong University, Shanghai 200240, China [2] Huawei
francis970625@gmail.com {chen19910528,nibingbing,stau7001,ljflnjz}@sjtu.edu.cn

Abstract

*While lightweight ViT framework has made tremendous progress in image super-resolution, its uni-dimensional self-attention modeling, as well as homogeneous aggregation scheme, limit its effective receptive field (ERF) to include more comprehensive interactions from both spatial and channel dimensions. To tackle these drawbacks, this work proposes two enhanced components under a new Omni-SR architecture. First, an Omni Self-Attention (OSA) block is proposed based on dense interaction principle, which can simultaneously model pixel-interaction from both spatial and channel dimensions, mining the potential correlations across omni-axis (i.e., spatial and channel). Coupling with mainstream window partitioning strategies, OSA can achieve superior performance with compelling computational budgets. Second, a multi-scale interaction scheme is proposed to mitigate sub-optimal ERF (i.e., premature saturation) in shallow models, which facilitates local propagation and meso-/global-scale interactions, rendering an omni-scale aggregation building block. Extensive experiments demonstrate that Omni-SR achieves record-high performance on lightweight super-resolution benchmarks (e.g., **26.95**dB@Urban100 ×4 with only **792K** parameters). Our code is available at https://github.com/Francis0625/Omni-SR.*

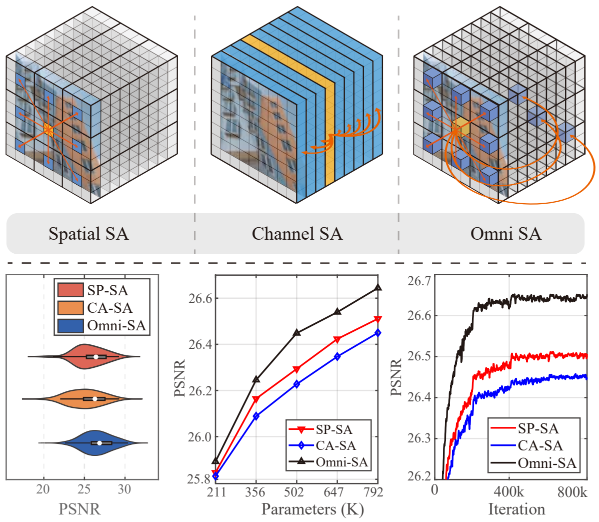


Figure 1. Typical self-attention schemes [31,59] can only perform uni-dimensional (e.g., spatial-only) interactions, and.

1. Introduction

Image super-resolution (SR) is a long-standing low-level problem that aims to recover high-resolution (HR) images from degraded low-resolution (LR) inputs. Recently, vision transformer [14, 51] based (i.e., ViT-based) SR frameworks [5, 31] have emerged, showing significant performance gains compared to previously dominant Convolutional Neural Networks (CNNs) [66]. However, most attempts [31] are devoted to improving the large-scale ViT-based models, while the development of lightweight ViTs (typically, less than 1M parameters) remains fraught with difficulties. This paper focuses on boosting the restoration performance of lightweight ViT-based frameworks.

Two difficulties hinder the development of lightweight ViT-based models: 1) **Uni-dimensional aggregation operators** (i.e., spatial [31] only or channel [59] only) imprisons the full potential of self-attention operators. Contemporary self-attention generally realizes the interaction between pixels by calculating the cross-covariance of the spatial direction (i.e., width and height) and exchanges context information in a channel-separated manner. This interaction scheme ignores the explicit use of channel information. However, recent evidences [59] and our practice show that self-attention in the channel dimension (i.e., computationally more compact than spatial self-attention) is also crucial in low-level tasks. 2) **Homogeneous aggregation schemes** (i.e., Simple hierarchical stacking of single operators, e.g., convolution, self-attention) neglect abundant texture patterns of multi-scales, which is urgently needed in SR task. Specifically, a single operator is only sensitive to information of one scale [6, 12], e.g., self-attention is sensitive to long-term information and pays little attention to lo-

[*] Equal Contribution.
[†] Corresponding author: Bingbing Ni.

cal information. Additionally, stacking of homogeneous operators proves to be inefficient and suffers from premature saturation of the interaction range [8], which is reflected as a suboptimal effective receptive field. The above problem is exacerbated in lightweight models because lightweight models cannot stack enough layers.

In order to solve the above problems and pursue higher performance, this work proposes a novel omni-dimension feature aggregation scheme called *Omni Self-Attention* (OSA) exploiting both spatial and channel axis information in a simultaneous manner (i.e., extends the interaction into three-dimensional space), which offers higher-order receptive field information, as shown in Figure 1. Unlike scalar-based (a group of important coefficient) channel interaction [19], OSA enables comprehensive information propagation and interaction by cascading computation of the cross-covariance matrices between spatial/channel dimensions. The proposed OSA module can be plugged into any mainstream self-attention variants (e.g., Swin [35], Halo [50]), which provides a finer granularity of important encoding (compared to the vanilla channel attention [19]), achieving a perceptible improvement in contextual aggregation capabilities. Furthermore, a multi-scale hierarchical aggregation block, named *Omni-Scale Aggregation Group* (i.e., OSAG for short), is presented to achieve tailored encoding of varying scales of texture patterns. Specifically, OSAG builds three cascaded aggregators: local convolution (for local details), meso self-attention (focusing on mid-scale pattern processing), and global self-attention (pursuing global context understanding), rendering an omni-scale (i.e., local-/meso-/global-scale simultaneously) feature extraction capability. Compared to the homogenized feature extraction schemes [28, 31], our OSAG is able to mine richer information producing features with higher information entropy. Coupling with the above two designs, we establish a new ViT-based framework for lightweight super-resolution, called *Omni-SR*, which exhibits superior restoration performance as well as covers a larger interaction range while maintaining an attractive model size, i.e., $792K$.

We extensively experiment with the proposed framework with both qualitative and quantitative evaluations on mainstream open-source image super-resolution datasets. It is demonstrated that our framework achieves state-of-the-art performance at the lightweight model scale (e.g., Urban100×4: 26.95dB, Manga109×4: 31.50dB). More importantly, compared to existing ViT-based super-solution frameworks, our framework shows superior optimization properties (e.g., convergence speed, smoother loss landscape), which endow our model with better robustness.

2. Related Works

Image Super-resolution. CNNs have achieved remarkable success in image SR task. SRCNN [13] is the first work to introduce CNNs into the SR field. Many methods [25, 48, 66] employ skip connection to speed up network convergence and improve the reconstruction quality. Channel Attention [66] is also proposed to enhance the representation ability of the SR model. In order to obtain better reconstruction quality with limited computing resources, several methods [23,38,42,47] explore lightweight architectural design. DRCN [26] utilizes the recursive operation to reduce the number of parameters. DRRN [47] introduces global and local residual learning on the basis of DRCN to accelerate training and improve the quality of details. CARN [1] employs cascading mechanism upon a residual network. IMDN [22] proposes an information multi-distillation block to archive better time performance. Another line of research is to utilize model compression techniques, e.g., knowledge distillation [15, 17, 65] and neural architecture search [11]) to reduce computing costs. Recently, a series of transformer-based SR models [5,8,31,37] emerge with superior performance. Chen*et al.* [5] develop a pre-trained model for the low-level computer vision task using the transformer architecture. Based on Swin transformer [35], SwinIR [31] proposes a three-stage framework, refreshing the state-of-the-art of SR task. More recently, some works [5, 30] explore ImageNet pre-training strategy to further enhance SR performance.

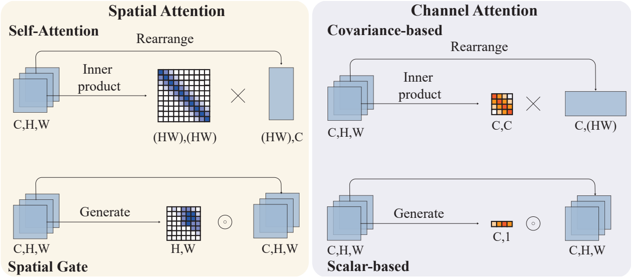


Figure 2. Illustration of spatial attention and channel attention. These typical attention paradigms only model uni-dimensional (i.e., spatial-only / channel-only) interaction.

Lightweight Vision Transformer. Due to the urgent demands for applying networks to resource-constrained devices, lightweight vision transformer [14, 51] has attracted widespread attention. Many attempts [7,9,10,37,41,43,57, 62] have been made to develop lightweight ViTs with comparable performance. A series of methods focus on combining convolutions with transformers to learn both local and global representations. For instance, LVT [57] introduces convolution in self-attention to enrich low-level features. MobileViT [41] replaces matrix multiplication in convolutions with transformer layers to learn global representations. Similarly, EdgeViTs [43] employs an information exchange bottleneck for full spatial interactions. Different from interpreting convolutions into vision transformers, LightViT [21] proposes aggregated self-attention for better

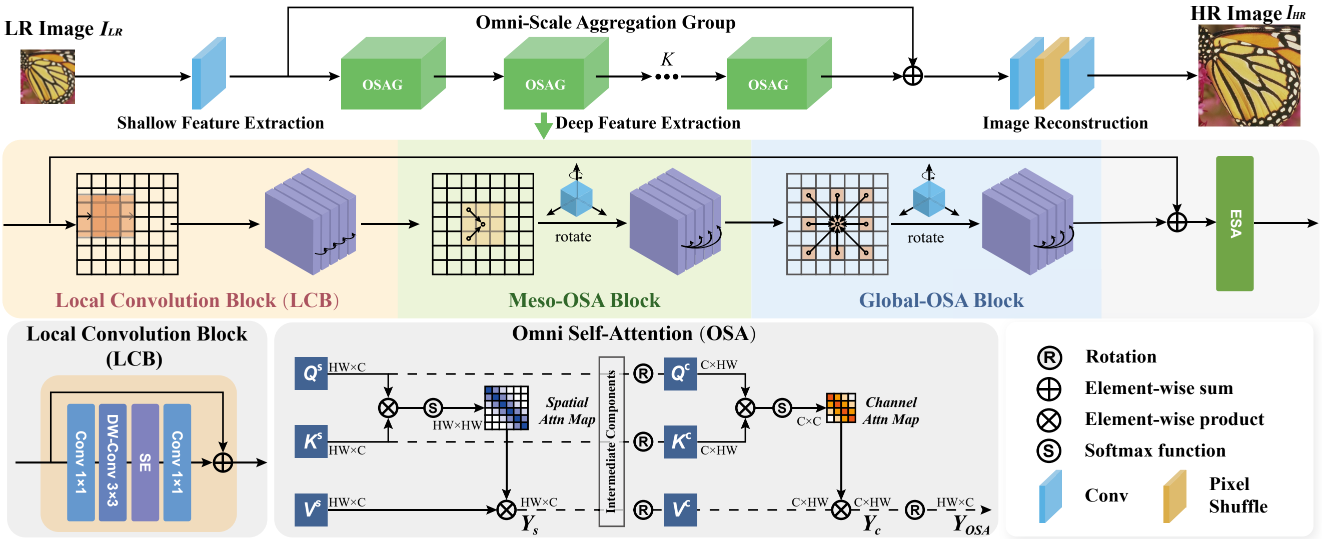


Figure 3. The overall architecture of the proposed Omni-SR framework and structure of OSAG and Omni Self-Attention (OSA).

information aggregation. In this work, we resort to ViT architecture to achieve lightweight and accurate SR.

3. Methodology

3.1. Attention Mechanisms in Super-Resolution

Two attention paradigms are widely adopted in SR to assist in analyzing and aggregating comprehensive patterns.

Spatial Attention. Spatial attention can be regarded as an anisotropic selection process. Spatial self-attention [37, 51] and spatial gate [10, 58] are predominantly applied. As shown in Figure 2, spatial self-attention calculates the cross-covariance along the spatial dimension, and the spatial gate generates the channel-separated masks. Neither of them can transmit information between channels.

Channel Attention. There are two categories of channel attention, i.e., scalar-based [19] and covariance-based [59], proposed to perform channel recalibration or transmit patterns among channels. As shown in Figure 2, the former predicts a group of importance scalars to weigh different channels, while the latter computes a cross-covariance matrix to enable channel re-weighting and information transmission simultaneously. Compared to spatial attention, channel attention handles spatial dimension isotropically, and thus, the complexity is significantly reduced, which also impairs the accuracy of aggregation.

Several attempts [44, 55] have proved that both spatial attention and channel attention are beneficial for SR task and their characteristics are complementary to each other, thus integrating them in a computationally compact way would bring notable benefits in expressive capability.

3.2. Omni Self-Attention Block

To mine all the correlations hidden in the latent variables, we propose a novel self-attention paradigm called Omni Self-Attention (OSA) block. Unlike existing self-attention paradigms (e.g., spatial self-attention [5, 37, 51]) that only indulge in unidimensional processing, OSA establishes the spatial and channel context simultaneously. The obtained two-dimensional relationship is highly necessary and beneficial, especially for lightweight models. On the one hand, as the network deepens, important information is scattered into different channels [19], and it is critical to deal with them in a timely manner. On the other hand, although spatial self-attention takes advantage of the channel dimension in calculating the covariance, it does not transmit the information between channels (refer to Sec. 3.1). Given the above conditions, our OSA is designed to transmit both spatial and dimensional information in a compact manner.

The proposed OSA calculates the score matrices corresponding to the space and channel direction through sequential matrix operations and rotation, as illustrated in Figure 3. Specifically, suppose $X \in \mathbb{R}^{HW \times C}$ denotes the input feature, where H and W are the width and height of the input, and C is the channel number. Firstly, X is embedded to query, key and value matrices $Q^s, K^s, V^s \in \mathbb{R}^{HW \times C}$ through linearly projection. We calculate the production of query and key to get the spatial attention map of size $\mathbb{R}^{HW \times HW}$. Then we perform the spatial attention to obtain the intermediate aggregated results. Note that window strategy is usually used to significantly reduce the resource overhead. Next stage, we **rotate** the input query and key matrices to get the transposed query and key matrices $Q^c, K^c \in \mathbb{R}^{C \times HW}$, and also **rotate** the value matrices to get the value matrix $V^c \in \mathbb{R}^{C \times HW}$ for the subsequent channel-wise self-attention. The obtained channel-wise attention map of size $\mathbb{R}^{C \times C}$ models channel-wise relationships. Finally, we get the final aggregated Y_{OSA} by the inverse rotation of the channel attention output Y_c. The whole OSA process is formulated as follows:

$$Q^s = X \cdot \mathcal{W}_q, \;\; K^s = X \cdot \mathcal{W}_k, \;\; V^s = X \cdot \mathcal{W}_v, \tag{1}$$

$$Y_s = \mathcal{A}^s(Q^s, K^s, V^s) = \text{SoftMax}(Q^s K^{sT}) \cdot V^s, \tag{2}$$

$$Q^c = \mathcal{R}(Q'), \quad K^c = \mathcal{R}(K'), \quad V^c = \mathcal{R}(V'), \tag{3}$$

$$Y_c = \mathcal{A}^c(Q^c, K^c, V^c) = \text{SoftMax}(K^c Q^{cT}) \cdot V^c, \tag{4}$$

$$Y_{OSA} = \mathcal{R}^{-1}(Y_c), \tag{5}$$

where $\mathcal{W}_q, \mathcal{W}_k, \mathcal{W}_v$ denote the linear projection matrices for the query, key, and value, respectively. Q', K', V' are the input embedding matrices of channel-wise self-attention, which are embedded from fore spatial self-attention or or copied directly from Q^s, K^s, V^s. $\mathcal{R}(\cdot)$ denotes the rotation operation around spatial axis and $\mathcal{R}^{-1}(\cdot)$ is the inverse rotation. Some normalization factors are omitted for the sake of simplicity. In particular, this design shows compelling properties that can integrate the element-wise results of two matrix operations (i.e., spatial-/channel-matrix operation), thereby enabling omni-axial interactions. Note that our proposed OSA paradigm can be a drop-in replacement of the Swin [31,35] attention block to higher performance with less parameters. Benefiting from the smaller attention map size of channel self-attention, the proposed OSA is less computationally intensive compared to the cascade shifted-window self-attention scheme in Swin.

Discussion with other hybrid attention paradigms. Compared to previous hybrid channel and spatial attention works like CBAM [55] and BAM [44], their scalar-based attention weights only reflect the relative degree of importance, without further inter-pixel information exchange, leading to limited relation modeling capability. Several recent works [8] also incorporate channel attention with spatial Self-attention, but these attempts only resort to scalar weights for channel recalibration, while our OSA paradigm enables channel-wise interaction to mine the potential correlations in omni-axis. Performance comparison of different attention paradigms can be found in Sec. 4.4.

3.3. Omni-Scale Aggregation Group

How to utilize the proposed OSA paradigm to build a high-performance and compact network is another key topic. Although hierarchical stacking of windows-based self-attention (e.g., swin [31, 35]) has become mainstream, various works have found that the window-based paradigms are very inefficient for large-range interactions, especially for shallow networks. It is worth pointing out that large-range interaction can provide a pleasing effective receptive field, which is crucial for improving image restoration performance [37]. Unfortunately, direct global interaction is resource-prohibitive and detracts from local aggregation capabilities. Taking these points into account, we propose an Omni-Scale Aggregation Group (i.e., OSAG for short) to pursue progressive receptive field feature aggregation with low computational complexity. As shown in Figure 3, OSAG mainly consists of three stages: local, meso and global aggregations. Specifically, a channel attention [19] enhanced inverted bottleneck [18] is introduced to fulfill the local pattern process with limited overhead. Based on the proposed OSA paradigm, we derive two instances (i.e., Meso-OSA and Global-OSA) responsible for the interaction and aggregation of meso and global information. Note that the proposed omni self-attention paradigm can be used for different purposes. Meso-OSA performs attention within a group of non-overlap patches, which restricts Meso-OSA to only focus on meso-scale pattern understanding. Global-OSA samples data point sparsely across the entire feature within an atrous manner, endowing Global-OSA with the ability to achieve global interactions at a compelling cost.

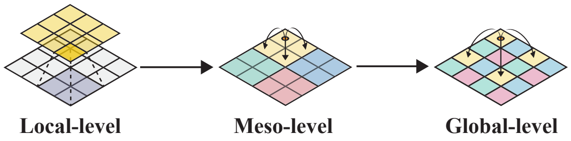


Figure 4. Illustration of omni-scale aggregation scheme. Our proposed Omni-SR contains three types of feature aggregation at local level, meso and global level, respectively.

The only difference between Meso-OSA and Global-OSA is the window partition strategy, as shown in Figure 4. In order to achieve meso-scale interaction, Meso-OSA split the input feature X into non-overlapping blocks with size $P \times P$. Note that after window partition, the block dimensions are gathered onto the spatial dimension (i.e., -2 axis): $(H, W, C) \rightarrow (\frac{H}{P} \times P, \frac{W}{P} \times P, C) \rightarrow (\frac{HW}{P^2}, P^2, C)$. While the Global-OSA divides the input feature into a uniform $G \times G$ grid, with each lattice having an adaptive size of $\frac{H}{G} \times \frac{W}{G}$. Similar to Meso-OSA, the grid dimension is also gathered on the spatial axis (*i.e.*-2 axis): $(H, W, C) \rightarrow (G \times \frac{H}{G}, G \times \frac{W}{G}, C) \rightarrow (G^2, \frac{HW}{G^2}, C) \rightarrow (\frac{HW}{G^2}, G^2, C)$.

3.4. Network Architecture

Overall Structure. Based on the Omni Self-Attention paradigm and the Omni-Scale Aggregation Group, we further develop a lightweight Omni-SR framework to achieve high-performance image super-resolution. As shown in Figure 3, Omni-SR consists of three parts, i.e., shallow feature extraction, deep feature extraction, and image reconstruction. Specifically, given the LR input $I_{LR} \in \mathbb{R}^{H \times W \times C_{in}}$, we first use a 3×3 convolution H_{SF} to extract shallow feature $X_0 \in \mathbb{R}^{H \times W \times C}$ as

$$X_0 = H_{SF}(I_{LR}), \tag{6}$$

where C_{in} and C denote the channel number of the input and shallow feature. The convolution layer provides a simple way to convert the input from image space into high-dimensional feature space. Then we use K stacked omni-scale aggregation groups (OSAG) and one 3×3 convolution layer H_{CONV} in a cascade manner to extract deep featureF_{DF}. Such a process can be expressed as

$$\begin{aligned} X_i &= H_{OSAG_i}(X_{i-1}), \quad i = 1, 2, \ldots, K, \\ X_{DF} &= H_{CONV}(X_K), \end{aligned} \tag{7}$$

where H_{OSAG_i} represents the i-th OSAG, $X_1, X_2, \ldots, X_K$ denote intermediate features. Following [31], we also use a convolutional layer at the end of feature extraction to get better feature aggregation. Finally we reconstruct the HR image I_{HR} by aggregating shallow and deep features as

$$I_{HR} = H_{Rec}(X_0 + X_{DF}), \tag{8}$$

where $H_{Rec}(\cdot)$ denotes the reconstruction module. In detail, PixelShuffle [46] is used to up-sample the fused feature.

Omni-Scale Aggregation Group (OSAG). As shown in Figure 3, each OSAG contains a local convolution block (LCB), a meso-OSA block, a global-OSA block, and an ESA block [28, 34]. The whole process can be formulated as

$$X_{res} = H_{Global\text{-}OSAB_i}(H_{Meso\text{-}OSAB_i}(H_{LCB_i}(X_{i-1}))), \tag{9}$$

$$X_i = H_{ESA_i}(H_{Conv_i}(X_{res} + X_{i-1})), \tag{10}$$

where X_{i-1} and X_i represents the input and output feature of i-th OSAG. After the mapping of convolution layers, we insert a Meso-OSAB for window-based self-attention and a Global-OSAB to enlarge the receptive field for better information aggregation. At the end of OSAG, we reserve the convolutional layer and ESA block following [28, 66].

In specific, LCB is implemented as a stack of pointwise and depthwise convolutions with a CA module [24] between them to adaptively re-weight channel-wise features. This block aims to aggregate local contextual information as well as to increase the trainability of the network [56]. Two types of OSA blocks (i.e., Meso-OSA block and Global-OSA block) are then followed to obtain interactions from different regions. Based on different window partition strategies, Meso-OSA block seeks inner-block interaction, and Global-OSA blocks aim for global mixing. OSA blocks follow typical Transformers designs with Feed-forward network (FFN) and LayerNorm [2], and the only difference is that the origin self-attention operation is replaced with our proposed OSA operator. For FFN, we adopt the GDFN proposed by Restormer [59]. Combining these individuals seamlessly, the designed OSAG enables information propagation between any pair of tokens in the feature map. We use the ESA module proposed in [28, 34] to further refine the fused feature.

Optimization Objective. Following prior works [31, 32, 53, 67], we train the model by minimizing a standard L_1 loss between model prediction $\hat{I}_{HR}$ and HR label I_{HR} as follows:

$$\mathcal{L} = \|I_{HR} - \hat{I}_{HR}\|_1. \tag{11}$$

4. Experiments

4.1. Experimental Setup

Datasets and Metrics. Following previous work [31, 32, 38, 49, 66], DIV2K [49] and Flickr2K [49] are used as training datasets. For a fair comparison, we employ two training protocols, i.e., training with DIV2K only and training with DF2K (DIV2K + Flickr2K). Note that the model trained with DF2K is marked with small †. For testing, we adopt five standard benchmark datasets: Set5 [4], Set14 [60], B100 [39], Urban100 [20] and Manga109 [40]. PSNR and SSIM [54] are adopted to evaluate the SR performance on the Y channel of the transformed YCbCr space.

Implementation Details. During training, we augment the data with random horizontal flips and 90/270-degree rotations. LR images are generated by bicubic downsampling [63] from HR images. OSAG number is set to 5, and channel number of the whole network is set to 64. The attention head number and window size are set to 4 and 8 for both Meso-OSAB and Global-OSAB. We use Adam [27] optimizer to train the model with a batch size of 64 for 800K iterations. The initial learning rate is set to 10^{-4} and halved for every 200k iterations. In each training batch, we randomly crop LR patches of size 64×64 as input. Our method is implemented with PyTorch [45], and all experiments are conducted on one NVIDIA V100 GPU. Note that no other data augmentation (e.g., Mixup [61], RGB channel shuffle) or training skills (e.g., pre-training [30], cosine learning schedule [36]) are employed. It should be pointed out that we maintain the consistency of model parameters in the ablation study by adjusting the channels of 1×1 convolution.

4.2. Comparison with the SOTA SR methods

To evaluate the effectiveness of Omni-SR, we compare our model with several advanced lightweight SR methods under a scale factor of 2/3/4. In particular, former works, VDSR [25], CARN [1], IMDN [22], EDSR [32], RFDN [33], MemNet [48], MAFFSRN [42], LatticeNet [38], RLFN [28], ESRT [37] and SwinIR [31] are introduced for comparison.

Quantitative results. In Table 1, the quantitative comparisons of different lightweight methods are presented on five benchmark datasets. With a similar model size, the performance of our Omni-SR surpasses existing methods with a notable margin on all benchmarks. In particular, compared to other transformer architectures with comparable parameters like SwinIR [31] and ESRT [37], the proposed Omni-SR obtains the best performance. The results

Table 1. Quantitative comparison (PSNR/SSIM) for **lightweight image SR** with state-of-the-art methods on benchmark datasets. The best and second-best results are marked in red and blue colors, respectively. "†" indicates that model is trained on DF2K.

Method	Years	Scale	Params (K)	Set5 PSNR / SSIM	Set14 PSNR / SSIM	BSD100 PSNR / SSIM	Urban100 PSNR / SSIM	Manga109 PSNR / SSIM
VDSR [25]	CVPR16	×2	666	36.66 / 0.9542	33.05 / 0.9127	31.90 / 0.8960	30.76 / 0.9140	37.22 / 0.9750
MemNet [48]	ICCV17		678	37.78 / 0.9597	33.28 / 0.9142	32.08 / 0.8978	31.31 / 0.9195	37.72 / 0.9740
SRMDNF [64]	CVPR18		1511	37.79 / 0.960	33.32 / 0.915	32.05 / 0.8985	31.33 / 0.9204	38.07 / 0.9761
CARN [1]	ECCV18		1,592	37.76 / 0.9590	33.52 / 0.9166	32.09 / 0.8978	31.92 / 0.9256	38.36 / 0.9765
IMDN [22]	MM19		694	38.00 / 0.9605	33.63 / 0.9177	32.19 / 0.8996	32.17 / 0.9283	38.88 / 0.9774
RFDN-L [33]	ECCV20		626	38.08 / 0.9606	33.67 / 0.9190	32.18 / 0.8996	32.24 / 0.9290	38.95 / 0.9773
MAFFSRN [42]	ECCV20		402	37.97 / 0.9603	33.49 / 0.9170	32.14 / 0.8994	31.96 / 0.9268	- / -
LatticeNet [38]	ECCV20		756	38.15 / 0.9610	33.78 / 0.9193	32.25 / 0.9005	32.43 / 0.9302	- / -
RLFN [28]	CVPRW22		527	38.07 / 0.9607	33.72 / 0.9187	32.22 / 0.9000	32.33 / 0.9299	- / -
SwinIR [31]	ICCVW21		878	38.14 / 0.9611	33.86 / 0.9206	32.31 / 0.9012	32.76 / 0.9340	39.12 / 0.9783
Omni-SR	Ours		772	38.22 / 0.9613	33.98 / 0.9210	32.36 / 0.9020	33.05 / 0.9363	39.28 / 0.9784
Omni-SR†	Ours		772	**38.29** / **0.9617**	**34.27** / **0.9238**	**32.41** / **0.9026**	**33.30** / **0.9386**	**39.53** / **0.9792**
VDSR [25]	CVPR16	×3	666	33.66 / 0.9213	29.77 / 0.8314	28.82 / 0.7976	27.14 / 0.8279	32.01 / 0.9340
MemNet [48]	ICCV17		678	34.09 / 0.9248	30.00 / 0.8350	28.96 / 0.8001	27.56 / 0.8376	32.51 / 0.9369
EDSR [32]	CVPRW17		1,555	34.37 / 0.9270	30.28 / 0.8417	29.09 / 0.8052	28.15 / 0.8527	33.45 / 0.9439
SRMDNF [64]	CVPR18		1,528	34.12 / 0.9254	30.04 / 0.8382	28.97 / 0.8025	27.57 / 0.8398	33.00 / 0.9403
CARN [1]	ECCV18		1,592	34.29 / 0.9255	30.29 / 0.8407	29.06 / 0.8034	28.06 / 0.8493	33.50 / 0.9440
IMDN [22]	MM19		703	34.36 / 0.9270	30.32 / 0.8417	29.09 / 0.8046	28.17 / 0.8519	33.61 / 0.9445
RFDN-L [33]	ECCV20		633	34.47 / 0.9280	30.35 / 0.8421	29.11 / 0.8053	28.32 / 0.8547	33.78 / 0.9458
MAFFSRN [42]	ECCV20		807	34.45 / 0.9277	30.40 / 0.8432	29.13 / 0.8061	28.26 / 0.8552	- / -
LatticeNet [38]	ECCV20		765	34.53 / 0.9281	30.39 / 0.8424	29.15 / 0.8059	28.33 / 0.8538	- / -
ESRT [37]	CVPRW22		770	34.42 / 0.9268	30.43 / 0.8433	29.15 / 0.8063	28.46 / 0.8574	33.95 / 0.9455
SwinIR [31]	ICCVW21		886	34.62 / 0.9289	30.54 / 0.8463	29.20 / 0.8082	28.66 / 0.8624	33.98 / 0.9478
Omni-SR	Ours		780	34.70 / 0.9294	30.57 / 0.8469	29.28 / 0.8094	28.84 / 0.8656	34.22 / 0.9487
Omni-SR†	Ours		780	**34.77** / **0.9304**	**30.70** / **0.8489**	**29.33** / **0.8111**	**29.12** / **0.8712**	**34.64** / **0.9507**
VDSR [25]	CVPR16	×4	666	31.35 / 0.8838	28.01 / 0.7674	27.29 / 0.7251	25.18 / 0.7524	28.83 / 0.8870
MemNet [48]	ICCV17		678	31.74 / 0.8893	28.26 / 0.7723	27.40 / 0.7281	25.50 / 0.7630	29.42 / 0.8942
EDSR [32]	CVPRW17		1,518	32.09 / 0.8938	28.58 / 0.7813	27.57 / 0.7357	26.04 / 0.7849	30.35 / 0.9067
SRMDNF [64]	CVPR18		1,552	31.96 / 0.8925	28.35 / 0.7787	27.49 / 0.7337	25.68 / 0.7731	30.09 / 0.9024
CARN [1]	ECCV18		1,592	32.13 / 0.8937	28.60 / 0.7806	27.58 / 0.7349	26.07 / 0.7837	30.47 / 0.9084
IMDN [22]	MM19		715	32.21 / 0.8948	28.58 / 0.7811	27.56 / 0.7353	26.04 / 0.7838	30.45 / 0.9075
RFDN-L [33]	ECCV20		643	32.28 / 0.8957	28.61 / 0.7818	27.58 / 0.7363	26.20 / 0.7883	30.61 / 0.9096
MAFFSRN [42]	ECCV20		830	32.20 / 0.8953	26.62 / 0.7822	27.59 / 0.7370	26.16 / 0.7887	- / -
LatticeNet [38]	ECCV20		777	32.30 / 0.8962	28.68 / 0.7830	27.62 / 0.7367	26.25 / 0.7873	- / -
RLFN [28]	CVPRW22		543	32.24 / 0.8952	28.62 / 0.7813	27.60 / 0.7364	26.17 / 0.7877	- / -
ESRT [37]	CVPRW22		751	32.19 / 0.8947	28.69 / 0.7833	27.69 / 0.7379	26.39 / 0.7962	30.75 / 0.9100
SwinIR [31]	ICCVW21		897	32.44 / 0.8976	28.77 / 0.7858	27.69 / 0.7406	26.47 / 0.7980	30.92 / 0.9151
Omni-SR	Ours		792	32.49 / 0.8988	28.78 / 0.7859	27.71 / 0.7415	26.64 / 0.8018	31.02 / 0.9151
Omni-SR†	Ours		792	**32.57** / **0.8993**	**28.95** / **0.7898**	**27.81** / **0.7439**	**26.95** / **0.8105**	**31.50** / **0.9192**

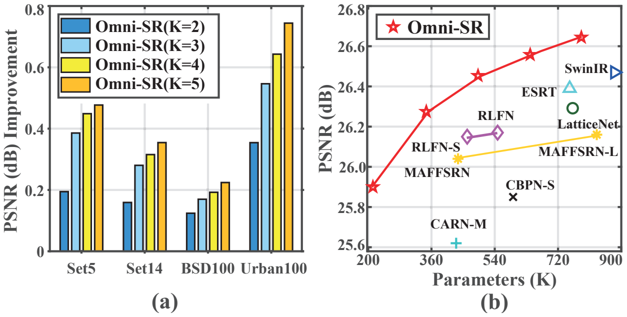


Figure 5. (a) PSNR improvement of Omni-SR variants with different OSAG number (K) over the smallest Omni-SR model (K=1) for 4 × SR. (b) The number of model parameters *vs.* PSNR of different lightweight methods on Urban100 dataset for 4 × SR.

showcase that the omni-axis (i.e., spatial + channel) interaction introduced by OSA can effectively boost the model's contextual aggregation capabilities, which promises superior SR performance. Coupling with large training dataset DF2K, the performance can be further improved, especially on Urban100. We suppose that such a phenomenon can be ascribed to the images in Urban100 have many similar patches, and the long-term relationship introduced by OSAG can bring great benefits for detail restoration. More importantly, with similar parameters, our model reduces 28% of computational complexity (Omni-SR: 36G FLOPs *vs.* SwinIR: 50G FLOPs @1280×720), showing its effectiveness and efficiency.

Visual comparison. In Figure 6, we also provide a visual comparison of different lightweight SR methods at ×4 scale. We can observe that the HR images constructed by Omni-SR contain more fine-grained details, while other methods generate blurred edges or artifacts in complicated areas. For instance, in the $1st$ row, our model is able to restore the detailed texture of the wall pleasantly, which all other methods fail to restore. Visual results also validate the effectiveness of the proposed OSA paradigm, which can perform omni-axis pixel-wise interaction modeling, thus obtaining a more powerful reconstruction capability.

Trade-off between Model Size and Performance. In experiments, we set the number of OSAG as 5 to make the model size around 800K for a fair comparison with other methods. We also explore model performance with smaller parameter sizes by reducing OSAG number K. As shown in Figure 5(a), compared to the smallest model variant with

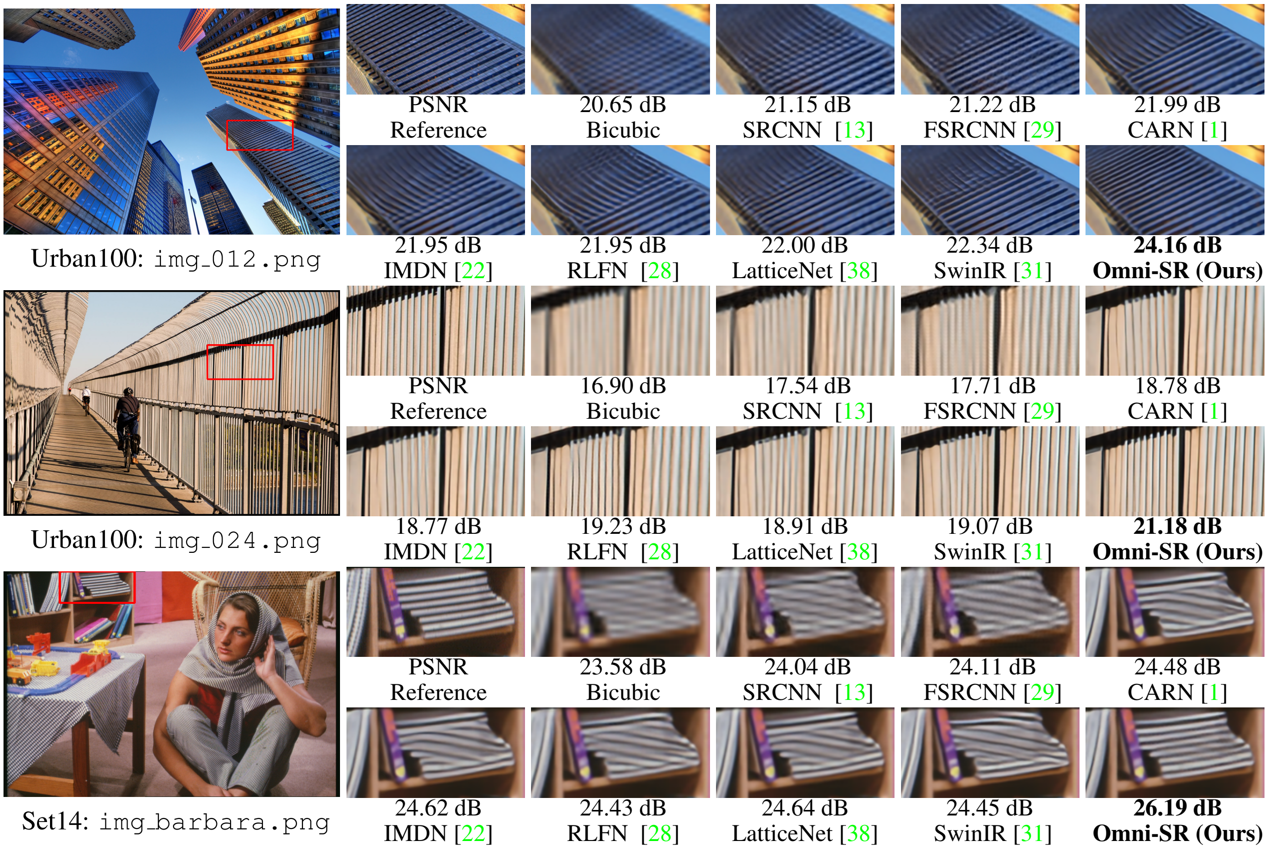


Figure 6. Visual comparison for ×4 SR methods. The patches for comparison are marked with red boxes. (Best viewed by zooming.)

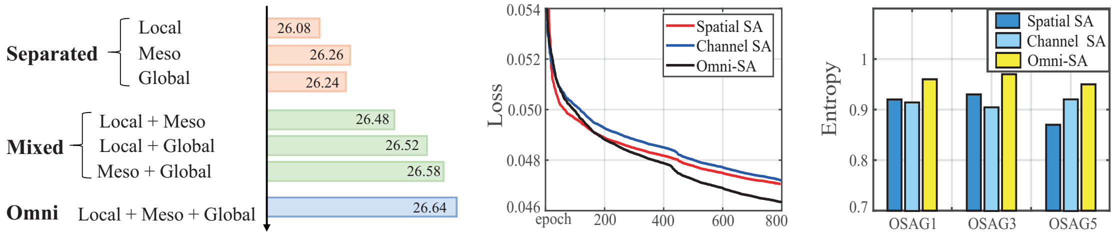
K=1, increasing the number of OSAG leads to stable performance improvements. In Figure 5(b), we present PSNR *vs.* parameters of different methods. It can be found that Omni-SR achieves the best results under various settings, showing its effectiveness and scalability.

4.3. Analysis of Omni Self-Attention

In this section, we illustrate the optimization features of OSA and further uncover its underlying mechanism. Self-attention is a low-bias operator, which makes its optimization difficult and requires more training epochs. For this, we introduce the additional channel-wise interaction to alleviate it. In Figure 7(a), we show the loss curves of different self-attention paradigms on the DIV2K training set, including spatial self-attention, channel self-attention, and the proposed omni self-attention. We can see that our OSA presents a obviously superior convergence speed. More importantly, the performance at the final epoch is also significantly ahead of them. The above phenomenon clearly shows that our OSA has superior good optimization characteristics. Further, we delve into why channel-wise interactions lead to these improvements. We calculate the normalized entropies [52] of the hidden layer features of the network composed of the above three computational primitives. We illustrate the entropy results in Figure 7(c). As shown in the figure, in all outgoing layers, our OSA-encoded features show higher entropy, indicating that our OSA encodes richer information. More information may come from various scales, and this information can help the operators to reconstruct the exact details faster. We speculate that this is the potential reason why our OSA shows better optimization properties. In addition, following previous works [8, 16], we also resort to LAM analysis. DI [16] metric can measure the furthest interaction distance of the model. From Figure 8 we can observe that Omni-SR generally has the highest max diffusion index than other methods, showing that our OSA paradigm can effectively capture long-range interactions.

4.4. Ablation Study

Effect of Omni Self-Attention. The core idea of our framework is to extend the vanilla self-attention with a channel-wise relationship to build omni-axis pixel-wise interaction. Based on Omni-SR framework, we design several variant models, and their SR results are reported in Table 2. We first simply remove the channel-wise component to form a spatial-only variant (Omni-SR$_{\text{sp}}$), and its performance is degraded by 0.13dB compared to the full model. Such a significant degradation justifies the importance of channel interactions. Note that Omni-SR$_{\text{sp}}$ still



(a) Ablation studies on OSAG (b) Training loss on DIV2K (c) Feature correlation

Figure 7. (a) Ablation studies on different components in OSAG. (b) Training loss of different attention paradigms. (c) Feature correlation analysis of different attention paradigms. All the results are reported on DIV2K dataset for 4× SR.

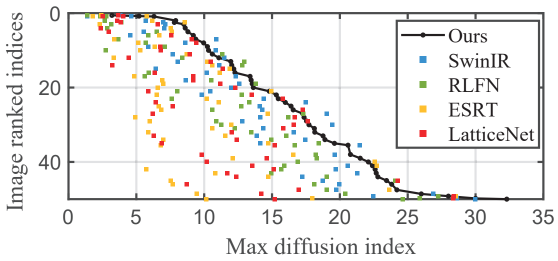


Figure 8. The distribution of DI values for different methods on Urban100. Top-left points with show a narrow area of interest, and right-bottom points show a large area of interest.

outperforms SwinIR by 0.04dB@Urban100 ×4, which benefits from the global interaction introduced by grid window partition. Similarly, we remove the spatial self-attention component to derive a channel self-attention variant, Omni-SR$_{\mathrm{ca}}$, and such a modification also leads to undesirable performance degradation. Besides, we use the most widely adopted channel and spatial attention configurations (i.e., SE [19] and CBAM [55]) to act as alternative operators for channel and spatial aggregation. Both substitutions (Omni-SR$_{\mathrm{SE}}$, Omni-SR$_{\mathrm{CBAM}}$) hurt PNSR performance compared to the full model. The above results show that the specific interaction paradigm (e.g., scalar-based, covariance-based) is equally important, and our channel interaction based on the covariance matrix shows great advantages.

Table 2. Ablation studies of omni self-attention on Urban100. Omni-SR$_{(*)}$ denotes different modifications. 'SA' and 'S-SA' denote spatial gate and spatial self-attention. 'CA' and 'C-SA' denote channel gate and channel self-attention. We maintain the consistency of model parameters by adjusting the channels of $1\times1 Conv$.

Model	SA	S-SA	CA	C-SA	FLOPs	×2	×3	×4
Omni-SR$_{\mathrm{sp}}$	✗	✓	✗	✗	33G	32.88	28.72	26.51
Omni-SR$_{\mathrm{SE}}$	✗	✓	✓	✗	34G	32.83	28.71	26.50
Omni-SR$_{\mathrm{CBAM}}$	✓	✗	✓	✓	34G	32.92	28.76	26.53
Omni-SR$_{\mathrm{ca}}$	✗	✗	✗	✓	33G	32.65	28.60	26.45
Omni-SR$_{\mathrm{full}}$	✗	✓	✗	✓	36G	33.05	28.84	26.64

Effect of Omni-Scale Aggregation Group. In Omni-SR, we propose a local-meso-global interaction scheme (i.e., OSAG) to pursue progressive feature aggregation. To investigate its effectiveness, we design three different kinds of interaction schemes based on Omni-SR framework: **Separated** scheme, **Mixed** scheme, and our fully designed **Omni** scheme (i.e., our proposed OSAG), and the ablation study results are shown in Figure 7(a). In the figure, we employ different words (e.g., "Local", "Meso+Global") to represent specific schemes, e.g., "Local" denotes using Local-Conv block to replace Meso-OSA and Global-OSA; "Local+Global" represents replace original cascaded Meso-OSA and Global-OSA with cascaded Local-Conv and Global-OSA. We can observe that single interaction schemes (e.g., "Local") perform the worst. Interestingly, the "Global" scheme is inferior to the "Meso" one due to its poor optimization properties of global self-attention [3, 35, 50]. Once two interaction operators are combined, the performance improves steadily. Among them, "Meso+ Global" setting achieves the second-best performance. Furthermore, combining all three interaction schemes together, we obtain the best performing scheme, i.e., "Omni". From the above experiments, we can infer that obvious performance gains can be obtained by introducing various-scale interactions, which also illustrates the feasibility and effectiveness of our proposed OSAG.

5. Conclusion

In this work, we propose Omni-SR, a lightweight framework for image SR. We propose the Omni Self-attention paradigm for simultaneous spatial and channel interactions, mining all the potential correlations across omni-axis. Furthermore, we propose an omni-scale aggregation scheme to effectively enlarge the receptive fields with low computational complexity, which encodes contextual relations in a progressively hierarchical manner. Extensive experiments on public benchmark datasets and comprehensive analytical studies validate its prominent SR performance.

6. Acknowledgement

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018. 2, 5, 6, 7

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. 5

[3] Irwan Bello, Barret Zoph, Quoc Le, Ashish Vaswani, and Jonathon Shlens. Attention augmented convolutional networks. In *ICCV*, 2019. 8

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

[5] 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*, 2021. 1, 2, 3

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016. 1

[7] Xuanhong Chen, Hang Wang, and Bingbing Ni. X-volution: On the unification of convolution and self-attention. *arXiv preprint arXiv:2106.02253*, 2021. 2

[8] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *CoRR*, abs/2205.04437, 2022. 2, 4, 7

[9] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobileformer: Bridging mobilenet and transformer. In *CVPR*, 2022. 2

[10] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Simple baselines for image restoration. In *ECCV*, 2022. 2, 3

[11] Xiangxiang Chu, Bo Zhang, Hailong Ma, Ruijun Xu, and Qingyuan Li. Fast, accurate and lightweight super-resolution with neural architecture search. In *ICPR*, 2020. 2

[12] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, 2022. 1

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015. 2, 7

[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 *ICLR*, 2021. 1, 2

[15] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong. Image super-resolution using knowledge distillation. In *ACCV*, 2018. 2

[16] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *CVPR*, 2021. 7

[17] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 2

[18] 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:1704.04861*, 2017. 4

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 3, 4, 8

[20] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 5

[21] Tao Huang, Lang Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Lightvit: Towards light-weight convolution-free vision transformers. *CoRR*, abs/2207.05557, 2022. 2

[22] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACMMM*, 2019. 2, 5, 6, 7

[23] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, 2018. 2

[24] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv:1602.07360*, 2016. 5

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2, 5, 6

[26] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 2

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

[28] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *CVPR Workshops*, 2022. 2, 5, 6, 7

[29] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *TPAMI*, 41(11):2599–2613, 2018. 7

[30] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 2, 5

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 1, 2, 4, 5, 6, 7

[32] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 5, 6

[33] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *ECCV*, 2020. 5, 6

[34] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 5

[35] 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*, 2021. 2, 4, 8

[36] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 5

[37] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *CVPR Workshops*, 2022. 2, 3, 4, 5, 6

[38] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *ECCV*, 2020. 2, 5, 6, 7

[39] 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 *ICCV*. IEEE, 2001. 5

[40] 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

[41] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. *CoRR*, abs/2110.02178, 2021. 2

[42] Abdul Muqeet, Jiwon Hwang, Subin Yang, JungHeum Kang, Yongwoo Kim, and Sung-Ho Bae. Multi-attention based ultra lightweight image super-resolution. In *ECCV*, 2020. 2, 5, 6

[43] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martínez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *ECCV*, 2022. 2

[44] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv:1807.06514*, 2018. 3, 4

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS Workshop*, 2017. 5

[46] 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 *CVPR*, 2016. 5

[47] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017. 2

[48] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, 2017. 2, 5, 6

[49] 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 *CVPR workshops*, 2017. 5

[50] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *CVPR*, 2021. 2, 8

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 1, 2, 3

[52] Pei Wang, Yijun Li, and Nuno Vasconcelos. Rethinking and improving the robustness of image style transfer. In *CVPR*, 2021. 7

[53] Xiaohang Wang, Xuanhong Chen, Bingbing Ni, Zhengyan Tong, and Hang Wang. Learning continuous depth representation via geometric spatial aggregator. In *AAAI*, 2023. 5

[54] 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

[55] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 3, 4, 8

[56] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick. Early convolutions help transformers see better. In *NeurIPS*, 2021. 5

[57] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan L. Yuille. Lite vision transformer with enhanced self-attention. In *CVPR*, 2022. 2

[58] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks, 2022. 3

[59] 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, 3, 5

[60] 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

[61] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 5

[62] Haokui Zhang, Wenze Hu, and Xiaoyu Wang. Parc-net: Position aware circular convolution with merits from convnets and transformer. In *ECCV*, 2022. 2

[63] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, 2017. 5

[64] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 6

[65] Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *CVPR*, 2021. 2

[66] 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*, 2018. 1, 2, 5

[67] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 5