Attention Guidance Distillation Network for Efficient Image Super-Resolution

Hongyuan Wang*, Ziyan Wei*, Qingting Tang*, Shuli Cheng†, Liejun Wang†, and Yongming Li
School of Computer Science and Technology, Xinjiang University
Ürümqi, China
{why5200, 107552204088, tqt}@stu.xju.edu.cn, {cslxju, wljxju, lym}@xju.edu.cn

Abstract

Over the past decade, neural network-based super-resolution techniques have been developed on a large scale with impressive achievements. Many novel solutions have been proposed, among which lightweight solutions based on convolutional neural networks have been designed for applications in edge devices. To better realize this application, we propose a more lightweight attention guidance distillation network (AGDN). We design the attention guidance distillation block (AGDB) with more efficient space, channel and self-attention as the infrastructure of AGDN. Specifically, multi-level variance-aware spatial attention (MVSA) is designed to better capture structurally information-rich regions with new multi-scale convolution and local variance alignment. Reallocated contrast-aware channel attention (RCCA) is designed to enhance the processing of common information in all channels while redistributing weights across channels. Sparse global self-attention (SGSA) is introduced for selecting the most useful similarity values for image reconstruction. Extensive experiments demonstrate that AGDN strikes a better balance between performance and complexity compared to other models, achieving SOTA performance on several benchmark tests. In addition, our AGDN-S ranks first in the FLOPs track and second in the Parameters track of the NTIRE 2024 Efficient SR Challenge. The code is available at https://github.com/daydreamer2024/AGDN.

1. Introduction

Single-image super-resolution (SISR) is a fundamental task in the field of computer vision that aims to generate high-resolution (HR) images from low-resolution (LR) images. As a critical component of low-level vision tasks, SR techniques find widespread use in various real-world applications, such as remote-sensing imaging, medical imaging, and security surveillance. In recent years, propelled by advancements in deep learning technology, SR techniques have made remarkable strides, leading to the emergence of numerous innovative network architectures. Beginning from the early convolutional neural networks [7] and progressing through residual networks [13], to transformer models [4] and diffusion models [26], the field of SR has witnessed a flourishing development. Nevertheless, with the continuous evolution of technology, SR networks have grown increasingly complex, and their network structures have expanded in size. While these sophisticated SR networks enhance the quality of image reconstruction, their deployment on edge devices with limited computational resources presents challenges due to the escalating model capacity and intensive computational demands.

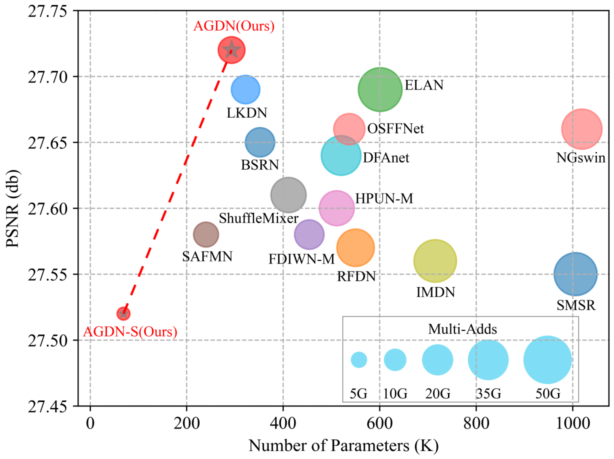


Figure 1. Comparison of model performance and complexity on B100 dataset for upsampling factor ×4.

To alleviate this situation, efficient SR networks based on convolutional neural networks have gradually gained attention. From early CARN [2] to IMDN [12], RFDN [19], BSRN [17], and then MDRN [22], efficient SR networks have improved greatly in terms of model structure and attention mechanisms, while the number of network parameters has been significantly reduced. In this research process, the

[*] Equal contributions to this work. [†] Corresponding author.

information distillation mechanism was proposed and considered a highly effective method, widely used as one of the efficient architectures to significantly boost the overall efficiency of the model. Therefore, based on the foundation of the information distillation mechanism, we summarize the factors that constrain the further development of efficient SR networks, among which precise attention guidance is a key issue. Building upon this insight, we adopt more effective attention modules to construct superior efficient SR networks.

In this paper, we propose a more lightweight attention guidance distillation network (AGDN), which leverages multiple attention mechanisms, including spatial attention, channel attention, and self-attention. These robust attention mechanisms play a pivotal role in steering the network towards more efficient selection of crucial information, resulting in superior reconstruction results. As depicted in Fig. 1, our AGDN not only achieves a reduction in model complexity but also surpasses other efficient methods in terms of performance metrics.

Specifically, we propose the attention guidance distillation block (AGDB) that utilizes multiple attention mechanisms as the foundational block of AGDN. We employ multi-level variance-aware spatial attention (MVSA) and reallocated contrast-aware channel attention (RCCA) as alternatives to enhanced spatial attention (ESA) [20] and contrast-aware channel attention (CCA) [12], respectively. Additionally, we introduce sparse global self-attention (SGSA) [14] to further enhance features. MVSA is adept at capturing structurally information-rich regions, RCCA can redistribute channel weights and handle common information, and SGSA selects the most relevant similarity values for image reconstruction. AGDN demonstrates superior performance while maintaining a leaner and more efficient model architecture. We developed AGDN-S based on AGDN for participation in the NTIRE 2024 Efficient SR Challenge [25], achieving first place in the FLOPs track and second place in the Parameters track.

Overall, our primary contributions can be succinctly summarized as follows:

1. We propose a more lightweight efficient super-resolution network named AGDN, which reconstructs higher-quality images with fewer parameters and multi-adds compared to other state-of-the-art methods on commonly available datasets.

2. Through the review and summary of existing distillation block designs, we have developed the AGDB, which improves and effectively utilizes attention modules to enhance the model's capabilities within limited computational resources.

3. We introduce SGSA as a novel addition, enhancing the utilization of critical global features for more effective image reconstruction.

2. Related work

2.1. Single image SR

The development of SISR has been remarkable. Dong et al. [7] first performed the nonlinear mapping of features by three-layer convolution, which achieved good reconstruction results compared to traditional mathematical methods. To enhance feature representation and reuse capabilities, VDSR [13] employed a very deep network structure to learn complex and abstract feature representations, while also accelerating training convergence. Huang et al. [9], by introducing dense connections, addressed gradient vanishing and feature sparsity issues in traditional networks, thereby enhancing model representation. For further advancements in image reconstruction, novel network structures have been introduced, such as Zhou et al. [37] introduced graph neural networks into SR, and Latticenet [21] utilized lattice filtering with varying fast fourier transforms.

In recent years, SR networks based on transformer or diffusion models have also started to drive further performance improvements. Network such as [4] achieved SOTA performance with the transformer model, while network like [26] excelled in visual generation using the diffusion model.

2.2. Efficient image SR

Researchers have begun to develop networks with lower computational requirements while aiming to maintain performance levels. CARN [2] replaced ordinary convolution in the residual block with group convolution, enhancing the network's expressive power while keeping the model lightweight. IDN [11] performed a splitting process on the feature map to better utilize layered features. IMDN [12] introduced a residual structure that extracted useful features in stages, with the remaining features processed downward through convolution layers, resulting in notable reconstruction results. RFDN [19] modified channel segmentation operation with two parallel 3×3 convolutions and used residual blocks with constant connectivity as feature extractors.

Several new networks have provided researchers with a new level of inspiration. BSRN [17] proposed by Kong et al. used BSConv instead of ordinary convolution. Choi et al. [5] introduced N-Gram context into the image SR task to expand the observed region and recover degraded pixels.

2.3. Attention model in SR

Attentional mechanisms can help models better focus on important parts. The channel attention (CA) mechanism was initially introduced to image SR by Zhang et al. [36], marking a advancement in the field. Hui et al. [12] replaced global average merging in the CA mechanism with the sum of standard deviation and average, enhancing the model's ability to capture long-term dependencies in sequences and improve modeling accuracy. However, CA mechanism still

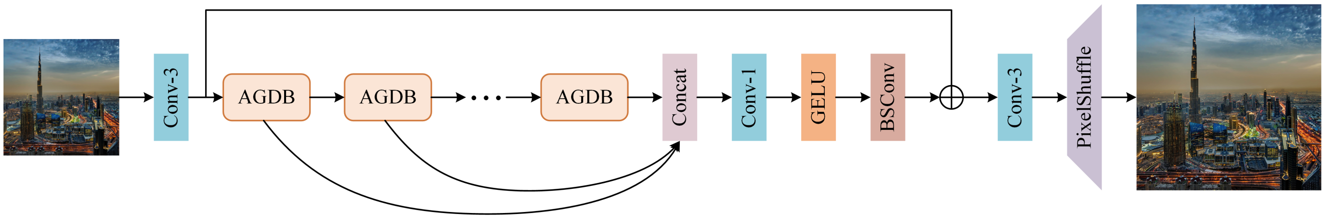


Figure 2. The overall architecture of attention guidance distillation network (AGDN).

extracts features globally, resulting in a lack of attention to other perspective features. Liu et al. [20] designed efficient spatial attention for SR aimed at generating more representative features. ELAN [35] utilized group multi-scale self-attention to better construct long-distance dependencies in images. Chen et al. [4] proposed a novel HAT that combines CA and self-attention, achieving good performance.

The excellent model structure and innovative attention module in AGDN enable a better balance between performance and model complexity. A manageable number of parameters and computational requirements are maintained while effectively increasing modeling capabilities. Making AGDN the ideal solution for high-efficiency and high-quality image SR tasks.

3. Method

3.1. Network Architecture

We propose a more lightweight attention guidance distillation network (AGDN) for efficient image SR, which is influenced by existing studies such as IMDN, RFDN, BSRN, and MDRN, and further improved based on these studies. As illustrated in Fig. 2, our network integrates four closely interconnected components: shallow feature extraction, deep feature extraction, feature fusion, and reconstruction. The effectiveness of this overall architecture has been extensively validated in previous studies.

Assuming the input and output of AGDN are represented as $I_{\text{LR}} \in \mathbb{R}^{3\times H\times W}$ and $I_{\text{SR}} \in \mathbb{R}^{3\times sH\times sW}$, where $H \times W$ denoted as the spatial size and s denoted as the upsampling factor. Images are fed into the network at the shallow feature extraction phase as

$$F_0 = \mathcal{H}_{\text{fe}}\left(I_{\text{LR}}\right), \tag{1}$$

where $\mathcal{H}_{\text{fe}}(\cdot)$ represents the shallow feature extraction phase consisting of a 3×3 convolutional layer. After extracting the obtained shallow features, deep feature extraction is carried out using AGDBs, so that the features are gradually refined and perfected. This process can be expressed as

$$F_k = \mathcal{H}_{\text{k}}\left(F_{k-1}\right), k = 1, \ldots, n, \tag{2}$$

where $\mathcal{H}_{\text{k}}(\cdot)$ denotes the k-th AGDB. F_{k-1} and F_k represent the input feature and output feature of the k-th AGDB, respectively. Following the deep feature extraction phase, feature fusion is performed to merge the multilevel features. Specifically, a 1×1 convolutional layer and the GELU activation function are utilized for feature mapping. Additionally, the features are refined using a BSConv. This entire process can be represented as

$$F_d = \mathcal{H}_{\text{fusion}}\left(\text{Concat}\left(F_1, \ldots, F_k\right)\right), \tag{3}$$

where $\mathcal{H}_{\text{fusion}}(\cdot)$ represents the feature fusion phase and F_d is the fused feature. This is followed by a reconstruction phase via a long skip connection, given by the formula as

$$I_{\text{SR}} = \mathcal{H}_{\text{re}}^{s}\left(F_d + F_0\right). \tag{4}$$

where $\mathcal{H}_{\text{re}}(\cdot)$ denotes the reconstruction phase, which consists of a 3×3 convolution and a pixelshuffle operation.

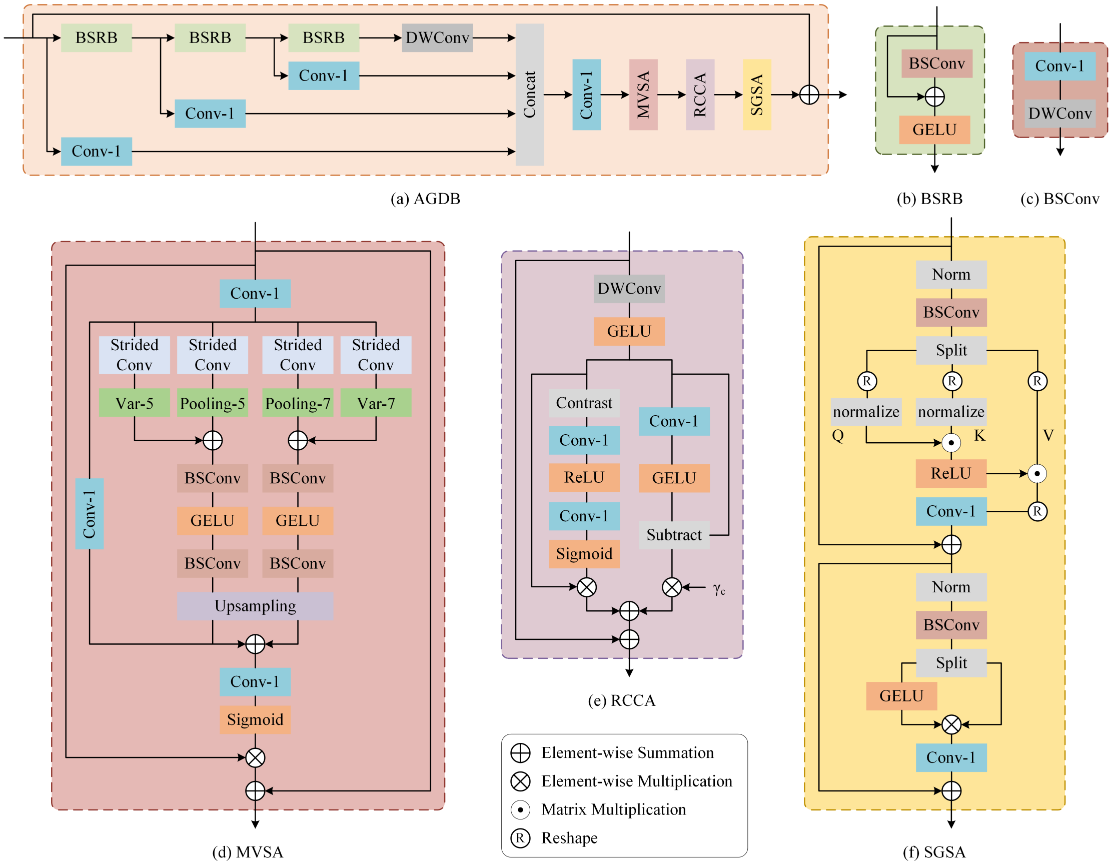
We train the above network by using the L_1 loss function. Given a training set with N pairs of LR images and HR counterparts, denoted by $\left\{I_{\text{LR}}^i, I_{\text{HR}}^i\right\}_{i=1}^N$, the loss can be obtained as

$$L_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left\|I_{\text{SR}}^i - I_{\text{HR}}^i\right\|_1 \tag{5}$$

where θ denotes the parameter sets of our proposed AGDN.

3.2. Rethinking the deep feature extraction phase

By compiling and analyzing previous studies, our study found that the deep feature extraction phase is still a key limiting factor for model performance, and this phase consists of a series of feature distillation blocks. By analyzing the structure in IMDN and LKDN, each block can be divided into two main parts: distillation in the pre-phase and enhancement in the post-phase. In the distillation part, it has been a developmental process from channel splitting operations to feature distillation connections to blueprint separable convolutions instead of traditional convolutions; in the enhancement part, it has evolved from channel attention to spatial attention to spatial channel attention fusion. Based on the above analysis, we believe that improving both distillation and enhancement can significantly boost network performance. Therefore, we center our work on improving the feature distillation blocks of other networks to further promote the model towards lightweight and efficiency.



(a) AGDB

(b) BSRB

(c) BSConv

(d) MVSA

(e) RCCA

(f) SGSA

Figure 3. The details of each component. (a) AGDB: Attention Guidance Distillation Block; (b) BSRB: Blueprint Shallow Residual Block; (c) BSConv: Blueprint Separable Convolution; (d) MVSA: Multi-level Variance-aware Spatial Attention; (e) RCCA: Reallocated Contrast-aware Channel Attention; (f) SGSA: Sparse Global Self-attention.

3.3. Attention Guidance Distillation Block

The information distillation mechanism is widely used in lightweight SR network design and has been proven effective. Therefore, we adopt the information distillation mechanism to propose a novel AGDB, as shown in Fig. 3(a).

The feature distillation phase involves a series of BSRBs and convolutional layers designed to iteratively enhance the initial input features F_{in}. The entire process can be represented as

$$\begin{aligned} F_{d_1}, F_{r_1} &= D_1\left(F_{in}\right), R_1\left(F_{in}\right), \\ F_{d_2}, F_{r_2} &= D_2\left(F_{r_1}\right), R_2\left(F_{r_1}\right), \\ F_{d_3}, F_{r_3} &= D_3\left(F_{r_2}\right), R_3\left(F_{r_2}\right), \\ F_{d_4} &= D_4\left(F_{r_3}\right), \end{aligned} \tag{6}$$

where D_i, R_i denote the ith distillation and ith refinement layer, respectively. F_{d_i}, F_{r_i} represents the ith distilled features and ith refined features, respectively. In the feature fusion phase, all the distilled features fused by a 1 × 1 convolutional layer as

$$F_{fusion} = \mathcal{H}_{\text{fusion}}\left(\text{Concat}\left(F_{d_1}, F_{d_2}, F_{d_3}, F_{d_4}\right)\right), \tag{7}$$

where $\mathcal{H}_{\text{fusion}}$ denotes the 1×1 convolutional layer, and F_{fusion} is the fused feature. For the feature enhancement phase, we use more efficient spatial attention, channeled attention and self-attention for cascading enhancements as

$$\begin{aligned} F_{mvsa} &= \mathcal{H}_{\text{mvsa}}\left(F_{fusion}\right), \\ F_{rcca} &= \mathcal{H}_{\text{rcca}}\left(F_{mvsa}\right), \\ F_{enhance} &= \mathcal{H}_{\text{sgsa}}\left(F_{rcca}\right), \end{aligned} \tag{8}$$

where $\mathcal{H}_{\text{mvsa}}$ denotes the multi-level variance-aware spatial attention, $\mathcal{H}_{\text{rcca}}$ denotes the reallocated contrast-aware channel attention, $\mathcal{H}_{\text{sgsa}}$ denotes the sparse global self-attention, and $F_{enhance}$ is the enhanced feature. Finally, a long skip connection is used to strengthen the residual learning ability of the model as

$$F_{out} = F_{enhance} + F_{in}. \tag{9}$$

We utilize the designed AGDB, which incorporates more efficient spatial, channel, and self-attention, as the foundational module of AGDN. This block enables us to achieve enhanced reconstruction results.

Table 1. The ablation analysis for MVSA, RCCA and SGSA in the AGDN on benchmark test datasets for ×4 SR.

Method	MVSA	RCCA	SGSA	Params	M-Adds	Set5	Set14	BSD100	Urban100	Manga109
						PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
1	×	×	×	127K	7.2G	32.00/0.8928	28.52/0.7801	27.53/0.7344	25.81/0.7770	30.24/0.9053
2	✓	×	×	145K	7.8G	32.20/0.8949	28.65/0.7824	27.58/0.7365	26.06/0.7850	30.53/0.9086
3	×	✓	×	132K	7.4G	32.09/0.8938	28.54/0.7802	27.53/0.7347	25.91/0.7799	30.35/0.9067
4	✓	✓	×	151K	7.9G	32.20/0.8948	28.64/0.7824	27.60/0.7367	26.10/0.7864	30.56/0.9097
5	✓	✓	✓	293K	16.1G	32.43/0.8980	28.80/0.7861	27.72/0.7412	26.50/0.7992	31.13/0.9148

Table 2. The ablation analysis for different components of attentional mechanisms on benchmark test datasets for ×4 SR.

Method	Params	M-Adds	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
ESA+CCA	286K	16.0G	32.42/0.8977	28.76/0.7850	27.71/0.7405	26.41/0.7970	31.04/0.9143
MVSA+RCCA	293K	16.1G	32.43/0.8980	28.80/0.7861	27.72/0.7412	26.50/0.7992	31.13/0.9148

Table 3. The ablation analysis for the TLC approach on benchmark test datasets for ×4 SR.

Method	Set5	Set14	BSD100	Urban100	Manga109
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
AGDN-woTLC	32.42/0.8977	28.77/0.7852	27.69/0.7399	26.33/0.7932	31.02/0.9131
AGDN	32.43/0.8980	28.80/0.7861	27.72/0.7412	26.50/0.7992	31.13/0.9148

Table 4. Performance results with varying number of AGDBs on benchmark test datasets for ×4 SR.

Number	Params	M-Adds	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
5	215K	11.9G	32.35/0.8969	28.75/0.7845	27.68/0.7399	26.36/0.7948	30.92/0.9132
6	254K	14.0G	32.42/0.8975	28.79/0.7856	27.71/0.7408	26.45/0.7974	31.01/0.9135
7	293K	16.1G	32.43/0.8980	28.80/0.7861	27.72/0.7412	26.50/0.7992	31.13/0.9148

Multi-level Variance-aware Spatial Attention. As shown in Fig. 3(d), we draw on research on multi-level dispersion spatial attention (MDSA) [22] and improve upon it. In MVSA, we consider the impact of multi-level branching and local variance on performance. Multi-level branches with small windows cannot cover a sufficient range of information, while using local variance in a single branch can lead to large differences in weights between branches. Therefore, we designed D5 and D7 branches that contain both local variance to better capture structurally information-rich regions while balancing performance and model complexity.

Specifically, on the one hand, we compute the local variance with the same kernel size and step size as the max-pooling layer while computing the max-pooling layer. We then perform element-wise summation on the outputs before continuing with the same operations as in ESA. On the other hand, we added a multi-level branch, which provides multi-level weight information. With this multi-level and local variance steering, we improve the reconstruction accuracy with a small increase in model complexity.

Reallocated Contrast-aware Channel Attention. As shown in Fig. 3(e), we design the reallocated contrast-aware channel attention (RCCA). In RCCA, we not only consider the reallocation of weights across channels by traditional channel attention but also enhance the treatment of common information across all channels. We added complementary branches with 1 × 1 convolution and GELU activation representations to reallocate complementary channel information, promoting the uniqueness of each channel.

Specifically, we first deepen each channel of the input individually by using the DWConv and GELU activation functions to enhance the uniqueness of each channel. Next, all channels of the input are augmented in two branches, redistributing the weights of each channel in the CCA branch, and redistributing the information of the complementary channels in the complementary branch. Finally, the results of the two branches are summed to ensure that each channel has a more unique importance.

Sparse Global Self-Attention. As shown in Fig. 3(f), we introduce SGSA to estimate the features in the channel direction and generate the scaled attention matrix. By introducing SGSA, the most important global features can be better utilized for image reconstruction. Here, we follow the original computational approach of SGSA, where global attention in image restoration usually has a gap between the training and testing phases. Therefore, we use the test-time localizer converter (TLC) [6] during the testing phase.

4. Experiments

4.1. Datasets and Metrics

We trained our model using a dataset consisting of 800 images from DIV2K [1] and 2650 images from Flickr2K [18]. For testing, we selected SR benchmark datasets such as Set5 [3], Set14 [34], B100 [23], Urban100 [10], and

Table 5. The quantitative results for super-resolution on five benchmark datasets are presented, highlighting the best result in red and the second-best result in blue. The Multi-Adds is calculated corresponding to a 1280 × 720 HR image.

Method	Scale	Venue	Params	M-Adds	Set5	Set14	BSD100	Urban100	Manga109
					PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
CARN [2]	×2	ECCV 18'	1592K	222.8G	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
IMDN [12]		ACM MM 19'	694K	158.8G	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
RFDN [19]		ECCVW 20'	534K	123.0G	38.05/0.9606	33.68/0.9184	32.16/0.8994	32.12/0.9278	38.88/0.9773
SMSR [30]		CVPR 21'	985K	131.6G	38.00/0.9601	33.64/0.9179	32.17/0.8990	32.19/0.9284	38.76/0.9771
ELAN [35]		ECCV 22'	582K	168.4G	38.17/0.9611	33.94/0.9207	32.30/0.9012	32.76/0.9340	39.11/0.9782
FDIWN-M [8]		AAAI 22'	433K	73.6G	38.03/0.9606	33.60/0.9179	32.17/0.8995	32.19/0.9284	-/-
ShuffleMixer [28]		NeurIPS 22'	394K	91.0G	38.01/0.9606	33.63/0.9180	32.17/0.8995	31.89/0.9257	38.83/0.9774
BSRN [17]		CVPRW 22'	332K	73.0G	38.10/0.9610	33.74/0.9193	32.24/0.9006	32.34/0.9303	39.14/0.9782
NGswin [5]		CVPR 23'	998K	140.4G	38.05/0.9610	33.79/0.9199	32.27/0.9008	32.53/0.9324	38.97/0.9777
SAFMN [29]		ICCV 23'	228K	52.0G	38.00/0.9605	33.54/0.9177	32.16/0.8995	31.84/0.9256	38.71/0.9771
HPUN-M [27]		AAAI 23'	492K	106.2G	38.04/0.9605	33.67/0.9190	32.22/0.9001	32.17/0.9290	38.89/0.9776
DFAnet [15]		ICASSP 23'	500K	137.8G	38.09/0.9609	33.80/0.9199	32.26/0.9009	32.55/0.9328	38.86/0.9778
LKDN [33]		CVPRW 23'	304K	69.1G	38.12/0.9611	33.90/0.9202	32.27/0.9010	32.53/0.9322	39.19/0.9784
OSFFNet [31]		AAAI 24'	516K	83.2G	38.11/0.9610	33.72/0.9190	32.29/0.9012	32.67/0.9331	39.09/0.9780
AGDN(Ours)			279K	61.4G	38.21/0.9613	34.02/0.9220	32.33/0.9017	32.76/0.9346	39.33/0.9784
CARN [2]	×3	ECCV 18'	1592K	118.8G	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.50/0.9440
IMDN [12]		ACM MM 19'	703K	71.5G	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
RFDN [19]		ECCVW 20'	541K	55.4G	34.41/0.9273	30.34/0.8420	29.09/0.8050	28.21/0.8525	33.67/0.9449
SMSR [30]		CVPR 21'	993K	67.8G	34.40/0.9270	30.33/0.8412	29.10/0.8050	28.25//0.8536	33.68/0.9445
ELAN [35]		ECCV 22'	590K	75.7G	34.61/0.9288	30.55/0.8463	29.21/0.8081	28.69/0.8624	34.00/0.9478
FDIWN-M [8]		AAAI 22'	446K	35.9G	34.46/0.9274	30.35/0.8423	29.10/0.8051	28.16/0.8528	-/-
ShuffleMixer [28]		NeurIPS 22'	415K	43.0G	34.40/0.9272	30.37/0.8423	29.12/0.8051	28.08/0.8498	33.69/0.9448
BSRN [17]		CVPRW 22'	340K	33.3G	34.46/0.9277	30.47/0.8449	29.18/0.8068	28.39/0.8567	34.05/0.9471
NGswin [5]		CVPR 23'	1007K	66.6G	34.52/0.9282	30.53/0.8456	29.19/0.8078	28.52/0.8603	33.89/0.9470
SAFMN [29]		ICCV 23'	233K	23.0G	34.34/0.9267	30.33/0.8418	29.08/0.8048	27.95/0.8474	33.52/0.9437
HPUN-M [27]		AAAI 23'	500K	48.1G	34.44/0.9271	30.37/0.8426	29.13/0.8056	28.18/0.8533	33.68/0.9450
DFAnet [15]		ICASSP 23'	508K	62.0G	34.52/0.9281	30.47/0.8449	29.17/0.8075	28.52/0.8604	33.72/0.9463
LKDN [33]		CVPRW 23'	311K	31.4G	34.54/0.9285	30.52/0.8455	29.21/0.8078	28.50/0.8601	34.08/0.9475
OSFFNet [31]		AAAI 24'	524K	37.8G	34.58/0.9287	30.48/0.8450	29.21/0.8080	28.49/0.8595	34.00/0.9472
AGDN(Ours)			285K	27.9G	34.63/0.9291	30.55/0.8460	29.26/0.8093	28.66/0.8633	34.30/0.9487
CARN [2]	×4	ECCV 18'	1592K	90.9G	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.47/0.9084
IMDN [12]		ACM MM 19'	715K	40.9G	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
RFDN [19]		ECCVW 20'	550K	31.6G	32.24/0.8952	28.61/0.7819	27.57/0.7360	26.11/0.7858	30.58/0.9089
SMSR [30]		CVPR 21'	1006K	41.6G	32.12/0.8932	28.55/0.7808	27.55/0.7351	26.11/0.7868	30.54/0.9085
ELAN [35]		ECCV 22'	601K	43.2G	32.43/0.8975	28.78/0.7858	27.69/0.7406	26.54/0.7982	30.92/0.9150
FDIWN-M [8]		AAAI 22'	454K	19.6G	32.17/0.8941	28.55/0.7806	27.58/0.7364	26.02/0.7844	-/-
ShuffleMixer [28]		NeurIPS 22'	411K	28.0G	32.21/0.8953	28.66/0.7827	27.61/0.7366	26.08/0.7835	30.65/0.9093
BSRN [17]		CVPRW 22'	352K	19.4G	32.35/0.8966	28.73/0.7847	27.65/0.7387	26.27/0.7908	30.84/0.9123
NGswin [5]		CVPR 23'	1019K	36.4G	32.33/0.8963	28.78/0.7859	27.66/0.7396	26.45/0.7963	30.80/0.9128
SAFMN [29]		ICCV 23'	240K	14.0G	32.18/0.8948	28.60/0.7813	27.58/0.7359	25.97/0.7809	30.43/0.9063
HPUN-M [27]		AAAI 23'	511K	27.7G	32.24/0.8950	28.66/0.7828	27.60/0.7371	26.12/0.7878	30.55/0.9089
DFAnet [15]		ICASSP 23'	520K	35.6G	32.29/0.8960	28.71/0.7835	27.64/0.7386	26.40/0.7955	30.71/0.9113
LKDN [33]		CVPRW 23'	322K	18.3G	32.39/0.8979	28.79/0.7859	27.69/0.7402	26.42/0.7965	30.97/0.9140
OSFFNet [31]		AAAI 24'	537K	22.0G	32.39/0.8976	28.75/0.7852	27.66/0.7393	26.36/0.7950	30.84/0.9125
AGDN-S(Ours)			69K	3.7G	32.03/0.8920	28.53/0.7793	27.52/0.7344	25.99/0.7821	30.29/0.9039
AGDN(Ours)			293K	16.1G	32.43/0.8980	28.80/0.7861	27.72/0.7412	26.50/0.7992	31.13/0.9148

Manga109 [24]. In line with standard practices, we assessed image SR performance using metrics like PSNR and SSIM [32]. Additionally, we conducted an analysis of the model complexity based on parameters and multi-adds.

4.2. Implementation Details

In line with previous approaches, we apply random rotation and horizontal flipping to augment the data. Our network is trained using the Adam optimizer. During training, we randomly sample 48 × 48 patches as input to the network. We set the total training iterations to 1000k with a batch size of 64. The initial learning rate is set to 2×10^{-3} and is halved at specific iterations: [100k, 500k, 800k, 900k, 950k]. All experiments were performed on one NVIDIA RTX 4090 GPU using the PyTorch framework.

The proposed AGDN-S has 4 AGDBs, in which the number of feature channels is set to 24. We start by pretraining the model on the DIV2K and Flickr2K datasets, we expand the network's input to 64 × 64 while keeping the rest consistent with AGDN. We then conduct fine-tuning on the DIV2K dataset and the first 10k images from LSDIR [16]. The input size is set to 96 × 96, with a batch size of 32. The fine-tuning process optimizes the model by minimizing the L2 loss function, starting with an initial learning rate of 5×10^{-4}, which is reduced by half at 50k iterations. The fine-tuning phase encompasses a total of 100k iterations.

4.3. Ablation Study

Efficacy of the Multi-level Variance-aware Spatial Attention. To demonstrate the effectiveness of MVSA, we performed ablation experiments on MVSA, and the results are shown in Table 1. After adding MVSA, the model complexity produces a small rise, but the reconstruction effect is greatly improved. It proves that MVSA is useful.

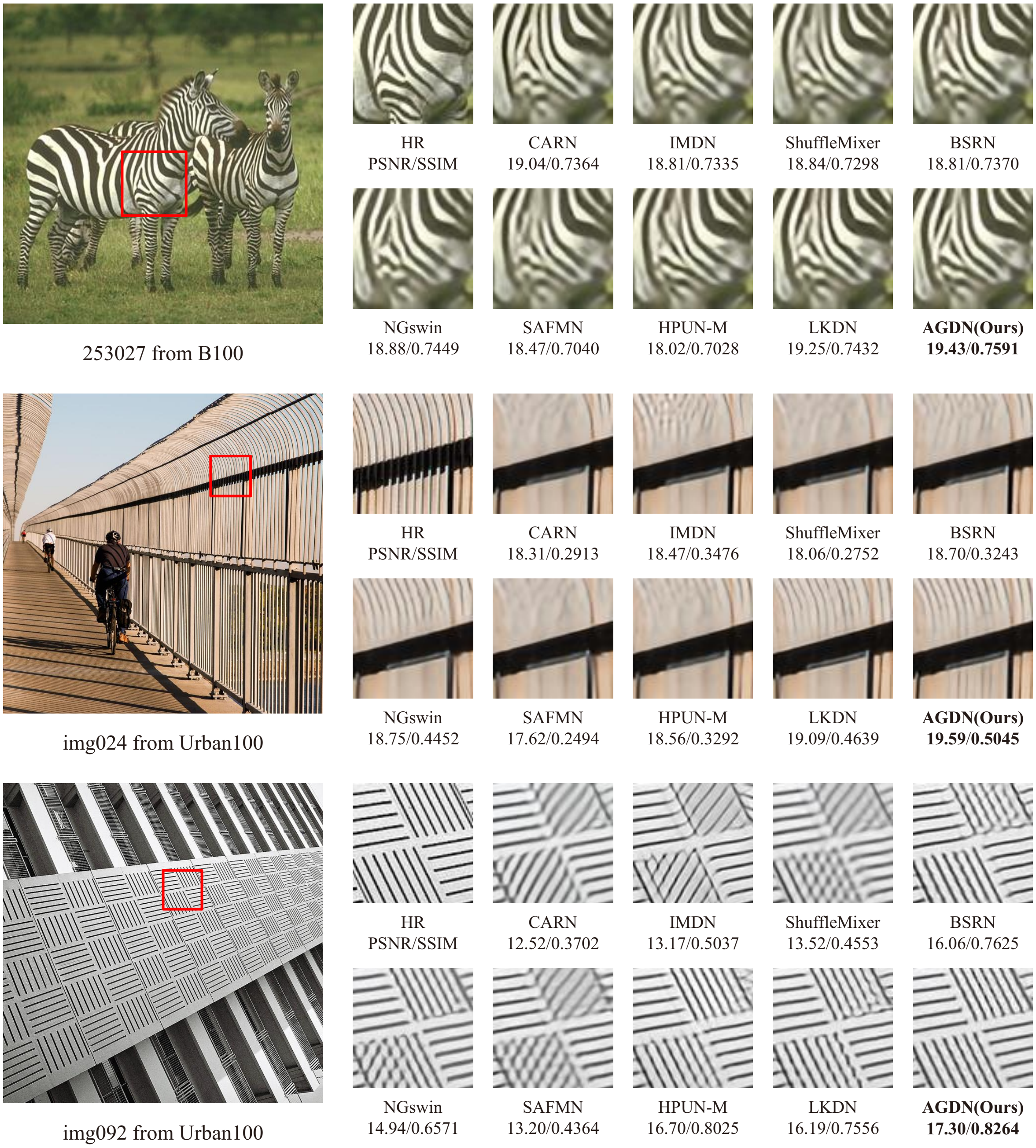


Figure 4. Result comparison of our method with other methods on the test datasets (×4).

Efficacy of the Reallocated Contrast-aware Channel Attention. As shown in Table 1, we performed ablation experiments on RCCA. Adding RCCA to the baseline improves the robustness of the network very well, and significant improvements are realized on several datasets. At the same time, we can still realize the enhancement of reconstruction performance with a small amount of model complexity increase by adding RCCA on top of MVSA. This proves that our RCCA is lightweight and practical.

Influence of the Attention Components. The model integrated with ESA and CCA serves as the benchmark for evaluating the effectiveness and performance improvements brought about by the proposed attention module. The comparative analysis is detailed in Table 2. We observed that AGDN, equipped with MVSA and RCCA, exhibits improved performance with only a 7K increase in parameters. Particularly notable improvements were observed on the Urban100 and Manga109 datasets.

Influence of the Sparse Global Self-Attention. To evaluate the importance of SGSA, we constructed the network structure without SGSA. As shown in Table 1, the introduction of SGSA resulted in a significant improvement in performance, despite a tremendous increase in network complexity. Even with this increase in complexity, the entire network remains at a lightweight level. We also experimented with the effect of the TLC approach on the final results. Experiments using the TLC approach in Table 3 showed better robustness, proving the practicality and effectiveness of maintaining the testing phase with the same patch size as the training phase in self-attention.

Influence of the number of AGDBs. In Table 4, we investigated the impact of varying the number of AGDBs. As the number of blocks increases, both the network's parameters and FLOPs increase, leading to a gradual improvement in reconstruction performance. Notably, increasing the number of blocks from 5 to 6 results in a significant performance boost, and further increasing the number from 6 to 7 enhances the network's robustness. Therefore, we have settled on using 7 AGDBs, which keeps the network lightweight while having good performance.

4.4. Comparison with the State-of-the-art Methods

We conducted a comparison between the proposed AGDN and several efficient SR methods, namely CARN [2], IMDN [12], RFDN [19], SMSR [30], ELAN [35], FDIWN-M [8], ShuffleMixer [28], BSRN [17], NGswin [5], SAFMN [29], HPUN-M [27], DFAnet [15], LKDN [33], and OSFFNet [31]. The results of this comparison are presented in Table 5. By comparing the results, we can find that our AGDN is able to achieve the SOTA effect on most of the publicly available common benchmark datasets with only a small number of parameters and multi-adds. Relative to contemporary other methods during that timeframe, our AGDN shows an intuitive improvement in performance on ×2, ×3, and ×4 SR tasks with better model robustness. These quantitative results demonstrate that our proposed AGDN is efficient and reliable.

Fig. 4 illustrates the visual comparisons of different SR methods on the B100 and Urban100 datasets for ×4 SR. AGDN makes the output clearer by enhancing texture and edge effects. Specifically, for images ”img024” and ”img092”, most comparison methods exhibit obvious artifacts and blurring effects. For example, CARN and IMDN show obvious misalignment on ”img092”, while our method retains more accurate lines. This intuitive visual demonstration further proves the competitiveness of our proposed AGDN in image SR.

4.5. AGDN-S for NTIRE 2024 Challenge

As shown in Table 6, our AGDN-S won first place in the FLOPs track and second place in the Parameters track in the NTIRE 2024 Efficient SR Challenge. Our AGDN-S achieved the performance required by the challenge with only 69K parameters and 4.39G FLOPs. Our method demonstrates significant improvements in terms of both parameters and FLOPs but still falls short in runtime efficiency. The issue of long runtime may be attributed to the slow computation of the attention mechanism and the additional matrix transformations introduced by the TLC approach. We plan to make further improvements in the future. For more details and results, please refer to [25].

Table 6. Results of NTIRE 2024 Efficient Super-Resolution Challenge

Team	Val PSNR	Test PSNR	Runtime[ms]	Params[M]	FLOPs[G]
XJU_100th Ann(Ours)	26.90	27.02	58.836	0.069	4.39
VPEG_C	26.90	27.03	16.032	0.084	4.97
ZHEstar	26.93	27.04	27.866	0.090	5.81
VPEG_E	26.90	27.01	18.476	0.093	5.89
PiXupt	26.91	27.00	48.562	0.060	9.84
MViC_SR	26.90	27.00	11.882	0.138	8.16

5. Conclusion

In this paper, we propose a more lightweight efficient super-resolution network, AGDN. We reevaluate the previous network structure, identifying the feature distillation block in the deep feature extraction phase as the key limiting factor of network performance. Therefore, we design AGDB with more efficient spatial, channel, and self-attention as the feature distillation block for AGDN. We enhance MDSA to obtain MVSA for better capturing structurally information-rich regions. RCCA is designed not only to redistribute weights across channels through traditional channel attention but also to improve the treatment of common information in all channels. SGSA is introduced for selecting the most useful similarity values to better utilize global features for image reconstruction. Extensive experiments demonstrate that our method achieves superior performance with fewer parameters and multi-adds compared to other efficient SR methods. In the future, we will focus on optimizing the network runtime to address network constraints. Additionally, we will explore the generality of the proposed method to apply it to other image restoration tasks, such as image denoising and enhancement.

Acknowledgments.

This work was supported by the Scientific and Technological Innovation 2030 Major Project under Grant 2022ZD0115800, the Basic Research Funds for Colleges and Universities in Xinjiang Uygur Autonomous Region under Grant XJEDU2023P008, the Key Laboratory Open Projects in Xinjiang Uygur Autonomous Region under Grant 2023D04028, and the Natural Science Foundation Program of Xinjiang Uygur Autonomous Region under Grant 2022D01C82.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 5

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision*, pages 252–268, 2018. 1, 2, 6, 8

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

[4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 1, 2, 3

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

[6] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *Proceedings of European Conference on Computer Vision*, pages 53–71. Springer, 2022. 5

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 1, 2

[8] Guangwei Gao, Wenjie Li, Juncheng Li, Fei Wu, Huimin Lu, and Yi Yu. Feature distillation interaction weighting network for lightweight image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 661–669, 2022. 6, 8

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 2

[10] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 5

[11] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 723–731, 2018. 2

[12] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the Acm International Conference on Multimedia*, pages 2024–2032, 2019. 1, 2, 6, 8

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 1, 2

[14] Xiang Li, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Dlgsanet: lightweight dynamic local and global self-attention networks for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12792–12801, 2023. 2

[15] Yanchun Li, Xinan He, Shujuan Tian, Zhetao Li, and Saiqin Long. Deep feature aggregation for lightweight single image super-resolution. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023. 6, 8

[16] 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 Workshops*, 2023. 6

[17] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 833–843, 2022. 1, 2, 6, 8

[18] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1132–1140, 2017. 5

[19] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 41–55, 2020. 1, 2, 6, 8

[20] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2359–2368, 2020. 2, 3

[21] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *Proceedings of the European Conference on Computer Vision*, pages 272–289. Springer, 2020. 2

[22] Yanyu Mao, Nihao Zhang, Qian Wang, Bendu Bai, Wanying Bai, Haonan Fang, Peng Liu, Mingyue Li, and Shengbo Yan. Multi-level dispersion residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1660–1669, 2023. 1, 5

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

[24] 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*, 76:21811–21838, 2017. 6

[25] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, et al. The ninth ntire 2024 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 2, 8

[26] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. 1, 2

[27] Bin Sun, Yulun Zhang, Songyao Jiang, and Yun Fu. Hybrid pixel-unshuffled network for lightweight image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2375–2383, 2023. 6, 8

[28] Long Sun, Jinshan Pan, and Jinhui Tang. Shufflemixer: An efficient convnet for image super-resolution. *Advances in Neural Information Processing Systems*, 35:17314–17326, 2022. 6, 8

[29] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13190–13199, 2023. 6, 8

[30] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4917–4926, 2021. 6, 8

[31] Yang Wang and Tao Zhang. Osffnet: Omni-stage feature fusion network for lightweight image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5660–5668, 2024. 6, 8

[32] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[33] Chengxing Xie, Xiaoming Zhang, Linze Li, Haiteng Meng, Tianlin Zhang, Tianrui Li, and Xiaole Zhao. Large kernel distillation network for efficient single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1283–1292, 2023. 6, 8

[34] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010. 5

[35] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 649–667, 2022. 3, 6, 8

[36] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, pages 286–301, 2018. 2

[37] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *Advances in neural information processing systems*, 33:3499–3509, 2020. 2