DVMSR: Distilled Vision Mamba for Efficient Super-Resolution

Xiaoyan Lei[1], Wenlong zhang[2]*, Weifeng Cao[1]*

[1] Zhengzhou University of Light Industry, [2] The HongKong Polytechnic University

xyan_lei@163.com, wenlong.zhang@connect.polyu.hk, weifeng_cao@163.com

Abstract

Efficient Image Super-Resolution (SR) aims to accelerate SR network inference by minimizing computational complexity and network parameters while preserving performance. Existing state-of-the-art Efficient Image Super-Resolution methods are based on convolutional neural networks. Few attempts have been made with Mamba to harness its long-range modeling capability and efficient computational complexity, which have shown impressive performance on high-level vision tasks. In this paper, we propose DVMSR, a novel lightweight Image SR network that incorporates Vision Mamba and a distillation strategy. The network of DVMSR consists of three modules: feature extraction convolution, multiple stacked Residual State Space Blocks (RSSBs), and a reconstruction module. Specifically, the deep feature extraction module is composed of several residual state space blocks (RSSB), each of which has several Vision Mamba Moudles(ViMM) together with a residual connection. To achieve efficiency improvement while maintaining comparable performance, we employ a distillation strategy to the vision Mamba network for superior performance. Specifically, we leverage the rich representation knowledge of teacher network as additional supervision for the output of lightweight student networks. Extensive experiments have demonstrated that our proposed DVMSR can outperform state-of-the-art efficient SR methods in terms of model parameters while maintaining the performance of both PSNR and SSIM. The source code is available at https://github.com/nathan66666/DVMSR.git

1. Introduction

Single image super-resolution (SR) is a key challenge in computer vision and image processing, aiming to reconstruct a high-resolution image from a low-resolution input. Effective super-resolution aims to improve the efficiency of the SR model while maintaining reconstruction perfor-
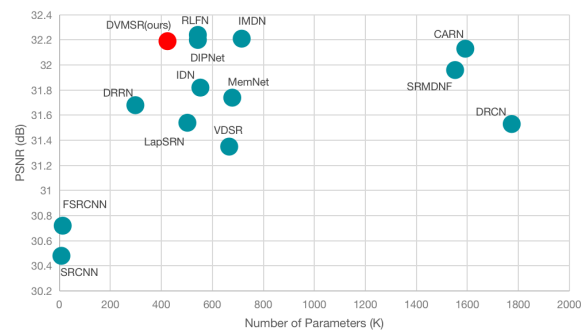

Figure 1. PSNR results v.s the total number of parameters of different methods for image SR on Set5.

mance. Since the introduction of deep learning into super-resolution tasks [9], many CNN-based methods have been proposed [8, 10, 11, 25, 26, 29, 34] to improve the performance. A series of approaches [10, 17, 19, 25, 26, 28, 30, 36, 56] have been proposed for building efficient models for image SR. The majority of these efficient models focus on five factors: runtime, parameters, FLOPS, activations, and depths. To further promote the development of efficient SR, ICCV holds the first competition in the AIM 2019 challenge [60]. The information multi-distillation network(IMDN) [19] proposes cascaded information multi-distillation blocks to improve the feature extraction module, which won first place in this competition. After that, The winning solution of the AIM 2020 challenge [61], residual feature distillation network(RFDN) [36], further improves the IMDN by residual learning in the main block. In the efficient SR track of NTIRE 2022 [24] challenge, the winning solution, residual local feature network(RLFN) [28], removes the hierarchical distillation connection of residual feature distillation block(RFDB) [36] to reduce the inference time. In the efficient SR track of NTIRE 2022 [56] challenge, the winning solution utilizes a multi-stage lightweight training strategy that combines distillation and

---

*Corresponding author

pruning to reduce both time consumption and model size.

The Transformer model, initially successful in natural language processing [49], has attracted interest from the computer vision community. Its effectiveness in high-level visual tasks (e.g., image classification [12, 39, 51]) has demonstrated the potential in super-resolution [6, 35]. Recently, Mamba [13] has demonstrated superior performance over Transformers across various sizes on large-scale real data and exhibits linear scalability with sequence length. Despite pioneering works adopting Mamba for vision tasks [13, 44, 55], it is still in its initial stages of exploring its potential (e.g., long-range modeling capability and efficiency) in low-level vision.

Different from the CNN-based and transformer-based methods, our goal is to explore the long-range modeling capability and efficiency of mamba-based methods for efficient SR. In this paper, we employ vision mamba as the basic architecture to enhance the model's long-range modeling capability and efficiency. Our DVMSR consists of several stacked Residual State Space Blocks (RSSB), each containing several Vision Mamba Modules (ViMM). The ViMM includes a unidirectional SSM, a residual connection, and SiLU activation function. These elements work together to accelerate model convergence and enhance model accuracy and efficiency. As shown in Figure 2, our method can achieve a larger perception range compared with other methods. Furthermore, we utilize a distillation strategy to enhance the model's efficiency. We introduce a Mamba network with a larger number of parameters as the teacher network to extract knowledge for the learning of the student network. Extensive experiments and ablation studies have shown the effectiveness of our proposed method.

Our contributions can be summarized as follows:

1. By leveraging the long-range modeling capability of Vision Mamba, we propose a lightweight model with unidirectional state space models (SSM) for efficient super-resolution.

2. We propose a special feature distillation strategy to enhance the efficiency ability of vision mamba for efficient super-resolution.

3. Extensive experiments have shown that our proposed method outperforms existing state-of-the-art (SOTA) methods in terms of parameters while maintaining comparable PSNR and SSIM performance.

2. Related Work

2.1. Lightweight Super Resolution

SRCNN [9] marks the inaugural application of deep learning algorithms in the Single Image Super-Resolution (SISR) [5, 6]. A series of works have been explored to apply the SR method in real scenarios, such as GAN-based SR [31, 52, 63, 64], degradation model [53, 62, 65], multi-task learning [67] and systematic evaluation [66]. In real-world SR model deployments, the computing power of the deployed devices is often limited, such as edge devices, etc. In this case, the efficiency of the SR network becomes an important aspect. Efficient Image Super-Resolution aims to reduce the computational effort and parameters of the SR network while achieving faster inference times and maintaining high performance. FSRCNN [10] reduces unnecessary computational costs by utilizing the deconvolution layer as the upsampling layer. VDSR [26] is introduced to further improve super-resolution (SR) performance. DRCN [25] achieves parameter reduction through deep recursive convolutional networks. LapSRN [30] employs a Laplacian pyramid super-resolution block for HR image reconstruction. DRRN [46] employs recursive and residual network architectures, surpassing DRCN in both performance and parameter reduction. MemNet [47] introduces a memory block to explicitly model long-term dependencies in CNN-based SR models. IDN [17] explicitly divides the preceding extracted features into two parts. IMDN [19] introduces a lightweight Information Multi-Distillation Network by constructing cascaded Information Multi-Distillation Blocks. RFDN [36] proposes the residual feature distillation network. RLFN [28] improves its speed by eliminating hierarchical distillation connections. DIP-Net [56] introduces the Reparameterization Residual Feature Block, which explores the potential of complex structures during optimization while maintaining computational efficiency. Besides, they achieve first place in the NTIRE 2023 Efficient Super-Resolution Challenge [33].

2.2. State space models in Vision

Recent researches have led to a surge of interest in the state space model (SSM), which has its origins in the classic Kalman filter model [23]. The linear scalability of State Space Models (SSMs) in handling long-range dependencies, exemplified by the Mamba architecture [13], contrasts with Transformers. While Mamba outperforms Transformers in natural language tasks, recent research endeavors extend its applicability to vision tasks. Specifically, Mamba models are designed to capture long-range temporal dependencies in video data, enhancing video classification performance [21, 22, 43, 50]. Additionally, other works explore Mamba's applicability in vision tasks, including image classification [38, 69], biomedical image segmentation [40], remote sensing image classification [4], and Multimodal Learning [44]. The research conducted by [14] emphasizes Mamba's utility as a straightforward and efficient baseline for image restoration in low-level vision tasks. Our work extends this by proposing a novel network architecture that combines Mamba with distillation, achieving a trade-off between super-resolution quality and computational efficiency.
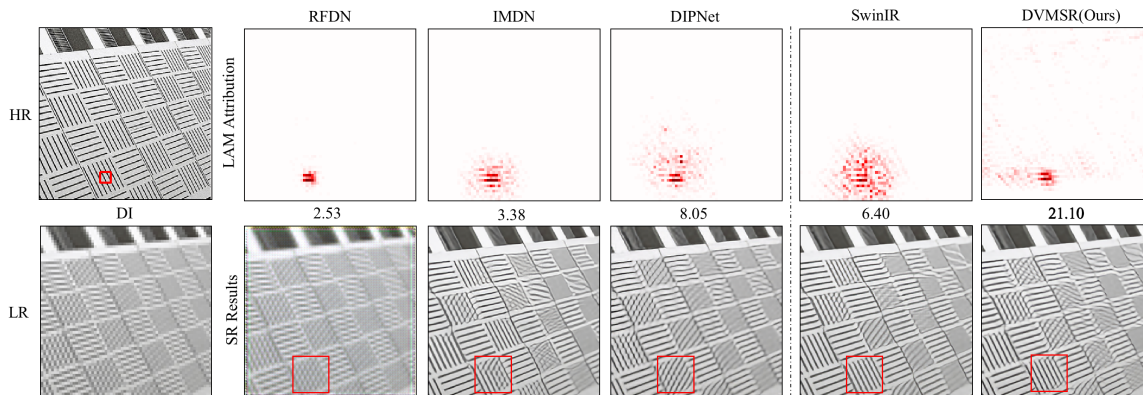
Figure 2. The LAM results are provided for various networks including both CNN-based and transformer-based methods. LAM attribution indicates the significance of each pixel in the input LR image during the reconstruction process of the patch highlighted by a box. The Diffusion Index (DI) denotes the extent of pixel involvement. A higher DI indicates a broader range of utilized pixels.

2.3. Feature Distillation

Knowledge distillation stands out as a straightforward yet powerful technique for enhancing the performance of smaller models, a necessity driven by the limited computing power of deployed devices. This method involves training a smaller network (student) under the guidance of a larger network (teacher), enabling effective knowledge transfer. Unlike other compression methods, knowledge distillation can reduce network size regardless of structural differences between the teacher and student networks. The seminal work by [15] introduced the knowledge distillation (KD) method, utilizing the softmax output of the teacher network. Notably, this method can be applied across various network architectures due to matching output dimensions. Over time, intermediate layer distillation methods have emerged, leveraging insights from the teacher network's convolutional or penultimate layers, preserving crucial feature-map localities [1, 15, 27, 57]. Moreover, there exists a wealth of research integrating distillation techniques into super-resolution tasks [18, 20, 37, 54, 68]. In this paper, we focus on adopting the output feature map of a pre-trained model as the distillation target. Through extensive experimentation, we demonstrate the effectiveness of our approach in enhancing model performance.

3. Methodology

3.1. Motivation

Efficient Super Resolution (SR) is designed to transform low-quality images into high-quality counterparts, leveraging a small parameter set and minimal computational power. ESR predominantly relies on CNNs for local feature extraction, but their limited long-range modeling hinders performance. Transformers, while proficient in global context, introduce computational complexities. Mamba

excels in high-level vision tasks, supported by prior research [4, 38, 40, 44, 55, 69]. Motivated by Mamba's long-range modeling capabilities, we investigate its performance in super-resolution (SR) tasks, comparing it to CNN-based ESR methods [19, 36, 56] and transformer-based method [35]. To elucidate Mamba's operational mechanisms, we employ a specialized diagnostic tool called LAM [7], designed specifically for SR tasks. Utilizing LAM enabled us to pinpoint the input pixels that contribute most significantly to the selected region. As depicted in Figure 2, the red-marked points denote informative pixels crucial for the reconstruction process. Notably, DVMSR exhibited a notably higher DI (Diffusion index) indication compared to other models, indicating its superior ability to leverage a broader range of pixel information and affirming its exceptional long-range modeling capability. The proposed DVMSR yields improved image details during the reconstruction process, thereby substantiating its efficacy for super-resolution tasks.

3.2. Preliminaries

State space models (SSMs), such as the Mamba deep learning model, hold potential for long sequence modeling. Inspired by continuous systems, SSMs map a 1-D function or sequence $x(t) \in \mathbb{R} \longmapsto y(t) \in \mathbb{R}$ via a hidden state $h(t) \in \mathbb{R}^N$. The formulation is as follows:

$$
\begin{aligned}
h'(t) &= Ah(t) + Bx(t), \\
y(t) &= Ch(t).
\end{aligned}
\tag{1}
$$

where N is the state size, $A \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^{N \times 1}$, $C \in \mathbb{R}^{1 \times N}$.

Mamba is the discrete versions of the continuous system, and it achieves this by utilizing Δ to convert continuous parameters A and B into their discrete counterparts, $\bar{A}$ and
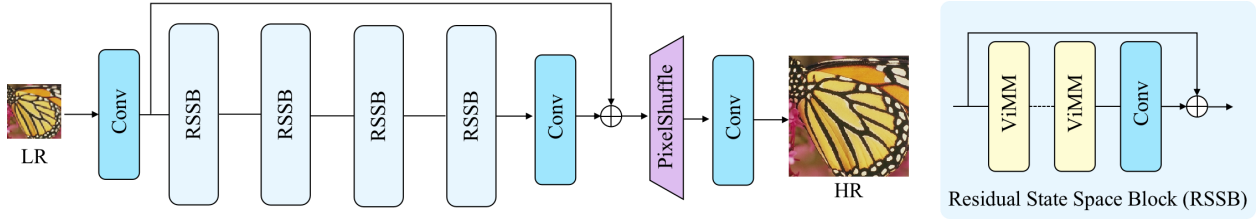
Figure 3. The overall network architecture of our DVMSR.

$\bar{B}$. The commonly used method for transformation is zero-order hold (ZOH), which is defined as follows:

$$\bar{A} = exp(\Delta A),$$
$$\bar{B} = (\Delta A)^{-1}(exp(\Delta A) - I) \cdot \Delta B. \tag{2}$$

After the discretization of $\bar{A}$, $\bar{B}$, the discretized version of Eq. 5 using a step size Δ can be rewritten as:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t,$$
$$y_t = Ch_t. \tag{3}$$

3.3. Overall network architecture

The overall network architecture of our proposed DVMSR is depicted in Figure 3. Our DVMSR mainly consists of three main modules: feature extraction convolution, multiple stacked Residual State Space Blocks (RSSBs), and a reconstruction module. Specifically, for a given low-resolution (LR) input $I_{LR} \in \mathbb{R}^{H \times W \times C_{in}}$, we exploit one convolution layer to extract the first feature $F_0 \in \mathbb{R}^{H \times W \times C}$, where C_{in} and C denote the channel number of the input and the intermediate feature. Then, a series of Residual State Space Block (RSSB) and one 3×3 convolution layer $HConv(\cdot)$ are utilized to perform the deep feature extraction. After that, we add a global residual connection to fuse shallow features F_0 and deep features $F_D \in \mathbb{R}^{H \times W \times C}$, and then reconstruct the high-resolution result via a reconstruction module. As depicted in Figure 3, each RSSB contains two Vision Mamba Module (ViMM) and a 3×3 convolution layer with a residual connection. For the reconstruction module, the pixel-shuffle method is adopted to up-sample the fused feature.

3.3.1 Mamba network

The design of mamba network is shown in Figure 4, which is Vision Mamba Module (ViMM) using unidirectional sequence modeling. The input token sequence $X \in \mathbb{R}^{H \times W \times C}$ is first normalized by the normalization layer. Next, we linearly project the normalized sequence, expanded the features channel to λC. We proceed by processing the projection layer through 1-D convolution, resulting
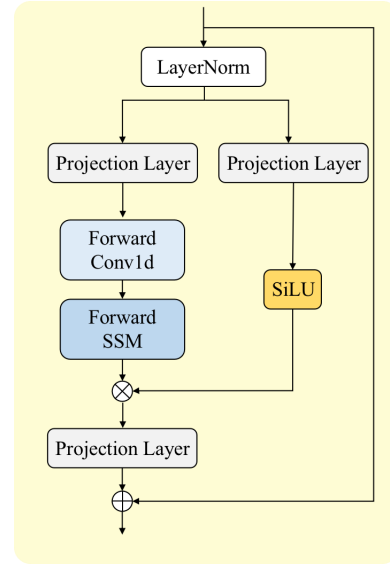

Figure 4. The structure of Vision Mamba Module(ViMM).

in the computation of X_1 via the SSM. The X_1 gated by the projection layer and a residual connection to get the output token sequence $X_{out} \in \mathbb{R}^{H \times W \times C}$, as follows:

$$X_1 = SSM(Conv1d(Linear(LN(X)))),$$
$$X_2 = SiLU(Linear(LN(X))),$$
$$X_{out} = Linear(X_1 \odot X_2) + X. \tag{4}$$

Where LN is the layer normalization and $\odot$ denotes the Hadamard product.

3.3.2 Distillation strategy

Our method introduces a deep feature distillation strategy (Fig. 5). During the distillation stage, the teacher network accumulates rich representation knowledge, maintaining a fixed state. By minimizing the $\mathcal{L}_1$ loss, we ensure alignment between student network features and those of the teacher. This formal process facilitates effective knowledge transfer from the teacher to the student network:
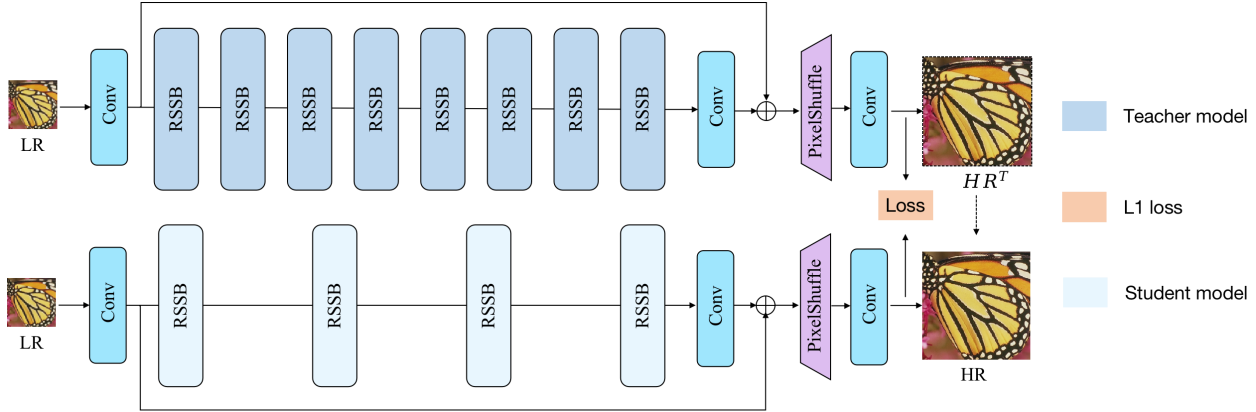
Figure 5. The deep feature distillation pipeline of our method.

$$\mathcal{L}_{out} = \lambda_{dis}\mathcal{L}_{dis} + \lambda_1\mathcal{L}_1,$$
$$\mathcal{L}_{dis} = \|\mathcal{T}(I_{LR}) - \mathcal{S}(I_{LR})\|_1, \quad (5)$$
$$\mathcal{L}_1 = \|I_{HR} - \mathcal{S}(I_{LR})\|_1,$$

where λ_{dis} and λ_1 represents the coefficient of the $\mathcal{L}_{dis}$ loss function and the coefficient of the $\mathcal{L}_1$ loss function, respectively. They are set 1. $\mathcal{T}$ represents the function of our teacher network and $\mathcal{S}$ denotes the function of our proposed network. I_{LR} and I_{HR} are the input LR images and the corresponding ground-truth HR images, respectively. More information of $\mathcal{L}_{dis}$ can be seen from Fig.6.

4. Experiments

4.1. Datasets and metrics

In this paper, DF2K (DIV2K + Flickr2K) [48] with 3450 images are used for training the proposed model from scratch. During testing, we select five standard benchmark datasets: Set5 [3], Set14 [58], BSD100 [41], Urban100 [16] and Manga109 [42]. The low-resolution images are generated from the ground truth images by the "bicubic" downsampling in MATLAB. PSNR/SSIM measured by discarding a 4-pixel boundary around the images, and calculated on the Y channel is reported for the quantitative metrics.

4.2. Implementation details

During training, we set the input patch size to 256×256 and use random rotation and horizontal flipping for data augmentation. The batch size is set to 128 and the total number of iterations is 500k. The initial learning rate is set to 2×10^{-4}. We adopt a multi-step learning rate strategy, where the learning rate will be halved when the iteration reaches 250000, 400000, 450000, and 475000, respectively. Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is used to train the model.

Distillation training. In the teacher learning phase, we utilize the DF2K dataset with 2K resolution to train the teacher network, which comprises 8 RSSB and 2 ViMM blocks with 192 channels. During the distillation training phase, we use DF2K datasets for the student network, which contains 4 RSSB and 2 ViMM blocks with 60 channels.

4.3. Comparison with State-of-the-art SR models

We compare DVMSR with several advanced efficient super-resolution model [2, 9, 10, 17, 19, 25, 26, 28, 30, 36, 46, 47, 56, 59]. The quantitative performance comparison on several benchmark datasets [3, 16, 41, 42, 58] is indicated in Table 1. Our experimental results showcase our ability to achieve smaller parameter counts while surpassing several previous methods on five benchmark datasets. Specifically, we attained higher SSIM scores on Set5, Set14, and BSD100. It's important to note that SSIM scores serve as a crucial metric, indicating how effectively our model preserves the structure and content of the images, ultimately resulting in reconstructions that closely resemble the original images. Additionally, we observed that PSNR values remain comparable across these five datasets. This comprehensive evaluation underscores the effectiveness of our approach in enhancing image quality while maintaining efficiency, making it a promising solution for various image enhancement tasks. It's worth emphasizing that in our current study, we directly utilize the final model architecture employed in the NTIRE competition. Remarkably, we manage to maintain excellent performance without unnecessarily inflating the parameter count. This strategic decision underscores our commitment to efficiency and effectiveness in model design, ensuring that our approach remains practical and scalable for real-world applications.

Model complexity comparisons between SwinIR and DVMSR. Our investigation focuses on Mamba's performance in super-resolution (SR) tasks. In Fig. 2, we show the excellent long-range modeling capabilities of our

Table 1. Average PSNR/SSIM for scale factor 4 on datasets Set5, Set14, BSD100, Urban100, and Manga109. The best and second best results are highlighted in red and blue respectively.

Method	Params	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	-	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [9]	8K	30.48/0.8626	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
FSRCNN [10]	13K	30.72/0.8660	27.61/0.7550	26.98/0.7150	24.62/0.7280	27.90/0.8610
VDSR [26]	666K	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524	28.83/0.8870
DRCN [25]	1774K	31.53/0.8854	28.02/0.7670	27.23/0.7233	25.14/0.7510	28.93/0.8854
LapSRN [30]	502K	31.54/0.8852	28.09/0.7700	27.32/0.7275	25.21/0.7562	29.09/0.8900
DRRN [46]	298K	31.68/0.8888	28.21/0.7720	27.38/0.7284	25.44/0.7638	29.45/0.8946
MemNet [47]	678K	31.74/0.8893	28.26/0.7723	27.40/0.7281	25.50/0.7630	29.42/0.8942
IDN [17]	553K	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632	29.41/0.8942
SRMDNF [59]	1552K	31.96/0.8925	28.35/0.7787	27.49/0.7337	25.68/0.7731	30.09/0.9024
CARN [2]	1592K	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.47/0.9084
IMDN [19]	715K	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
RFDN [36]	550K	32.24/0.8952	28.61/0.7819	27.57/0.7360	26.11/0.7858	30.58/0.9089
RLFN [28]	543K	32.240.8952	28.62/0.7813	27.60/0.7364	26.17/0.7877	-/-
DIPNet [56]	543K	32.20/0.8950	28.58/0.7811	27.59/0.7364	26.16/0.7879	30.53/0.9087
DVMSR (Ours)	424K	32.19/0.8955	28.61/0.7823	27.58/0.7379	26.03/0.7838	30.48/0.9084

DVMSR using LAM. Additionally, we compare DVMSR with SwinIR, a transformer-based model, in terms of model complexity. SwinIR outperforms DVMSR by 0.23 dB in PSNR, but at the cost of approximately twice the number of parameters, significantly higher FLOPS, and about 20 times longer inference time. These findings suggest that Mamba-based models hold promise for efficient SR.

Table 2. Model complexity comparisons between SwinIR and DVMSR. Times represent the average inference time measured on the DIV2K dataset with an Nvidia RTX 3090 in seconds (s). FLOPS and memory is measured when the input is 256×256. PSNR is the result of testing on DIV2K.

Method	PSNR	Time (s)	Params[M]	FLOPS[G]	Activations	Memory[M]
SwinIR	29.20 dB	0.865	0.9296	70.7828	26.7387	1454.458
DVMSR	28.97 dB	0.048	0.4244	20.1680	26.7387	1094.245

4.4. Ablation Study

4.4.1 Model Parameter Analysis

Here, we train DVMSR on DIV2K for classical image SR (×4) and test it on Set5 and Set14.

Impact of ViMM number. We show the effects of ViMM number in each RSSB on model performance in Table 3. In experiments 1 - 3, it is observed that the PSNR/SSIM is negatively correlated with the number of ViMMs. However, when we set the ViMM number to 1, as presented in experiment 4, the PSNR in Set5 and Set14 decreased by 0.09 dB compared to when the ViMM number is set to 2. Therefore, there may be a balance point for the ViMM number, where

it should not be too large to avoid over-complexity of the model, nor too small to limit the model's ability to represent the data. Experimental results indicate that setting the ViMM number to 2 is appropriate.

Table 3. Impact of ViMM number in each RSSB on the Set5 and Set14 datasets with scale factor of ×4. The number of RSSB is fixed at 4 and keep other parameter settings consistent. The best results are highlighted.

Exp.	Params[M]	ViMM number	Set5 PSNR/SSIM	Set14 PSNR/SSIM
1	7.222	6,6,6,6	31.99/0.8926	28.44/0.7785
2	5.214	2,2,9,2	32.17/0.8959	28.63/0.7834
3	3.651	2,2,2,2	**32.30/0.8972**	**28.68/0.7847**
4	2.758	1,1,1,1	32.21/0.8954	28.59/0.7821

Impact of RSSB number. In Table 4, In Experiments 1-3, as the RSSB number increases, the parameter count increases, with the channel number set to 180. Along with the increase in RSSB number, the PSNR in Set5 shows a significant improvement. Compared to Experiment 1, Experiment 2 shows an increase of 0.26 dB, and relative to Experiment 2, Experiment 3 shows an increase of 0.13 dB. When we set the RSSB number to 10, the improvement is moderated, with Experiment 4 showing an increase of 0.01 dB relative to Experiment 3.

Impact of channel number. We maintained the ViMM number and RSSB number while examining the influence of channel numbers on model performance, as detailed in Table 5. Notably, our analysis revealed a diminishing improvement in model performance when the channel number

Table 4. Impact of RSSB number on the Set5 and Set14 datasets with scale factor of ×4. The number of ViMM is fixed at 2 and keeps other parameter settings consistent. The best results are highlighted.

Exp.	Params[M]	RSSB number	Set5 PSNR/SSIM	Set14 PSNR/SSIM
1	2.175	2	32.04/0.8938	28.51/0.7799
2	3.651	4	32.30/0.8972	28.68/0.7847
3	5.128	6	32.43/0.8987	28.75/0.7866
4	8.080	10	**32.44/0.8990**	**28.77/0.7874**

was set to 210. Thus, we conclude that setting the channel number to 192 is more suitable for optimal model performance.

Table 5. Impact of channel number on the Set5 and Set14 datasets with scale factor of ×4. keep other parameter settings consistent. The best results are highlighted.

Exp.	Params[M]	channel number	Set5 PSNR/SSIM	Set14 PSNR/SSIM
1	2.664	150	32.32/0.8971	28.65/0.7838
2	3.651	180	32.30/0.8972	28.68/0.7847
3	4.089	192	32.37/0.8977	**28.71/0.7851**
4	4.809	210	**32.39/0.8976**	28.71/0.7850

Table 6. Comparison of unidirectional SSM or bidirectional SSM. Times represent the average inference time measured on the DIV2K dataset with an Nvidia RTX 3090 in seconds (s). FLOPS and memory are measured when the input is 256×256. PSNR is the result of testing on DIV2K.

Method	PSNR	Time (s)	Params[M]	FLOPS[G]	Activations	Memory[M]
unidirectional SSM	28.87 dB	0.048	0.4244	20.1680	26.7387	1094.245
bidirectional SSM	28.88 dB	0.087	0.4849	23.9429	26.7387	1451.680

4.4.2 Distillation Learning

Distillation loss. To investigate the effectiveness of distillation loss, we tried multiple distillation strategies. Mid-level feature distillation and end-level feature distillation are presented in Figure 6. As shown in Table 7, using the end-level feature distillation method tends to increase the PSNR and SSIM on Set5 and Set14 datasets. This suggests that **the features towards the end of the model might be closer to the target output of the SR task.** When attempting to alter the weights and types of distillation loss in the mid-level feature distillation method, there were no changes observed in PSNR and SSIM values on Set5 and Set14 datasets. This indicates that it is difficult for the student model to benefit from the features of the middle layer of the teacher model, as even after modifying the weights and types of distillation loss, there were no significant changes in the results. When we increase the weight of distillation loss in the end-level
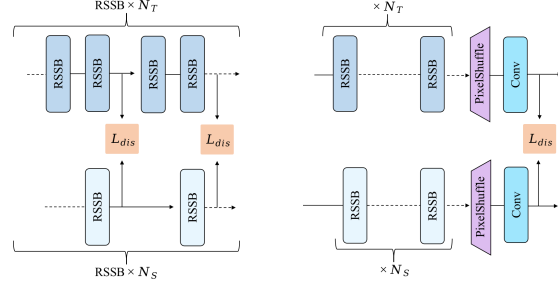

Figure 6. Left: The structure of mid-level feature distillation; Right: The structure of end-level feature distillation

feature distillation method, there is a slight decrease in the PSNR and SSIM on Set5 and Set14 datasets. This could be **because excessively high weights on distillation loss might introduce too many constraints, thereby affecting the model's performance.**

Table 7. Impact of the distillation loss. "✗" signifies that distillation is not used, and "✔" signifies that distillation is used. "mid" and "end" represent mid-level feature distillation and end-level feature distillation, respectively. $\mathcal{L}_{dis} : \mathcal{L}_1$ represents the weight ratio of the distillation loss and $\mathcal{L}_1$ loss.

distillation strategy	distillation position	distillation loss	$\mathcal{L}_{dis} : \mathcal{L}_1$	Set5 PSNR/SSIM	Set14 PSNR/SSIM
✗	-	-	-	32.04/0.8940	28.50/0.7801
✔	mid	$\mathcal{L}_1$	1:1	32.11/0.8949	28.56/0.7811
✔	mid	$\mathcal{L}_1$	5:1	32.11/0.8949	28.56/0.7811
✔	mid	$\mathcal{L}_2$	1:1	32.11/0.8949	28.56/0.7811
✔	end	$\mathcal{L}_1$	1:1	**32.12/0.8951**	**28.57/0.7813**
✔	end	$\mathcal{L}_1$	5:1	32.11/0.8950	28.57/0.7813

Teacher model. When the teacher model has more parameters and richer representation capability, the knowledge it transfers to the student model will be more abundant, leading to a more significant performance improvement of the student model on the task. To verify this conclusion, we attempted two teacher models with different parameters. They exhibited a PSNR difference of 0.27dB on the Set5 dataset. However, as shown in Table 8, the performance of the student model remained unchanged. This could indicate that the student model's capacity or architecture may not be sufficiently expressive to fully utilize the additional knowledge provided by the larger teacher model. Therefore, finding the balance point between the performance of the teacher model and the student model is a worthwhile exploration.

4.4.3 Unidirectional v.s. Bidirectional SSM

To investigate the effectiveness of bidirectional SSM in ESR, we evaluate its performance in ESR based on several aspects: PSNR, Time, Params, FLOPS, Activations, and Memory. The architecture of unidirectional SSM and

Table 8. Design of the teacher model. PSNR is the result of testing on Set5. Params is the parameter of teacher model, and the parameter of student model is fixed.

Method	Params[M]	Teacher model PSNR/SSIM	Student model PSNR/SSIM
DVMSR	-	-	32.04/0.8940
DVMSR	4.089	32.38/0.8977	32.12/0.8950
DVMSR	7.432	32.65/0.9011	32.12/0.8950
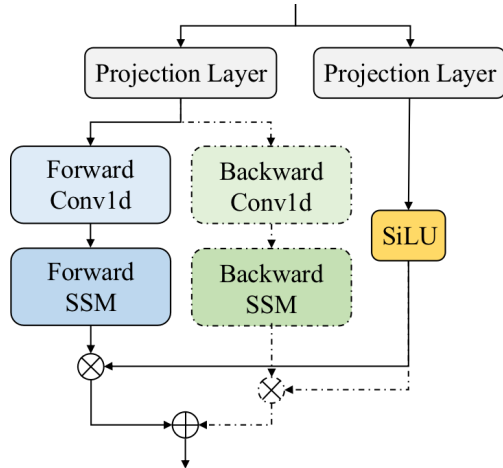
Figure 7. Unidirectional SSM or bidirectional SSM in ViMM.

bidirectional SSM are presented in Figure 7. As shown in Table 6, compared to Unidirectional SSM, the improvement of bidirectional SSM in PSNR is limited (increased by 0.01dB), while the inference time has increased by 0.039s. This cost is significant. Therefore, Unidirectional SSM is more suitable for the ESR task.

4.4.4 NTIRE 2024 Challenge on Efficient SR

We actively participate in the NTIRE 2024 Efficient Super-Resolution Challenge [45]. The model structure and training strategy are slightly different from the above. This competition aims to procure solutions that excel in overall performance metrics, encompassing inference runtime, FLOPS, and parameter optimization on the NVIDIA GeForce RTX 3090 GPU. This challenge also requires the maintenance or enhancement of threshold PSNR results, underscoring the importance of efficiency without compromising on image quality benchmarks.

During the teacher learning phase, we train the teacher network using the DIV2K dataset with a resolution of 2K. Our teacher architecture consists of 6 RSSB (Residual Scaling and Shifting Block) and 2 ViMM (Vision Mamba Modules), each configured with 180 channels.

In the subsequent distillation training phase, we amalgamated data from both the DIV2K and LSDIR datasets to train the student network. This student model comprises 2 RSSB and 2 ViMM blocks, tailored with 60 channels to maintain computational efficiency while preserving performance standards. Notably, the teacher network remains unchanged.

We employ DIV2K [48] and LSDIR [32] to construct the training dataset. The High-Resolution (HR) images are cropped to 256×256 patches for the training procedure.

During network optimization, we employ the $\mathcal{L}_1$ loss function in conjunction with the Adam optimizer, a widely adopted optimization algorithm in deep learning tasks. Our optimization regimen commenced with an initial learning rate of 2×10^{-4}, evolving through a multi-step learning rate strategy. Specifically, the learning rate halved at key iterations: 250000, 400000, 450000, and 475000, respectively, throughout the 500k total iterations. This adaptive learning rate scheme enhances model convergence and stability over the training period, crucial for achieving superior performance.

Through extensive experiments, we refine our model's architecture and training process, aiming for excellence in both efficiency and performance, as evidenced by our results in Table 9. Our approach employs a novel architecture that differs from both CNN and transformer, providing a reference for the development of mamba in Efficient Super-Reslution.

Table 9. NTIRE 2024 ESR Challenge results.

Model	Val PSNR (dB)	Test PSNR (dB)	Val Time (ms)	Test Time (ms)	FLOPS (G)	Params (M)
RLFN_baseline	26.96	27.07	14.348	9.194	19.67	0.317
DVMSR	26.93	27.04	40.75	34.634	20.17	0.424

5. Conclusion

In this paper, we propose DVMSR, a novel lightweight Image SR network that incorporates Vision Mamba and a distillation strategy. It consists of three main modules: feature extraction convolution, multiple stacked Residual State Space Blocks (RSSBs), and a reconstruction module. In particular, we use a stack of residual state space blocks (RSSB) for deep feature extraction, and each RSSB is composed of Vision Mamba Moudles, a convolution layer and a residual connection. Specifically, we leverage the larger teacher model as additional supervision, which effectively enhances the performance of the student model. DVMSR demonstrates the potential for efficient and long-range dependency modeling in SR tasks, but our work merely offers a preliminary insight. We still need further to explore the potential of Mamba in ESR tasks.

References

[1] Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. Fitnets: Hints for thin deep nets. *Proc. ICLR*, 2(3):1, 2015. 3

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV (10)*, pages 256–272. Springer, 2018. 5, 6

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

[4] Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *arXiv preprint arXiv:2403.19654*, 2024. 2, 3

[5] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration. *arXiv preprint arXiv:2309.05239*, 2023. 2

[6] 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. 2

[7] 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. 3

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 1

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV (4)*, pages 184–199. Springer, 2014. 1, 2, 5, 6

[10] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV (2)*, pages 391–407. Springer, 2016. 1, 2, 5, 6

[11] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[14] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024. 2

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

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

[17] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, pages 723–731. IEEE Computer Society, 2018. 1, 2, 5, 6

[18] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018. 3

[19] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM Multimedia*, pages 2024–2032. ACM, 2019. 1, 2, 3, 5, 6

[20] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019. 3

[21] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022. 2

[22] Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758, 2023. 2

[23] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 2

[24] Fahad Shahbaz Khan and Salman Khan. Ntire 2022 challenge on efficient super-resolution: Methods and results. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1061–1101, 2022. 1

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645. IEEE Computer Society, 2016. 1, 2, 5, 6

[26] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654. IEEE Computer Society, 2016. 1, 2, 5, 6

[27] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 3

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–776, 2022. 1, 2, 5, 6

[29] Xiangtao Kong, Hengyuan Zhao, Yu Qiao, and Chao Dong. Classsr: A general framework to accelerate super-resolution networks by data characteristic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12016–12025, 2021. 1

[30] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, pages 5835–5843. IEEE Computer Society, 2017. 1, 2, 5, 6

[31] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[32] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 8

[33] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1959, 2023. 2

[34] 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*, pages 833–843, 2022. 1

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2, 3

[36] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 41–55. Springer, 2020. 1, 2, 3, 5, 6

[37] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 41–55. Springer, 2020. 3

[38] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 2, 3

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[40] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2, 3

[41] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–425, 2001. 5

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

[43] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022. 2

[44] Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024. 2, 3

[45] 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. 8

[46] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, pages 2790–2798. IEEE Computer Society, 2017. 2, 5, 6

[47] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, pages 4549–4557. IEEE Computer Society, 2017. 2, 5, 6

[48] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 5, 8

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[50] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023. 2

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[52] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *ECCV Workshops (5)*, pages 63–79. Springer, 2018. 2

[53] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 2

[54] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 777–785, 2022. 3

[55] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*, 2024. 2, 3

[56] Lei Yu, Xinpeng Li, Youwei Li, Ting Jiang, Qi Wu, Haoqiang Fan, and Shuaicheng Liu. Dipnet: Efficiency distillation and iterative pruning for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1701, 2023. 1, 2, 3, 5, 6

[57] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 3

[58] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730. Springer, 2010. 5

[59] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, pages 3262–3271. IEEE Computer Society, 2018. 5, 6

[60] Kai Zhang, Shuhang Gu, Radu Timofte, Zheng Hui, Xiumei Wang, Xinbo Gao, Dongliang Xiong, Shuai Liu, Ruipeng Gang, Nan Nan, et al. Aim 2019 challenge on constrained super-resolution: Methods and results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3565–3574. IEEE, 2019. 1

[61] Kai Zhang, Martin Danelljan, Yawei Li, Radu Timofte, Jie Liu, Jie Tang, Gangshan Wu, Yu Zhu, Xiangyu He, Wenjie Xu, et al. Aim 2020 challenge on efficient super-resolution: Methods and results. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 5–40. Springer, 2020. 1

[62] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[63] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3096–3105, 2019. 2

[64] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Super resolution generative adversarial networks with learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7149–7166, 2021. 2

[65] Wenlong Zhang, Guangyuan Shi, Yihao Liu, Chao Dong, and Xiao-Ming Wu. A closer look at blind super-resolution: Degradation models, baselines, and performance upper bounds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 527–536, 2022. 2

[66] Wenlong Zhang, Xiaohui Li, Xiangyu Chen, Yu Qiao, Xiao-Ming Wu, and Chao Dong. Seal: A framework for systematic evaluation of real-world super-resolution. *arXiv preprint arXiv:2309.03020*, 2023. 2

[67] Wenlong Zhang, Xiaohui Li, Guangyuan Shi, Xiangyu Chen, Yu Qiao, Xiaoyun Zhang, Xiao-Ming Wu, and Chao Dong. Real-world image super-resolution as multi-task learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[68] Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2021. 3

[69] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2, 3