Contrast: A Hybrid Architecture of Transformers and State Space Models for Low-Level Vision

Aman Urumbekov
Kyrgyz State Technical University
Kyrgyzstan
amanurumbekov@gmail.com

Zheng Chen
Shanghai Jiao Tong University
Shanghai
zhengchen.cse@gmail.com

Abstract

*Transformers have become increasingly popular for image super-resolution (SR) tasks due to their strong global context modeling capabilities. However, their quadratic computational complexity necessitates the use of window-based attention mechanisms, which restricts the receptive field and limits effective context expansion. Recently, the Mamba architecture has emerged as a promising alternative with linear computational complexity, allowing it to avoid window mechanisms and maintain a large receptive field. Nevertheless, Mamba faces challenges in handling long-context dependencies when high pixel-level precision is required, as in SR tasks. This is due to its hidden state mechanism, which can compress and store a substantial amount of context but only in an approximate manner, leading to inaccuracies that transformers do not suffer from. In this paper, we propose **Contrast**, a hybrid SR model that combines **Con**volutional, **Tra**nsformer, and **St**ate Space components, effectively blending the strengths of transformers and Mamba to address their individual limitations. By integrating transformer and state space mechanisms, **Contrast** compensates for the shortcomings of each approach, enhancing both global context modeling and pixel-level accuracy. We demonstrate that combining these two architectures allows us to mitigate the problems inherent in each, resulting in improved performance on image super-resolution tasks.*

1. Introduction

Single image super-resolution (SR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) input, serving as a fundamental task in low-level vision. This problem is inherently ill-posed, as multiple plausible solutions can exist for any given LR input. In recent years, a variety of methods have been explored to improve SR performance, balancing computational efficiency with the ability to capture fine-grained details.

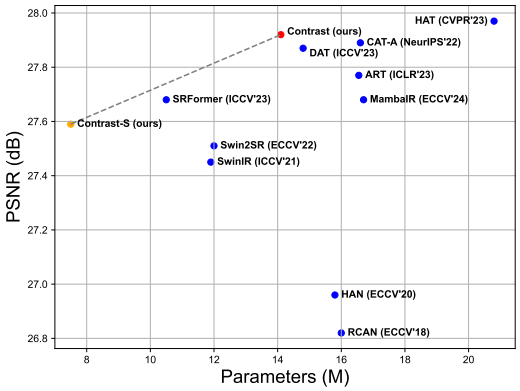

Figure 1. Model comparison on the Urban100 dataset for $\times 4$ SR. The plot illustrates the trade-off between PSNR and parameter count, where higher PSNR values and fewer parameters (upper-left) indicate better performance. This highlights the effectiveness of **Contrast** on the Urban100 benchmark.

Convolutional Neural Networks (CNNs) [15, 16, 33] were the early choice for SR tasks [8, 10, 27, 45], effectively capturing local structures in images. However, due to the local nature of convolutions, CNNs struggle to model long-range dependencies, a key component for capturing contextual information in complex images. This limitation has prompted the search for alternative methods that can better handle global context in SR.

Transformers [38], originally developed for natural language processing (NLP), have shown strong potential in computer vision, particularly for high-level tasks [11, 36, 37], due to their self-attention (SA) mechanism, which directly models long-range dependencies. By capturing global context, transformers help overcome the local limitations of CNNs. However, their quadratic computational complexity with image resolution becomes prohibitive in high-resolution SR tasks. To mitigate this, SR models typi-

cally employ window-based self-attention [19, 24], limiting the attention mechanism to local windows to reduce computational costs. While this approach effectively addresses the quadratic complexity, it introduces a constrained receptive field that is difficult to expand, limiting the model's ability to capture broader image context essential for high-quality SR outcomes.

Recently, the **Mamba** architecture [9, 12] has emerged as a promising alternative, emphasizing efficient global modeling with linear complexity scaling. Mamba is capable of extending receptive fields across the entire image without relying on window mechanisms, thereby alleviating issues associated with limited receptive fields in window-based attention. This makes it particularly attractive for SR tasks, where capturing global context is crucial. Furthermore, increasing Mamba's efficiency does not require the introduction of window mechanisms, allowing it to maintain a global receptive field inherently.

However, Mamba faces challenges when high pixel-level precision is required, as in SR tasks. Its hidden state mechanism allows it to compress and store substantial context information, but the representation is approximate rather than exact. This approximation can lead to inaccuracies in modeling long-context dependencies, especially when fine details are essential for accurate image reconstruction. In contrast, transformers do not suffer from this problem due to their precise attention mechanisms.

In this work, we demonstrate that by combining the strengths of transformers and Mamba, we can address the limitations inherent in each architecture. We propose **Contrast**, a hybrid SR model that integrates **Con**volutional, **Tra**nsformer, and **St**ate Space components, effectively blending the advantages of both transformers and Mamba. The hybrid Contrast model leverages the receptive field expansion and efficient complexity scaling of Mamba, while simultaneously benefiting from the transformer's capacity to model complex spatial dependencies with high pixel-level accuracy. By merging these capabilities, Contrast enhances the modeling of both local and global context, providing a balanced solution for SR that neither architecture achieves alone.

Figure 1 demonstrates that Contrast outperforms other models on the Urban100 [17] dataset in terms of PSNR relative to the number of parameters. Urban100 comprises high-resolution images with repetitive patterns, posing a challenge ideal for models with extensive receptive fields. Notably, Contrast achieves a PSNR of 27.92 with 14.1 million parameters, compared to MambaIR (ECCV'24) [14] which attains 27.68 PSNR with 16.7 million parameters, and HAT (CVPR'23) [3] which reaches 27.97 PSNR with 20.8 million parameters. This highlights the efficiency of Contrast, a hybrid model combining transformer and Mamba architectures in a ratio of one transformer block to six Mamba blocks, in delivering high-quality image reconstruction with fewer parameters, especially when contrasted with MambaIR and HAT.

In summary, adding window mechanisms to transformers can significantly alleviate the issue of quadratic complexity but at the cost of a limited receptive field that is challenging to expand. Mamba, with its linear complexity, avoids the need for window mechanisms altogether, inherently possessing a global receptive field. By integrating the two architectures, Contrast effectively compensates for their respective shortcomings, achieving superior performance in image super-resolution tasks.

2. Related Work

Image Super-Resolution. Deep learning has revolutionized image super-resolution (SR), with Convolutional Neural Networks (CNNs) initially setting the benchmark for performance. Pioneering work like SRCNN [10] introduced CNNs to SR, achieving significant improvements over traditional methods. Advanced architectures such as RCAN [45] utilized deep residual networks exceeding 400 layers to enhance feature extraction. Additionally, attention mechanisms [40, 45] were integrated to focus on important spatial and channel-wise information. Despite these advancements, CNNs inherently struggle to capture long-range dependencies due to the local nature of convolution operations, limiting their ability to model global context effectively.

Vision Transformers in Low-Level Vision. Transformers [38], originally designed for natural language processing, have shown great promise in computer vision tasks [11, 36, 37] due to their self-attention mechanism, which can model long-range dependencies. In SR, Transformers help overcome the limitations of CNNs by capturing global context. However, the quadratic computational complexity of self-attention with respect to input size poses significant challenges for high-resolution images typical in SR tasks. To alleviate this, models like SwinIR [19, 24] employ window-based self-attention, where attention is calculated within local windows to reduce computational costs. While this approach effectively mitigates quadratic complexity, it introduces a limited receptive field that is difficult to expand, constraining the model's ability to capture broader image context essential for high-quality SR outcomes.

State Space Models. State space models (SSMs) have recently emerged as a promising alternative for modeling long-range dependencies with linear computational complexity. Architectures like Mamba [9, 12] exploit SSMs to efficiently capture global context, inherently providing a global receptive field. This makes them particularly attractive for SR tasks, where understanding the entire image context is crucial. However, SSM-based models like

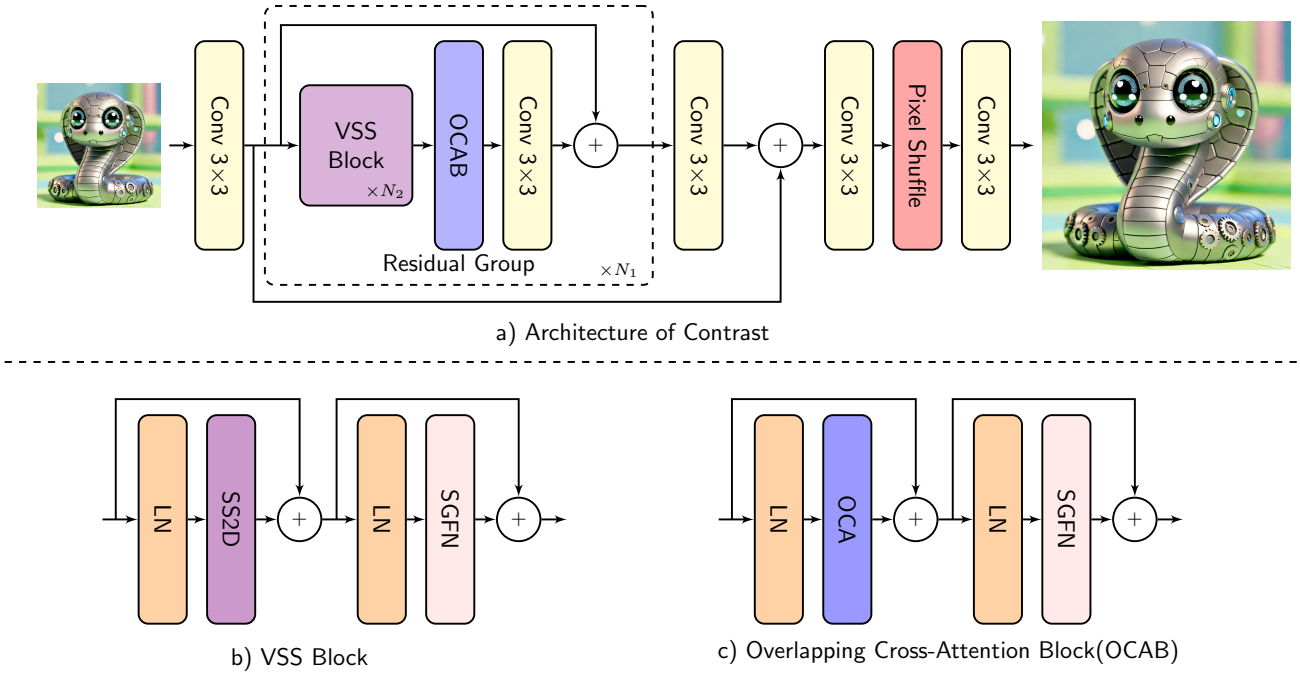

Figure 2. The network architecture of our method. (a) The architecture of Contrast. (b) Illustration of VSS Block. (c) Illustration of Overlapping Cross-Attention Block (OCAB).

Mamba face challenges when precise pixel-level accuracy is required. The hidden state mechanism compresses context information in an approximate manner, which can lead to inaccuracies in modeling long-context dependencies. Additionally, their sequential scanning mechanism makes it difficult to capture diagonal dependencies, which are important for reconstructing detailed spatial patterns.

Hybrid Models. Numerous architectures have explored hybrid models that integrate Transformers with SSMs like Mamba, often employing full attention mechanisms in the Transformer components [39]. However, few studies have delved into providing a comprehensive explanation for the effectiveness of such integrations. Our work distinguishes itself by offering greater clarity and laying the groundwork for the development of these hybrid ideas. Specifically, our proposed Contrast model combines Transformer blocks with windowed attention mechanisms and Mamba blocks to harness their respective strengths. The Mamba blocks provide a global receptive field, effectively capturing long-range dependencies essential for super-resolution tasks. In parallel, Transformer blocks with windowed attention refine the global information from the Mamba blocks, enhancing the model's ability to capture detailed features. Additionally, we integrate convolutional layers within the MLP layers to model local pixel relationships, particularly focusing on diagonal pixels. This convolutional enhancement ensures that Contrast maintains high pixel-level accuracy while effectively capturing both local and global dependencies. By synergizing the global context capabilities of Mamba with the refinement capabilities of Transformer blocks and enhancing local pixel relationships through convolutional layers, Contrast overcomes the individual limitations of Transformers and SSMs when used in isolation.

3. Methodology

3.1. Architecture

The Contrast model is designed as a hybrid framework with three primary modules: shallow feature extraction, deep feature extraction, and image reconstruction, as illustrated in Fig. 2. This architecture balances computational efficiency and performance in super-resolution (SR) tasks by combining Visual State Space (VSS) Blocks[23], Overlapping Cross-Attention Blocks (OCAB)[3], and Spatial Gated Feed-Forward Networks (SGFN)[5]. Each module is tailored to extract and refine features at multiple levels, progressively enhancing the image from low resolution to high resolution.

3.1.1 Shallow Feature Extraction

Given a low-resolution (LR) input image $I_{LR} \in \mathbb{R}^{H\times W\times 3}$, the shallow feature extraction module first applies a convolutional layer to map I_{LR} to an initial feature space:

$$F_S = \text{Conv}(I_{LR}), \tag{1}$$

where $F_S \in \mathbb{R}^{H\times W\times C}$ represents the shallow feature map, with H and W denoting the spatial dimensions of the input

image, and C the number of feature channels. This initial step provides a feature-rich representation for further processing.

3.1.2 Deep Feature Extraction

The deep feature extraction module builds upon the shallow features F_S to capture complex spatial and channel relationships, outputting a deeper feature representation $F_D \in \mathbb{R}^{H\times W\times C}$. This module is composed of N_1 stacked Residual Groups (RGs), each containing N_2 Visual State Space (VSS) Blocks followed by an Overlapping Cross-Attention Block (OCAB).

In each Residual Group (RG), features are processed through a sequence of VSS Blocks, arranged to optimize computational efficiency and performance in hybrid architectures. Specifically, we utilize a structure of six Mamba blocks followed by a single Transformer block, a ratio that has demonstrated effectiveness and was empirically validated in prior work [39]. This setup aligns with the original HAT architecture, where we replaced the (S)W-MSA modules with Mamba blocks while preserving the original six-to-one block ratio. Although we considered adding more Transformer layers, the Overlapping Cross-Attention Block (OCAB) required for cross-window information flow is computationally intensive due to its overlapping nature. This increased the training time substantially and slowed the model to a point where further improvements in metrics could not justify the trade-off in efficiency, prompting us to maintain the initial six-to-one configuration.

The nested structure of the deep feature extraction module, composed of N_1 RGs, progressively refines the shallow features F_S. The output of each RG can be represented as:

$$F_{RG}^{(j)} = \text{Conv}(\text{OCAB}(\text{VSS}_{N_2}(F_{RG}^{(j-1)}))) + F_{RG}^{(j-1)}, \quad (2)$$

where j denotes the index of the current RG, $F_{RG}^{(j-1)}$ is the input to the RG, and VSS_{N_2} represents the sequential application of N_2 VSS Blocks within the RG. Each VSS Block applies a state-space model to capture global dependencies with reduced computational overhead, as follows:

$$F_{VSS}^{(i)} = \text{VSS}(F_{VSS}^{(i-1)}), \quad (3)$$

where i denotes the index of the VSS Block within an RG.

Following the VSS Blocks, the OCAB operation integrates cross-window information to enhance spatial coherence, ensuring a more continuous representation across feature windows. This is formulated as:

$$F_{OCAB} = \text{OCAB}(F_{VSS}^{(N_2)}), \quad (4)$$

where F_{OCAB} represents the output after the OCAB step, which is then refined by a convolution layer to produce the final output of the RG.

Thus, after passing through all N_1 Residual Groups, the deep feature extraction module yields the deep feature representation F_D, where each RG applies the combined operations of VSS Blocks, OCAB, and residual connections to progressively enrich the feature map.

3.1.3 Spatial Gated Feed-Forward Network (SGFN)

Instead of conventional MLP layers within each VSS Block, we employ the Spatial Gated Feed-Forward Network (SGFN) based on [5]. The SGFN adds a spatial gate to the Feed-Forward Network (FFN) layers, addressing the limitations of traditional FFNs in spatial modeling. The SGFN applies non-linear activation and two linear projections, with an additional spatial gate for selective channel-wise information flow. Overall, given the input $\hat{X}\in\mathbb{R}^{H\times W\times C}$, SGFN is formulated as

$$\begin{aligned} \hat{X}' &= \sigma(W_p^1\hat{X}), \quad [\hat{X}_1', \hat{X}_2'] = \hat{X}', \\ \text{SGFN}(\hat{X}) &= W_p^2(\hat{X}_1' \odot (W_d\hat{X}_2')), \end{aligned} \quad (5)$$

where W_p^1 and W_p^2 indicate linear projection layers, σ is the GELU activation function, and W_d represents the learnable parameters for the depth-wise convolution. Both $\hat{X}_1'$ and $\hat{X}_2'$ lie in $\mathbb{R}^{H\times W\times \frac{C'}{2}}$ space, with C' denoting SGFN's hidden dimension. Unlike the standard FFN, SGFN is designed to capture non-linear spatial information and reduce channel redundancy in fully-connected layers.

3.1.4 Image Reconstruction

The image reconstruction module takes the deep feature representation F_D and upscales it to produce the final high-resolution (HR) output image $I_{HR} \in \mathbb{R}^{H_{out}\times W_{out}\times 3}$. The upsampling is performed using the pixel shuffle method [32], ensuring efficient spatial upscaling:

$$F_{up} = \text{PixelShuffle}(F_D), \quad (6)$$

where F_{up} is the upscaled feature map. After upsampling, a convolutional layer aggregates the upscaled features to produce the HR image:

$$I_{HR} = \text{Conv}(F_{up}). \quad (7)$$

This series of transformations enhances the input LR image into a high-quality SR output, maintaining spatial continuity and ensuring fine-grained detail preservation across the reconstructed image.

4. Experiments

4.1. Experimental Settings

Implementation Details. We build two variants of the Contrast model with different complexity levels, called Contrast and Contrast-S. For the main Contrast model, we use

6 Residual Groups (RGs) with an embedding dimension of 210, a window size of 32, an MLP ratio of 2, an SSM state dimension of 1, and an SSM ratio of 1, as shown in the VMamba architecture. In the Contrast-S variant, we also use 6 RGs but with a reduced embedding dimension of 150 and a smaller window size of 16, while keeping the MLP ratio, SSM state dimension, and SSM ratio the same as in the main model.

In both models, we use the Mamba-1 module rather than Mamba-2. The authors of Mamba [9, 12] have noted that Mamba-2 provides minimal improvements in metrics over Mamba-1, with the primary enhancements focused on speeding up large language models (LLMs). For our super-resolution (SR) task, where we employ smaller embedding dimensions, Mamba-2 performed significantly slower than Mamba-1 in our experiments. Thus, Mamba-1 was chosen for Contrast and Contrast-S to balance efficiency and performance in SR.

Data and Evaluation. We follow the standard practices established in previous works [4, 5] for training and evaluating our models. Specifically, we use two large-scale datasets for training: DIV2K [35] and Flickr2K [21]. For evaluation, we test on five widely-used benchmark datasets: Set5 [2], Set14 [42], B100 [25], Urban100 [17], and Manga109 [26].

Our experiments cover upscaling factors of $\times 2$, $\times 3$, and $\times 4$. Low-resolution (LR) images are generated from high-resolution (HR) images using bicubic degradation. To evaluate super-resolution (SR) performance, we use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [41], both calculated on the Y channel (luminance) of the YCbCr color space. These metrics provide objective assessments of image quality and structural preservation in SR results.

Training Settings. We train the models with a patch size of 64×64 and a batch size of 32 for 500K iterations. The optimization is performed by minimizing the L1 loss using the Adam optimizer [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 1×10^{-4} for the main Contrast model and 2×10^{-4} for Contrast-S. The learning rate is halved at milestones: [250K, 400K, 450K, 475K]. During training, data augmentation is applied by randomly rotating images by 90°, 180°, and 270°, along with horizontal flips. Our model is implemented in PyTorch[31] and Triton[34] and trained on 4 A6000 GPUs.

4.2. Ablation Study

In this section, we conduct a series of ablation studies to investigate the effectiveness of different components in our proposed **Contrast** model. We aim to understand how the integration of Transformer and State Space components contributes to the model's performance, and we analyze the impact of architectural choices on training dynamics and final results. All experiments are conducted with models of comparable parameter sizes to ensure a fair comparison.

Effect of Replacing Transformer Blocks with SS2D. We begin by exploring the impact of replacing the (Shifted) Window Multi-Head Self-Attention ((S)W-MSA) blocks in the baseline Transformer model HAT [3] with the State Space 2D (SS2D) blocks from Mamba [12]. This modification results in an initial hybrid architecture, allowing us to evaluate whether SS2D provides a more effective solution for capturing dependencies and spatial context compared to traditional self-attention mechanisms.

Training Dynamics of Pure Mamba, Transformer, and Hybrid Models. To further investigate the individual contributions of the Transformer and Mamba architectures, we trained three models:

- Pure Mamba: A model where all Transformer blocks are replaced with SS2D blocks, and Channel Attention Blocks (CABs) are removed.
- Pure Transformer: The baseline HAT model without modifications.
- Hybrid Contrast: The baseline HAT model with (S)W-MSA replaced by SS2D blocks.

All models have approximately the same number of parameters to ensure a fair comparison. We observed an interesting dynamic during training, as illustrated in Figure 3. The pure Mamba model achieves high PSNR values early in training but exhibits slow progress afterward. In contrast, the Transformer model starts with lower PSNR values but continues to improve steadily, eventually surpassing the pure Mamba model. The hybrid Contrast model combines the strengths of both architectures: it starts with high PSNR values like Mamba and continues to improve steadily like the Transformer, ultimately outperforming both.

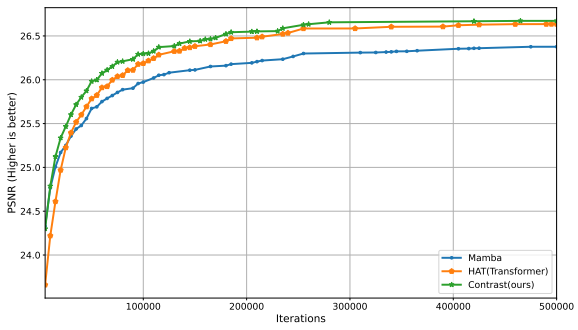

Figure 3. PSNR (dB) during training of pure Mamba [12], pure Transformer (HAT [3]), and our proposed Hybrid Contrast model. Results are evaluated on the Urban100 [17] dataset for $\times 4$ SR.

This training behavior is consistently observed across all five datasets: Set5, Set14, B100, Urban100, and Manga109. The hybrid model's ability to combine the rapid initial

learning of Mamba with the sustained improvement of the Transformer suggests that the integration of SS2D blocks with Transformer components effectively leverages the advantages of both architectures.

Impact of Channel Attention Blocks. We also investigated the role of Channel Attention Blocks (CABs) in our hybrid model. CABs were originally included in the baseline model, but we wanted to assess whether they are essential in our hybrid architecture or if removing them could lead to a more efficient model without compromising performance. To this end, we compared two versions of the hybrid model:

- With CAB: The hybrid model including CABs.
- Without CAB: The hybrid model without CABs, with additional layers and adjusted embedding dimensions to compensate for the removal and maintain an equal number of parameters.

The results, presented in Table 1, indicate that removing the CABs not only simplifies the model but also leads to improved PSNR values across all evaluated datasets. This suggests that the hybrid model benefits more from increased depth and embedding capacity than from the inclusion of CABs.

Table 1. Ablation study on the impact of Channel Attention Blocks (CABs). We compare models with and without CABs, adjusting the number of layers and embedding dimensions to keep the parameter count approximately equal. Results are reported in PSNR (dB) for $\times 4$ SR.

Model	Set5	Set14	B100	Urban100	Manga109
With CAB	32.44	28.78	27.68	26.57	31.06
Without CAB	**32.53**	**28.88**	**27.69**	**26.65**	**31.16**

Replacing MLP with SGFN. We further explored the effect of replacing the standard Multi-Layer Perceptron (MLP) layers in our model with the Spatial Gated Feed-forward Network (SGFN). The SGFN is designed to enhance the model's ability to capture spatial information by integrating gating mechanisms. We trained two versions of the model:

- With MLP: The hybrid model using standard MLP layers.
- With SGFN: The hybrid model with MLP layers replaced by SGFN.

The results, shown in Table 2, demonstrate that using SGFN leads to improved performance on most datasets.

Table 2. Comparison of models using MLP and SGFN layers. Results are reported in PSNR (dB) for $\times 4$ SR.

Model	Set5	Set14	B100	Urban100	Manga109
With MLP	32.44	28.84	27.73	26.72	31.22
With SGFN	**32.50**	**28.87**	27.73	**26.74**	**31.24**

LAM Visualization Analysis. To better understand how each model captures dependencies and spatial context, we utilize local attribution map (LAMs) [13]. Figure 4 presents the LAM visualizations for pure Mamba, pure Transformer, and our hybrid Contrast model.

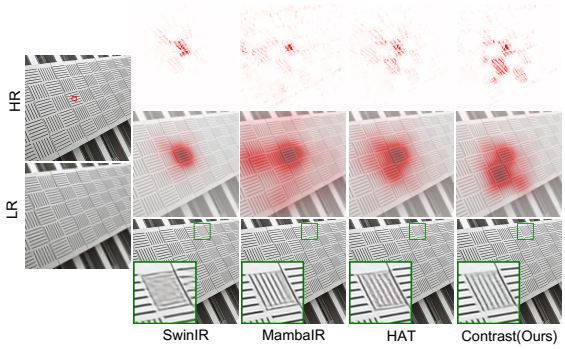

Figure 4. LAM visualizations and prediction results. The first and second rows show LAM visualizations for MambaIR [14], the Transformer model, and our hybrid Contrast model. The third row presents the corresponding SR predictions. The red square on the HR image marks the region used for LAM calculation. MambaIR exhibits a more global receptive field with a preference for features aligned horizontally and vertically, while the Transformer focuses more locally. Our Contrast model effectively combines these characteristics, capturing both global context and precise spatial details.

From the LAMs, we observe that MambaIR has a global receptive field, but with a tendency to emphasize horizontal and vertical features over diagonal ones. The Transformer model shows a more localized attention pattern. Our hybrid Contrast model successfully captures a broad receptive field while maintaining focus on important spatial regions, especially those with higher resolution or closer to the camera. This visualization underscores the advantage of our hybrid approach in balancing global context modeling and precise spatial attention, supporting our quantitative findings.

4.3. Comparison with State-of-the-Art Methods

We compare our proposed models, **Contrast-S** and **Contrast**, with several state-of-the-art (SOTA) image super-resolution methods. Consistent with prior studies [19, 45], we employ a self-ensemble strategy during testing, denoted by the symbol "+". The quantitative results are presented in Table 3, and visual comparisons are provided in Figure 5.

Quantitative Results. As shown in Table 3, our models achieve competitive performance, particularly excelling on the Urban100 [17] dataset. Our **Contrast** model attains a PSNR of 27.92 dB on Urban100, closely approaching the performance of HAT [3], which achieves 27.97 dB but with approximately 50% more parameters. Compared

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [20]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [45]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [7]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
RFANet [22]	×2	38.26	0.9615	34.16	0.9220	32.41	0.9026	33.33	0.9389	39.44	0.9783
HAN [30]	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
CSNLN [28]	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
NLSA [29]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
ELAN [44]	×2	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
DFSA [1]	×2	38.38	0.9620	34.33	0.9232	32.50	0.9036	33.66	0.9412	39.98	0.9798
SwinIR [19]	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
Swin2SR [6]	×2	38.43	0.9623	34.48	0.9256	32.54	0.9050	33.89	0.9431	39.88	0.9798
ART [43]	×2	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
CAT-A [4]	×2	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
DAT [5]	×2	38.58	0.9629	34.81	0.9272	32.61	0.9051	34.37	0.9458	40.33	0.9807
SRFormer [46]	×2	38.51	0.9627	34.44	0.9253	32.57	0.9046	34.09	0.9449	40.07	0.9802
HAT [3]	×2	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
MambaIR [14]	×2	38.57	0.9627	34.67	0.9261	32.58	0.9048	34.15	0.9446	40.28	0.9806
Contrast-S (ours)	×2	38.49	0.9624	34.58	0.9256	32.54	0.9041	34.06	0.9438	40.06	0.9802
Contrast (ours)	×2	38.58	0.9630	34.68	0.9261	32.57	0.9046	34.45	0.9467	40.30	0.9808
EDSR [20]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN [45]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [7]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
RFANet [22]	×3	34.79	0.9300	30.67	0.8487	29.34	0.8115	29.15	0.8720	34.59	0.9506
HAN [30]	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
CSNLN [28]	×3	34.74	0.9300	30.66	0.8482	29.33	0.8105	29.13	0.8712	34.45	0.9502
NLSA [29]	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
ELAN [44]	×3	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517
DFSA [1]	×3	34.92	0.9312	30.83	0.8507	29.42	0.8128	29.44	0.8761	35.07	0.9525
SwinIR [19]	×3	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
ART [43]	×3	35.07	0.9325	31.02	0.8541	29.51	0.8159	30.10	0.8871	35.39	0.9548
CAT-A [4]	×3	35.06	0.9326	31.04	0.8538	29.52	0.8160	30.12	0.8862	35.38	0.9546
DAT [5]	×3	35.16	0.9331	31.11	0.8550	29.55	0.8169	30.18	0.8886	35.59	0.9554
SRFormer [46]	×3	35.02	0.9323	30.94	0.8540	29.48	0.8156	30.04	0.8865	35.26	0.9543
HAT [3]	×3	35.07	0.9329	31.08	0.8555	29.54	0.8167	30.23	0.8896	35.53	0.9552
MambaIR [14]	×3	35.08	0.9323	30.99	0.8536	29.51	0.8157	29.93	0.8841	35.43	0.9546
Contrast-S (ours)	×3	35.02	0.9322	30.97	0.8541	29.49	0.8151	29.93	0.8838	35.25	0.9538
Contrast (ours)	×3	35.06	0.9324	31.00	0.8541	29.51	0.8158	30.17	0.8884	35.45	0.9549
EDSR [20]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [45]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [7]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
RFANet [22]	×4	32.66	0.9004	28.88	0.7894	27.79	0.7442	26.92	0.8112	31.41	0.918
HAN [30]	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
CSNLN [28]	×4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201
NLSA [29]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
ELAN [44]	×4	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
DFSA [1]	×4	32.79	0.9019	29.06	0.7922	27.87	0.7458	27.17	0.8163	31.88	0.9266
SwinIR [19]	×4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
Swin2SR [6]	×4	32.92	0.9039	29.06	0.7946	27.92	0.7505	27.51	0.8271	31.03	0.9256
ART [43]	×4	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
CAT-A [4]	×4	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
DAT [5]	×4	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343	32.51	0.9291
SRFormer [46]	×4	32.93	0.9041	29.08	0.7953	27.94	0.7502	27.68	0.8311	32.21	0.9271
HAT [3]	×4	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
MambaIR [14]	×4	33.03	0.9046	29.20	0.7961	27.98	0.7503	27.68	0.8287	32.32	0.9272
Contrast-S (ours)	×4	32.92	0.9038	29.11	0.7947	27.93	0.7489	27.59	0.8280	32.21	0.9264
Contrast (ours)	×4	32.94	0.9030	29.20	0.7973	27.98	0.7508	27.92	0.8357	32.38	0.9283
Contrast+ (ours)	×4	33.09	0.9055	29.23	0.7980	28.00	0.7520	28.07	0.8385	32.61	0.9299

Table 3. Quantitative comparison with state-of-the-art methods. The best and second-best results are coloured red and blue.

to MambaIR [14], which is based on the Mamba architecture we aimed to enhance, our **Contrast** model surpasses it by 0.24 dB on Urban100 while using fewer parameters. Our smaller model, **Contrast-S**, also performs comparably, trailing MambaIR by only 0.09 dB despite having nearly half the number of parameters.

It's important to note that our primary focus was on enhancing performance for datasets like Urban100, which contain high-resolution images with complex structures and repetitive patterns. These characteristics make Urban100 particularly challenging and ideal for models capable of capturing extensive global context. Our hybrid approach effectively addresses this, leveraging Mamba's global receptive field and the Transformer's precise contextual modeling to achieve superior results on Urban100.

On other datasets such as Set5, Set14, and B100, which consist of low-resolution images without significant patterns or contain low-quality photographs, our models do not outperform all existing methods. This outcome aligns with our expectations, as these datasets often favor deeper

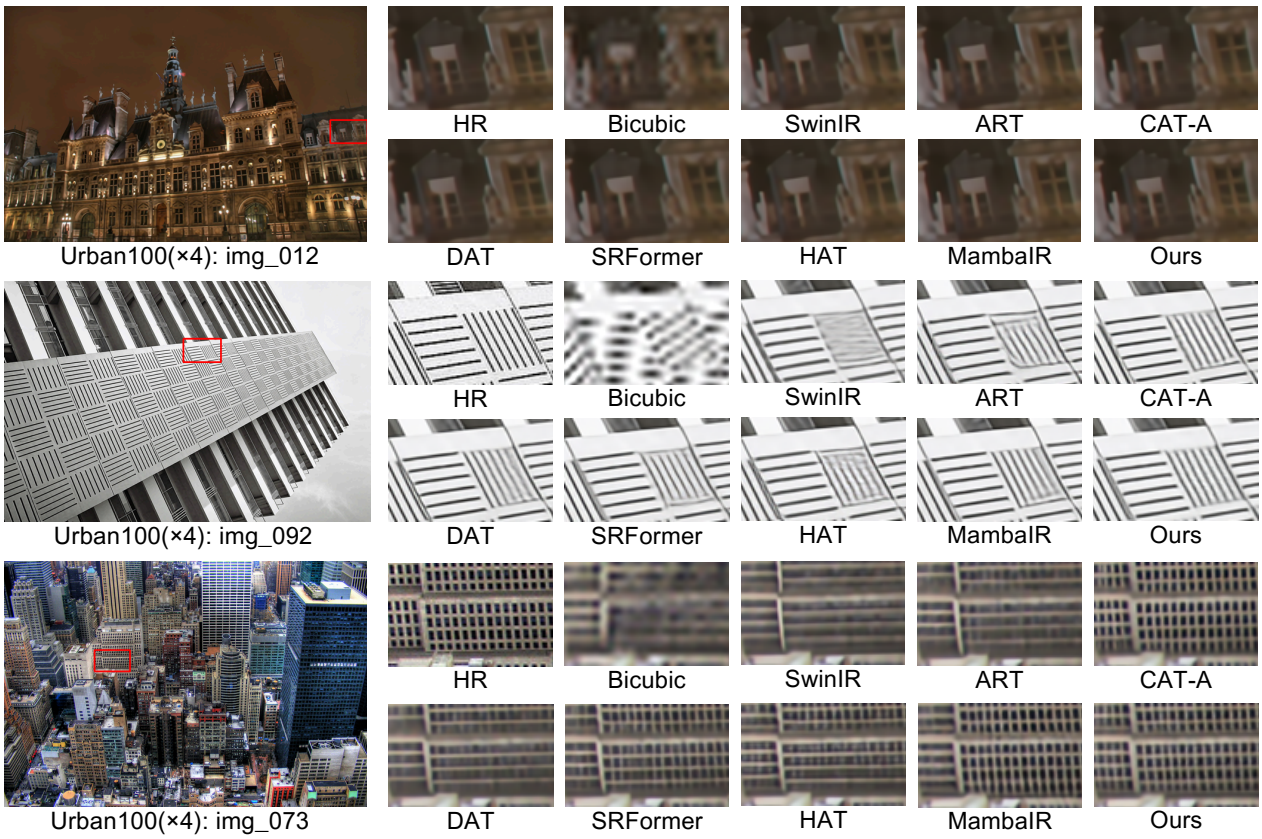

Figure 5. Visual comparison for $\times 4$ SR on the Urban100 dataset. Our Contrast model produces sharper and more detailed images, effectively reconstructing fine textures and patterns.

networks with larger parameter counts that can capture fine-grained details through extensive depth. Our goal was not to surpass SOTA results across all benchmarks but to demonstrate the effectiveness of our hybrid model in scenarios where capturing global context is crucial.

Visual Results. Figure 5 showcases visual comparisons on challenging samples from the Urban100 dataset. Our **Contrast** model effectively reconstructs intricate details and repetitive patterns, such as building facades and window grids, providing sharper and more detailed images than both MambaIR and other Transformer-based models.

5. Conclusion

In this paper, we proposed Contrast, a novel hybrid super-resolution model that effectively combines convolutional layers, Transformer blocks, and state space components. By integrating the global receptive field of Mamba with the precise contextual modeling capabilities of Transformers, our model addresses the limitations inherent in each individual approach.

Our competitive performance on the Urban100 dataset validates this hybrid approach, demonstrating that Contrast can effectively capture complex spatial patterns and deliver high-quality reconstructions, especially in images with repetitive structures and high-resolution details. While not leading on all benchmarks, our models offer a compelling trade-off between performance and model size, which is advantageous in practical applications where computational resources are limited and efficiency is crucial.

Future works could explore scaling our model and optimizing it for low-resolution images to improve performance on datasets like Set5 and Set14. By enhancing network depth, incorporating more advanced feature extraction techniques, and refining the balance between Transformer and Mamba components. Additionally, extending our approach to other low-level vision tasks could further demonstrate the versatility and effectiveness of hybrid architectures.

In summary, Contrast highlights the potential of hybrid architectures in advancing super-resolution performance where capturing extensive global context is essential. Our work opens new avenues for developing efficient and powerful super-resolution models that balance computational efficiency with high-quality image reconstruction, contributing to the ongoing progress in the field of low-level computer vision.

References

[1] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4288–4297, 2021. 7

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2012. 5

[3] 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 (CVPR)*, pages 22367–22377, 2023. 2, 3, 5, 6, 7

[4] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NeurIPS*, 2022. 5, 7

[5] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, 2023. 3, 4, 5, 7

[6] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2022. 7

[7] 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 (CVPR)*, 2019. 7

[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 (CVPR)*, pages 11065–11074, 2019. 1

[9] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024. 2, 5

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015. 1, 2

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

[12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 5

[13] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 6

[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. In *ECCV*, 2024. 2, 6, 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 1

[17] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 2, 5, 6

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

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 2, 6, 7

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, pages 1132–1140, 2017. 7

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

[22] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution, 2020. 7

[23] 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. 3

[24] Ze Liu, Yutong Lin, Yixuan 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 (ICCV)*, pages 10012–10022, 2021. 2

[25] 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. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 416–423, 2001. 5

[26] Yusuke Matsui, Kenji Ito, Yuki Aramaki, Aya Fujimoto, Toshihiko Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109

dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 5

[27] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5690–5699, 2020. 1

[28] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining, 2020. 7

[29] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Zhiwei Huang, Hao Tian, Humphrey Shi, Zhaowen Wu, and Thomas S. Huang. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3526, 2021. 7

[30] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network, 2020. 7

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

[32] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 4

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[34] Philippe Tillet, David Kung, Florent Png, Cliff Chou, and Nicolas Belanger. Triton: An intermediate language and compiler for tiled linear algebra. *arXiv preprint arXiv:1909.09788*, 2019. 5

[35] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Dataset and study. arXiv preprint arXiv:1708.08132, 2017. 5

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021. 1, 2

[37] Zizhao Tu, Hossein Talebi, Han Zhang, Feng Yang, Liang-Chieh Lau, Anna Goldenberg, Huisheng Mao, and Dmitry Kalenichenko. Maxvit: Multi-axis vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1, 2

[39] Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, Garvit Kulshreshtha, Vartika Singh, Jared Casper, Jan Kautz, Mohammad Shoeybi, and Bryan Catanzaro. An empirical study of mamba-based language models, 2024. 3, 4

[40] 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 *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, pages 63–79, 2018. 2

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

[42] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proceedings of the International Conference on Curves and Surfaces*, pages 711–730. Springer, 2010. 5

[43] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer, 2023. 7

[44] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 649–667, 2022. 7

[45] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 1, 2, 6, 7

[46] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12780–12791, 2023. 7