Fair and Faithful: A Diffusion-Enhanced Dataset and Hybrid State-Space Mamba for Face Super-Resolution

Tao Wang[1*], Peiwen Xia[2‡], Bowen Tang[3], Jinwei Chen[1], Kaihao Zhang[4], and Bi Li[1‡]

[1] vivo BlueImage Lab, vivo Mobile Communication Co., Ltd.
[2] Nanjing University
[3] Harbin Institute of Technology, Shenzhen
[4] Australian National University

Abstract. Face super-resolution (FSR) aims to enhance low-resolution facial images into high-resolution versions. Existing FSR methods face significant challenges, including the lack of publicly available and racially diverse datasets, which hinders reproducibility and subgroup-aware benchmarking. Additionally, many methods struggle to balance the restoration of fine-grained local details with the preservation of global facial structures, often producing results that are either geometrically inconsistent or lack realistic details. To address these issues, we introduce SFHQFSR, a new FSR dataset offering high-quality, racially diverse facial images generated from generative models, addressing challenges in reproducibility, fairness, and dataset construction. Building on this, we propose GLASNet, a hybrid state-space network that integrates global-local adaptive scanning with multi-domain refinement for high-quality facial image reconstruction. GLASNet combines global semantic reasoning with region-specific local modeling through a Region-Adaptive Scan Module and enhances restoration with a Multi-domain Refiner using channel, spatial, and frequency attention. Extensive experiments demonstrate that GLASNet achieves state-of-the-art performance.

Keywords: Face Super-Resolution · Face Dataset · Mamba · Benchmark Study

1 Introduction

Face super-resolution (FSR) aims to reconstruct high-resolution facial images from low-resolution inputs. Low-resolution (LR) facial images are often degraded by noise, motion blur, and poor quality, which hinder tasks like face recognition [54], attribute analysis [55], and detection [9]. Thus, FSR has received significant research attention in recent years. Early FSR methods [17, 37] mainly adopt facial priors (*e.g.,* facial key points, face heatmaps, or face parsing maps) to help the model restore high-resolution face images. However, the facial prior requires additional annotation efforts, which is subject to environmental factors.

With the success of deep learning, convolutional neural networks (CNNs) have become the foundation of many FSR methods [7, 32, 50, 59]. While they excel at enhancing texture details, their limited receptive fields hinder global facial structure modeling, often leading to inconsistent faces. Transformers, using self-attention [35], model

* Project lead (taowangzj@vivo.com).
‡ Corresponding authors (peiwenxia19@gmail.com, libra@vivo.com).

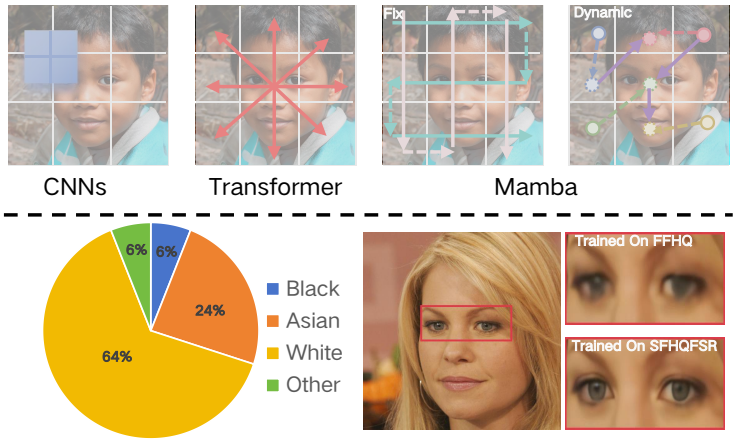

Fig. 1: Architectural Comparison and Dataset Analysis. Top: CNN (local grids), Transformer (global attention), Mamba (state-space). Bottom left: the details of the racial composition in our SFHQFSR dataset, which aims to improve racial diversity rather than enforce exact demographic parity. Bottom right: SFMNet [36] results trained on FFHQ vs. SFHQFSR. Our SFHQFSR enables the recovery of finer details and more visually pleasing outputs.

long-range dependencies but incur high quadratic computational costs. Mamba [12] introduced a state-space model (SSM) with linear complexity for global dependency modeling. As shown in Fig. 1, CNNs focus on local grids, Transformers use global attention, and Mamba uses state-space scanning. CNNs neglect global context, while Transformers introduce high computational complexity. Mamba, by modeling context through state transitions, improves efficiency. However, existing Mamba-based methods [6, 12, 28, 43, 58] use fixed scan paths, ignoring semantic variations across facial regions.

Another persistent challenge in the FSR community lies in dataset construction and evaluation. Recent methods [4, 16] rely on non-public synthetic datasets, where paired low- and high-resolution images are generated from existing datasets such as FFHQ [20], CelebA [19], or Helen [21], followed by random splits into training and testing sets. This practice makes it difficult to reproduce previous experiments because the data partitions are unreleased, and it forces subsequent works to re-synthesize datasets and retrain prior models for fair comparison, leading to unnecessary workload and poor reproducibility [4, 16, 31, 36]. Furthermore, most synthetic datasets still suffer from significant racial imbalance, overrepresenting certain groups while underrepresenting others, which introduces fairness and ethical concerns in FSR evaluation.

To address these limitations, we introduce SFHQFSR, a new FSR dataset offering high-quality, racially diverse facial images generated from the generative model [51] and corresponding low-resolution counterparts. As shown in the pie chart of Fig. 1, the dataset features diverse racial representation across multiple groups, although it is not designed to enforce exact demographic parity. Furthermore, we propose GLASNet, a

hybrid state-space network that combines global-local adaptive scanning with multi-domain refinement for high-fidelity FSR. The key idea is to integrate global semantic reasoning with region-adaptive local modeling. GLASNet incorporates a Region-Adaptive Scan Module to dynamically adjust scanning based on spatial content, capturing both global dependencies and fine details with linear complexity. Additionally, a Multi-domain Refiner, using channel, spatial, and frequency attention, enhances semantic identity, geometric structure, and spectral sharpness. Extensive experiments show GLASNet achieves state-of-the-art performance. To verify the advantage of SFHQFSR over FFHQ for FSR, we train three representative FSR methods, including SPARNet [4], SISN [31], and SFMNet [36] on both datasets and compare the results. With SFHQFSR, SFMNet [36] produces more faithful results, as shown in Fig. 1.

The contributions of this paper are four-fold: (1) We build a new dataset called SFHQFSR for FSR. SFHQFSR contains 70K face images generated from the generative model, which provide fixed training and testing groups, pairs of low-resolution and high-resolution images with different scale factors and degradation types (*e.g.,* BI $4\times$, BI $8\times$, BD $4\times$, and DN $4\times$). Based on SFHQFSR, we evaluate several state-of-the-art methods to better understand their potential and limitations. (2) We propose a novel framework called GLASNet, a hybrid state-space network that efficiently models both global context and local adaptivity for high-quality facial reconstruction. (3) In GLASNet, we design a Region-Adaptive State Space Module to dynamically capture region-specific structural and textural patterns. In addition, we develop a Multi-domain Refiner that harmonizes semantic, spatial, and frequency cues for balanced and perceptually realistic restoration. (4) Our GLASNet establishes new state-of-the-art FSR performance on standard evaluation metrics.

2 Related work

Face Datasets. The progress of FSR has been driven by high-quality facial datasets. Early datasets like LFW [15] and CelebA [29] focused on recognition, while high-resolution datasets like CelebA-HQ [19] and FFHQ [20] improved visual fidelity. Datasets such as FairFace [18] and EDFace-Celeb-1M [48] offer more balanced coverage across race, gender, and age. Synthetic datasets like SFHQ [1] provide high realism but lack attribute balance and paired supervision. Despite these advances, no unified open-source paired dataset for FSR exists. To address this, we introduce SFHQFSR, a dataset with paired high- and low-resolution images across various degradations.

FSR Methods. Early FSR methods, such as subspace [27] and component-based [3] models, struggled with large upscaling due to limited representational capacity. Deep learning methods, including CNN-based [56] and GAN-based [45], improved fidelity, while techniques like spatial transformers [46], reinforcement learning [2], and attention mechanisms [4] enhanced detail quality. Methods incorporating structural priors, such as CBN [59] and FSRNet [7], focus on preserving identity. Transformer and frequency-domain methods like SISN [31] and FaceFormer [40] capture global context and fine details. Despite these advances, most methods treat global and local cues separately, often compromising identity or texture sharpness. Our method addresses

this with a hybrid state-space formulation that models both global semantics and local adaptivity with linear complexity.

Vision Mamba. State-space models (SSMs) offer an efficient alternative to Transformers for visual sequence modeling. The original Mamba [12] introduced selective SSMs to capture long-range dependencies with linear complexity. Vision Mamba [58] and its variants, such as U-Mamba [33], MambaIR [13], and DAMamba [23], have shown strong performance in image classification, segmentation, and detection. These methods combine state evolution and input selection for efficient global context modeling. In image restoration, Mamba-inspired frameworks like MambaIR [13] and FMSR [42] balance efficiency and quality but often overlook spatial importance in identity-critical areas for FSR. To address this, we propose a hybrid state-space design that alternates between global and region-adaptive scanning, enabling fine-grained reconstruction while maintaining linear complexity.

3 Benchmark

This section outlines the multi-stage process used to build the SFHQFSR dataset. The pipeline includes 1) generating raw face images with ControlNet [51], 2) selecting high-quality images based on sharpness and face attribute classification, 3) manual filtering, and 4) synthesizing low-resolution (LR) face images. Using this pipeline, we obtain 70K paired images. Below are the key stages.

Generating the Face Images. Face images in existing datasets are typically collected from the Internet, which often raises privacy concerns and racial bias, while large-scale collection remains time-consuming and labor-intensive. To address these issues, we adopt ControlNet [51], which introduces spatial conditioning to Stable Diffusion, and use face parsing maps as structural guidance. These maps provide pixel-wise semantic information of facial components, enabling precise alignment and controllable synthesis. Specifically, we train ControlNet on 70K FFHQ and 27K CelebA-HQ training images, ensuring no overlap with the evaluation datasets. All images are resized to 512×512, and corresponding parsing maps are generated using a pretrained parsing network [5]. Specifically, the ControlNet [51] is trained for 200K iterations, and during inference, 50 sampling steps are used with parsing maps from a pretrained unconditional model. In this stage, we use diverse text prompts and varied parsing maps to generate 1M face images spanning multiple racial groups (*e.g.,* Asian, Black, White, Indian, Latino), two genders (male and female), and a wide range of ages. The pipeline of our high-quality face generation is shown in Fig. 2.

Selecting and Classifying. We first assess the sharpness of synthetic face images using Laplacian variance from OpenCV, retaining only those with values above 160. Since Laplacian variance alone cannot filter all distorted images, we further manually verify the quality of the remaining faces. To ensure accurate annotation, we employ a prediction model [18] trained on the FairFace dataset, which is balanced across race, gender, and age. This model classifies each image into four racial groups (Asian, Black, White, and Other, including Indian and Latino), two genders (male and female), and six age groups (0–9, 10–19, 20–29, 30–39, 40–49, and 50+ years). From each race–gender–age combination, we randomly select images for the training and testing

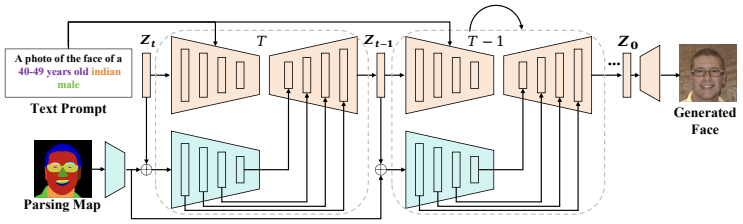

Fig. 2: The structure of the high-quality face generation process using ControlNet [51].

sets, resulting in 60K training images and 10K testing images. This process is intended to improve diversity, controllability, and subgroup awareness, rather than to enforce exact demographic parity.

Synthesize Pairs of Images. We follow the degradation simulation strategy of [47, 49,53]. Specifically, we define four degradation settings for face super-resolution (FSR): 4× BI, 4× BD, 4× DN, and 8× BI. The number denotes the downsampling factor, while "B" represents a blur operation, "D" indicates downsampling, and "N" stands for additive Gaussian noise. The order of letters reflects the sequence of operations. For example, "BD" means blur is applied before downsampling. Bicubic interpolation (BI) is used for the downsampling step. The Gaussian blur kernel is 7×7 with a standard deviation of 1.6, and Gaussian noise is added with $\sigma = 30$ to simulate realistic sensor noise.

To unify these settings, we formulate the degradation process as:

$$\mathbf{y} = \big((\mathbf{x} * \mathbf{k}) \downarrow_s^{\text{bicubic}}\big) + \mathbf{n}, \tag{1}$$

where $\mathbf{x}$ is the high-resolution face image, $\mathbf{y}$ is the low-resolution face image, $\mathbf{k}$ is the blur kernel (set to a delta function when no blur is applied), $s \in \{4, 8\}$ is the downsampling scale, and $\mathbf{n} \sim \mathcal{N}(0, \sigma^2)$ is additive Gaussian noise with $\sigma = 30$ when noise is included (otherwise $\mathbf{n} = \mathbf{0}$). We further use high-quality face images from CelebA [19] and Helen [21] to synthesize corresponding low-resolution counterparts according to the above degradation model, enabling comprehensive evaluation under diverse realistic conditions.

4 Methodology

4.1 Overview of GLASNet

Motivation. FSR aims to restore high-quality, identity-preserving facial images from degraded inputs. Existing methods face three challenges: 1) global-local inconsistency, where CNNs capture textures but miss long-range semantics, and global attention smooths details; 2) non-uniform importance, where critical regions like eyes and mouth are treated equally to others; and 3) multi-domain correlations, where semantic, spatial, and frequency features are handled separately, causing artifacts. To address these, we

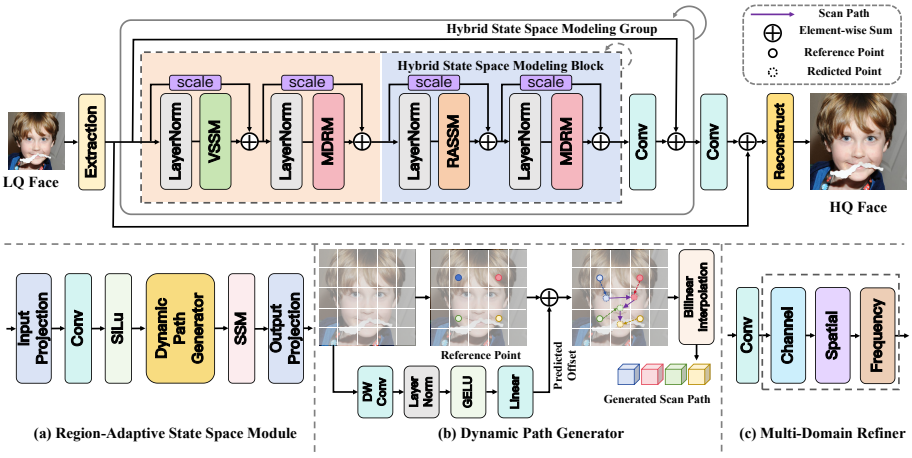

Fig. 3: Overview of our method. The framework includes multiple Hybrid State Space Modeling Groups, each with Hybrid State Space Modeling Blocks. Key modules: (a) Region-Adaptive State Space Module for adaptive modeling, (b) Dynamic Path Generator for creating adaptive scan paths, and (c) Multi-Domain Refinement Module enhancing facial representations across channel, spatial, and frequency domains.

introduce GLASNet, a Global-Local Adaptive State-space Network that restores faces through a progressive semantic-to-detail reasoning process.

Overall Pipeline. The GLASNet pipeline, shown in Fig. 3, consists of three stages: shallow feature extraction, deep feature learning, and high-quality reconstruction. Given a low-quality input image $\mathbf{I}$, shallow representations $\mathbf{F}_0$ are extracted using a 3×3 convolution layer. These are then passed through Hybrid State Space Modeling Groups (HSSMGs) to capture hierarchical deep features, with each group consisting of Hybrid State Space Modeling Blocks (HSSMBs) that perform global scanning via the Visual State Space Module (VSSM) and local adaptive scanning via the Region-Adaptive State Space Module (RASSM). The Multi-Domain Refiner (MDR) further refines facial representations. Four residual connections with trainable scaling factors are used in the HSSMBs. Finally, $\mathbf{F}_L$ is processed through a convolution layer, added to $\mathbf{F}_0$, and passed through pixel shuffle layers to reconstruct the super-resolved image $\hat{\mathbf{I}}$.

4.2 Hybrid State Space Modeling

Preliminaries. Recent state-space-based image restoration models are inspired by the Mamba [12] framework, which expresses token relationships using a discrete state-space formulation:

$$\begin{aligned} h_i &= \overline{\mathbf{A}} h_{i-1} + \overline{\mathbf{B}} x_i, \\ y_i &= \mathbf{C} h_i + \mathbf{D} x_i, \end{aligned} \tag{2}$$

where $\overline{\mathbf{A}} = \mathbf{exp}(\Delta\mathbf{A})$ controls information propagation, $\overline{\mathbf{B}} \approx \Delta\mathbf{B}$ represents input transformation, and $\mathbf{C}$ is the output projection matrix. Unlike attention mechanisms

with costly pairwise interactions, this approach captures long-range dependencies with linear complexity, making it suitable for high-fidelity super-resolution tasks that require modeling spatial dependencies across large feature maps.

Hybrid State-Space Modeling. Previous Mamba-based methods [13, 22, 42] rely on a single global scanning path to capture long-range dependencies. While effective for general image modeling, this uniform scanning treats all spatial locations equally, neglecting the structural hierarchy in human faces. In FSR, facial components like eyes, nose, and mouth carry richer identity information than surrounding regions, and equal attention to all areas wastes computation and struggles to recover fine details.

To address this, we propose the Hybrid State-Space Modeling mechanism, combining global and region-adaptive reasoning in a unified framework. As shown in Fig. 3, each Hybrid State-Space Modeling Block (HSSMB) processes features through Layer Normalization, State-Space Model layer, and Multi-domain Refiner, applied twice to enhance global semantics and local structures. Residual connections with learnable scaling factors stabilize gradients and balance residual feature contributions. The HSSMB is expressed as,

$$\begin{aligned}
\mathbf{F}_s^g &= \mathbf{VSSM}(\mathbf{LN}(\mathbf{F}_{in})) + \alpha_g \mathbf{F}_{in}, \\
\mathbf{F}_r^g &= \mathbf{MDR}(\mathbf{LN}(\mathbf{F}_s^g)) + \beta_g \mathbf{F}_s^g, \\
\mathbf{F}_s^a &= \mathbf{RASSM}(\mathbf{LN}(\mathbf{F}_r^g)) + \alpha_s \mathbf{F}_r^g, \\
\mathbf{F}_r^a &= \mathbf{MDR}(\mathbf{LN}(\mathbf{F}_s^a)) + \beta_s \mathbf{F}_s^a,
\end{aligned} \tag{3}$$

where $\mathbf{F}_{in}$ is the input feature map, $\mathbf{VSSM}$ and $\mathbf{RASSM}$ are the visual and region-adaptive state-space modules, $\mathbf{MDR}$ is the multi-domain refiner, and α, β are learnable scaling factors for residual blending.

In the first stage, the Visual State-Space Module (VSSM) [28] performs global feature propagation to capture long-range semantic dependencies across the face. It models contextual continuity via a dual-branch interaction,

$$\begin{aligned}
\mathbf{X}_1 &= \mathbf{SiLU}(\mathbf{Linear}(\mathbf{X}_{in})), \\
\mathbf{X}_2 &= \mathbf{LN}(\mathbf{2DSSM}(\mathbf{SiLU}(\mathbf{DWConv}(\mathbf{Linear}(\mathbf{X}_{in}))))), \\
\mathbf{X}_{out} &= \mathbf{Linear}(\mathbf{X}_1 \odot \mathbf{X}_2),
\end{aligned} \tag{4}$$

where $\odot$ is element-wise multiplication, and $\mathbf{DWConv}$ refers to depth-wise convolution. The first branch encodes semantic projection, while the second evolves spatial context through two-dimensional state-space propagation. Their fusion filters redundant responses, producing globally coherent features.

In the second stage, the Region-Adaptive State-Space Module (RASSM) adjusts its propagation path based on local feature complexity. RASSM allocates more capacity to regions with dense identity cues, and less to homogeneous or background areas. This hybrid reasoning mechanism ensures a balance between global semantic integrity and local perceptual fidelity, forming the core reasoning unit of GLASNet. The formulation of RASSM is detailed in the following subsection.

4.3 Region-Adaptive Scan

Motivation. Existing State Space Models [6, 12, 28, 43, 58] use fixed global scanning paths. However, image regions vary in information density, especially in face super-

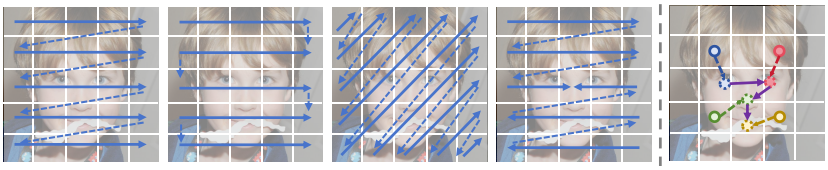

Fig. 4: Comparison between previous scans and our proposed scan. Unlike previous methods [6, 12, 28, 43, 58] with fixed scanning paths, our method uses a dynamic scanning path.

resolution, where critical areas like the eyes, nose, and mouth contain richer structural cues than the background. A uniform scan treats all regions equally, weakening the representation of important features and degrading reconstruction quality.

Region-Adaptive Scan. To overcome this, we propose the Region-Adaptive Scan. As illustrated in Fig. 4, it aims to enhance scanning adaptability by dynamically adjusting the scanning paths according to region content importance. The specific steps of Region-Adaptive Scan are as follows: (1) Given the input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times \lambda C}$, the importance of each patch is estimated, and spatial offsets $\Delta\mathbf{p} \in \mathbb{R}^{H\times W\times 2}$ are predicted to indicate how patches should shift toward informative regions,

$$\begin{aligned} \mathbf{Z} &= \mathbf{GELU}(\mathbf{LN}(\mathbf{DWConv}(\mathbf{X}))), \\ \Delta\mathbf{p} &= \mathbf{Linear}(\mathbf{Z}). \end{aligned} \tag{5}$$

(2) These offsets Δp are then processed by DCNv3 [38], and added to the original 2D patch coordinates p to generate content-adaptive target coordinates $\mathbf{p}' = \mathbf{p} + \Delta\mathbf{p}$, with the coordinates normalized to $[-1, 1]$.

(3) Bilinear interpolation is subsequently employed to extract features from input $\mathbf{X}$ at the target positions, forming a dynamic scanning path $\mathbf{X}'$ that emphasizes critical facial structures,

$$\begin{aligned} \mathbf{X}'[\mathbf{h}, \mathbf{w}] &= \mathcal{B}(\mathbf{p}'[\mathbf{h}, \mathbf{w}], \mathbf{X}), \\ \mathcal{B}(\mathbf{p}, \mathbf{X}) &= \sum_{r\in\mathcal{N}(p)} w(\mathbf{p}, r)\mathbf{X}[r_x, r_y, :], \end{aligned} \tag{6}$$

where $\mathcal{N}(p)$ denotes the four nearest neighbors of $\mathbf{p}$, and $w(p, r)$ is the corresponding bilinear interpolation weight.

(4) The generated path $\mathbf{X}'$ is fed into the SSM for long-range dependency modeling.

Region-Adaptive State Space Module. Based on the proposed Region-Adaptive Scan, we introduce the RASSM. As shown in Fig. 3 (a), RASSM comprises a sequence of convolutional and state-space operations, with the Dynamic Path Generator (DPG) guiding regional sampling.

During the RASSM, the input $\mathbf{X}_{in} \in \mathbb{R}^{H\times W\times C}$ is first projected into a latent space with λC channels, and processed with a convolution followed by a SiLU activation to enhance local representations,

$$\mathbf{X} = \mathbf{SiLU}(\mathbf{Conv}(\mathbf{Proj_{in}}(\mathbf{X}_{in}))), \tag{7}$$

Table 1: Quantitative comparison under bicubic degradation (×4/×8). SISR: general single-image SR; FSR: face SR. The 1st-place results are bolded, and the 2nd-place results are underlined.

Type	Method	Deg	SFHQ			CelebA			Helen		
			PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Bicubic Degradation (BI), Scale ×4											
SISR	RCAN [52]	BI	33.50	0.9033	0.2130	32.18	0.8618	0.2021	28.21	0.8471	0.1674
	EDSR [25]	BI	33.58	0.9041	0.2142	32.28	0.8622	0.2012	28.31	0.8534	0.1674
	SwinIR [24]	BI	33.60	0.9043	0.2112	32.30	0.8699	0.2190	28.41	0.8661	0.1692
	PFT [30]	BI	33.68	0.9054	0.2001	32.33	0.8671	0.2161	28.52	0.8682	0.1662
FSR	WaveletSR [16]	BI	32.94	0.9017	0.2015	31.00	0.8493	0.2098	28.90	0.8688	0.1738
	SPARNet [4]	BI	33.38	0.9010	0.2276	31.84	0.8597	0.2181	28.49	0.8663	0.2048
	HiFaceGAN [44]	BI	32.06	0.8689	**0.1418**	30.66	0.8245	0.2284	28.19	0.8470	0.1653
	SISN [31]	BI	33.58	0.9004	0.2264	32.35	0.8704	0.2087	29.43	0.8864	0.1773
	SFMNet [36]	BI	34.42	0.9132	0.2019	32.39	0.8701	0.2003	29.50	0.8864	0.1711
	Ours	BI	**34.70**	**0.9158**	0.1999	**32.51**	**0.8725**	**0.1993**	**29.57**	**0.8878**	**0.1614**
Bicubic Degradation (BI), Scale ×8											
SISR	RCAN [52]	BI	28.65	0.8315	0.3312	28.33	0.7812	0.3255	23.48	0.6433	0.3844
	EDSR [25]	BI	28.80	0.8340	0.3385	28.39	0.7820	0.3225	23.56	0.6871	0.3824
	SwinIR [24]	BI	28.86	0.8351	0.3398	28.44	0.7837	0.3203	23.63	0.6904	0.3779
	PFT [30]	BI	28.96	0.8384	0.3381	28.67	0.7845	0.3123	23.83	0.6964	0.3743
FSR	WaveletSR [16]	BI	28.13	0.8252	0.3180	27.70	0.7693	0.3175	24.04	0.7015	0.3575
	SPARNet [4]	BI	30.48	0.8403	0.3233	29.11	0.7865	0.3203	24.09	0.7098	0.3812
	HiFaceGAN [44]	BI	27.95	0.7679	**0.2451**	26.98	0.7108	0.3138	23.30	0.6459	0.3684
	SISN [31]	BI	30.11	0.8313	0.3386	29.04	0.7841	0.3249	24.07	0.7060	0.3772
	SFMNet [36]	BI	30.64	0.8413	0.3224	29.28	0.7905	0.3124	24.34	0.7211	0.3676
	Ours	BI	**30.86**	**0.8451**	0.3200	**29.44**	**0.7938**	**0.3099**	**24.58**	**0.7274**	**0.3543**

where $\mathbf{Proj}_{in}(\cdot)$ denotes input projection. Next, the DPG predicts the dynamic scanning path $\mathbf{X}'$,

$$\mathbf{X}' = \mathbf{DPG}(\mathbf{X}). \tag{8}$$

Following this, the adaptively sampled features X' are arranged in the original patch order (top-to-bottom, left-to-right) and fed into the SSM for feature extraction. Finally, a projection produces the module output,

$$\mathbf{X}_{out} = \mathbf{Proj_{out}}(\mathbf{SSM}(\mathbf{X}')), \tag{9}$$

where $\mathbf{Proj}_{out}(\cdot)$ denotes output projection.

4.4 Multi-domain Refiner

The core motivation. Faithful face super-resolution requires joint reconstruction of semantic identity, geometric structure, and frequency content. Conventional methods fail to coordinate these cues, often amplifying textures at the cost of identity coherence [39, 57] or suppressing high-frequency details like hair strands [24]. This limitation arises from treating attention as a single operation without disentangling the roles of channels, spatial regions, and spectral bands. Our Multi-domain Refiner addresses this by modeling channel-wise semantics, spatial geometry, and spectral characteristics in a unified framework.

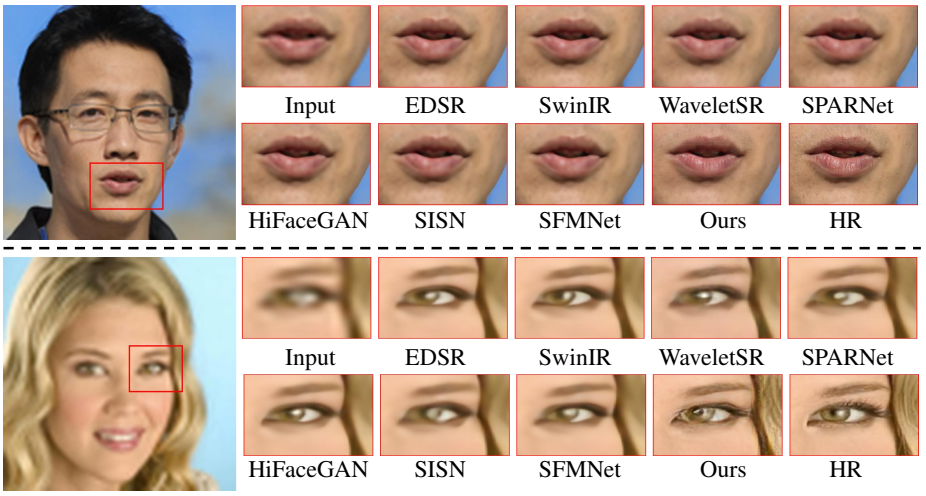

Fig. 5: Visual Comparison. The top results are tested on the SFHQFSR dataset with $\times 4$ bicubic degradation, and the bottom results are on the CelebA-test dataset with $\times 8$ bicubic degradation.

Multi-domain Refiner. Based on this, we design a refiner that recovers structural and spectral details by refining across semantic, spatial, and frequency domains. Specifically, the input features are first fused by a lightweight convolution and then refined through three attention branches: channel attention (CA) [14] for identity-related semantics, spatial attention (SA) [41] for key geometric regions, and frequency attention (FA) [8] for high-frequency details. The process is expressed as,

$$\mathbf{X}_{out} = \mathbf{FA}(\mathbf{SA}(\mathbf{CA}(\mathbf{Conv}(\mathbf{X}_{in})))), \tag{10}$$

where the channel attention branch highlights identity-related semantics, the spatial attention branch enhances geometric alignment around key landmarks, and the frequency attention branch restores perceptual sharpness by amplifying high-frequency textures while suppressing artifacts.

5 Experiments

5.1 Experimental Settings

Implementation details. Our network consists of 6 HSSMGs, each containing 6 HSSMBs. The input and output channel dimension in all blocks is set to 180. For both VSSM and RASSM, the channel scaling factor λ is fixed to 1. We adopt the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 2×10^{-4} and decayed by a factor of 0.5 when the training iteration reaches specific milestones. For a fair comparison, the batch size is set to 1 in all experiments. The model is trained for a total of 800k iterations. Data augmentation is applied by randomly performing horizontal flips

Table 2: Quantitative comparison under complex degradations (BD $\times 4$, DN $\times 4$). SISR: general single-image SR; FSR: face SR. The 1st-place results are bolded, and the 2nd-place results are underlined.

Type	Method	Deg	SFHQ			CelebA			Helen		
			PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Blur-downscale Degradation (BD), Scale $\times 4$											
SISR	RCAN [52]	BD	33.46	0.9031	0.2235	31.47	0.8638	0.2197	28.72	0.8832	0.1765
	EDSR [25]	BD	33.52	0.9012	0.2237	31.46	0.8635	0.2199	28.75	0.8840	0.1769
	SwinIR [24]	BD	33.63	0.9062	0.2219	31.46	0.8636	0.2198	28.78	0.8849	0.1792
	PFT [30]	BD	33.74	0.9073	0.2031	31.67	0.8645	0.2178	28.98	0.8869	0.1782
FSR	WaveletSR [16]	BD	33.14	0.9046	0.2033	31.23	0.8534	0.2047	28.90	0.8716	0.1723
	SPARNet [4]	BD	33.47	0.9045	0.2237	31.91	0.8644	0.2175	28.37	0.8699	0.2056
	HiFaceGAN [44]	BD	31.80	0.8721	**0.1419**	30.57	0.8307	0.2246	28.04	0.8507	**0.1664**
	SISN [31]	BD	33.54	0.9016	0.2261	32.32	0.8715	0.2083	29.29	0.8862	0.1767
	SFMNet [36]	BD	34.31	0.9136	0.2021	32.37	0.8723	0.2011	29.42	0.8880	0.1756
	Ours	BD	**34.66**	**0.9164**	0.2008	**32.50**	**0.8741**	**0.1988**	**29.78**	**0.8944**	0.1691
Noise Degradation (DN), Scale $\times 4$											
SISR	RCAN [52]	DN	28.19	0.8010	0.3996	27.22	0.7625	0.3998	23.48	0.6879	0.4326
	EDSR [25]	DN	28.17	0.8006	0.4010	27.20	0.7630	0.3923	23.46	0.6865	0.4399
	SwinIR [24]	DN	28.26	0.8026	0.3989	27.26	0.7642	0.3918	23.49	0.6880	0.4390
	PFT [30]	DN	28.37	0.8039	0.3904	27.32	0.7657	0.3901	23.71	0.6883	0.4384
FSR	WaveletSR [16]	DN	28.86	0.8110	0.3883	27.88	0.7579	0.3759	24.29	0.7126	**0.3865**
	SPARNet [4]	DN	28.73	0.8129	0.3899	27.72	0.7583	0.3812	23.71	0.6980	0.4227
	HiFaceGAN [44]	DN	26.62	0.7403	**0.3070**	25.67	0.6832	0.3672	22.93	0.6363	0.4290
	SISN [31]	DN	28.53	0.8075	0.4066	27.77	0.7613	0.3851	24.12	0.7143	0.4028
	SFMNet [36]	DN	29.04	0.8183	0.3766	28.06	0.7696	0.3679	24.33	0.7222	0.3909
	Ours	DN	**29.30**	**0.8234**	0.3749	**28.30**	**0.7754**	**0.3582**	**24.50**	**0.7292**	0.4004

and $90°$, $180°$, and $270°$ rotations on the training images. For benchmarking, we compare 8 representative methods including 3 SISR methods and 5 FSR methods. For each method, we use the publicly available code and train each learning-based method for 800k iterations. For fairness, all baselines are retrained on SFHQFSR using the same training iterations, instead of using pretrained weights. All experiments are conducted on four NVIDIA A100 GPUs.

5.2 Benchmark Experiments

We evaluate existing methods and our GLASNet on SFHQFSR under multiple degradation settings: Bicubic (BI), Blur-downscale (BD), and Downsampling and Noise (DN), with two upscaling factors ($\times 4$ and $\times 8$). BI represents the classical FSR degradation setting, while BD and DN simulate more realistic degradations.

Comparison in Bicubic Degradation. We compare GLASNet with representative SISR methods (RCAN [52], EDSR [25], SwinIR [24], PFT [30]) and FSR methods (WaveletSR [16], SPARNet [4], HiFaceGAN [44], SISN [31], SFMNet [36]) under bicubic degradation at $\times 4$ and $\times 8$ scales. The results are shown in Table 1. At $\times 4$ setting, GLASNet achieves the highest PSNR and SSIM on all datasets, outperforming the best prior FSR method (SFMNet [36]) by up to 0.28 dB in PSNR and 0.0026 in SSIM. It also obtains the lowest LPIPS on CelebA-test and Helen. General SISR models lag behind, particularly in preserving facial structure. At $\times 8$ setting, GLASNet remains

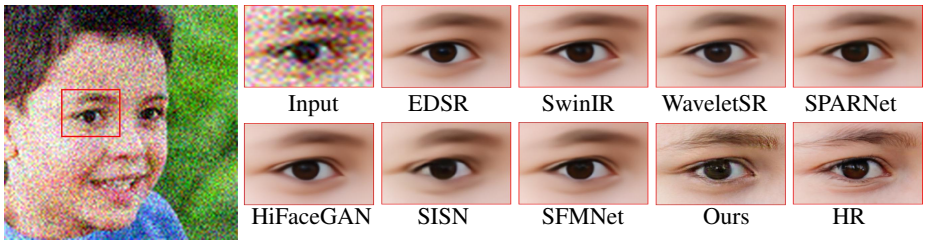

Fig. 6: Visual comparison with existing methods on our SFHQFSR dataset with $\times 4$ noise degradation setting.

superior across all metrics, improving upon SFMNet [36] by 0.22 dB in PSNR and achieving the lowest LPIPS on CelebA-test and Helen.

Fig. 5 shows results under $\times 4$ bicubic degradation. GLASNet produces sharper lip contours and natural skin texture, while SISR methods like EDSR [25] and SwinIR [24] produce overly smooth outputs. FSR methods exhibit artifacts: HiFaceGAN [44] and SISN [31] introduce unnatural gloss, SFMNet [36] shows slight lip misalignment, and others blur fine details. GLASNet preserves accurate geometry and consistent shading. As for $\times 8$ setting (see the bottom in Fig. 5), competing methods degrade significantly. For example, EDSR [25] and SwinIR [24] lose eyeliner and iris details, HiFaceGAN [44] distorts eye shape, and SFMNet [36] blurs eyelid structures. Our GLASNet retains crisp eyeliner, well defined irises, and realistic skin tone.

Comparison in Complex Degradation. We evaluate GLASNet under BD and DN degradations with an upscaling factor of $\times 4$. Quantitative results are shown in Table 2. Under BD setting, GLASNet achieves the highest PSNR and SSIM on all three datasets. It outperforms the best prior FSR method, SFMNet, by 0.35 dB in PSNR and 0.0028 in SSIM on Helen, and obtains the lowest LPIPS on CelebA-test. As for DN setting, GLASNet again leads in PSNR and SSIM across all benchmarks, with a 0.26 dB gain over SFMNet on SFHQFSR. It also achieves the best LPIPS on CelebA-test. While HiFaceGAN shows strong LPIPS under BD and DN, its PSNR and SSIM are significantly lower, indicating a fidelity-perception trade-off. GLASNet maintains high performance in both aspects. In Fig. 6, we compare reconstructions under $\times 4$ noise degradation. EDSR [25] and SwinIR [24] produce overly smooth outputs, losing fine details. WaveletSR [16] and SPARNet [4] introduce blurring, while HiFaceGAN [44] causes color shifts and halo artifacts. SISN [31] and SFMNet [36] preserve geometry but soften details. Our GLASNet restores sharp eyeliner, defined iris boundaries, and natural skin tone.

5.3 Comparison with FFHQ

FFHQ [20] is a widely used high-quality human face dataset commonly employed for FSR. To compare the quality of SFHQFSR with FFHQ, we randomly select 60K images from FFHQ, apply the same BI $\times 8$ degradation used in SFHQFSR, and generate low-resolution inputs. We then train three representative FSR methods, including SPAR-

Table 3: FSR Performance comparison on CelebA-test and Helen datasets. Average PSNR/SSIM values for BI×8 are reported.

Methods	Training Dataset	CelebA-test		Helen	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑
SPARNet [4]	FFHQ	29.09	0.7838	23.83	0.7083
	SFHQFSR	**29.11**	**0.7865**	**24.09**	**0.7098**
SISN [31]	FFHQ	28.80	0.7835	23.91	0.7009
	SFHQFSR	**29.04**	**0.7841**	**24.07**	**0.7060**
SFMNet [36]	FFHQ	28.90	0.7819	24.02	0.7075
	SFHQFSR	**29.28**	**0.7905**	**24.34**	**0.7211**

Table 4: Ablation study of scan strategy in the network.

Models	Global Scan	Region-adaptive Scan	PSNR↑	SSIM↑
Base	✓	✗	30.13	0.8319
Variant 1	✗	✓	30.41	0.8425
Variant 2	✓	✓	30.73	0.8431
Ours	✓	✓	**30.86**	**0.8451**

Net [4], SISN [31], and SFMNet [36], on both FFHQ and SFHQFSR datasets under same settings. The models are evaluated on the CelebA-test and Helen datasets to ensure a fair comparison. As shown in Table 3, models trained on SFHQFSR achieve comparable or slightly better performance than those trained on FFHQ. In particular, SFMNet trained on SFHQFSR attains the highest PSNR and SSIM scores on both benchmarks, indicating that SFHQFSR provides higher-quality and more diverse training data for facial detail reconstruction. This improvement can be attributed to the broader variations in facial pose, expression, illumination, and texture present in SFHQFSR, which enhance the model's generalization ability to diverse conditions. In summary, the quality of SFHQFSR is comparable to that of FFHQ, demonstrating its reliability and effectiveness as a training source for FSR.

5.4 Discussion

In this section, we provide further analysis of GLASNet and SFHQFSR. We first study the effects of different architectural components through ablation experiments, then compare GLASNet with recent state-of-the-art methods in terms of accuracy, perceptual quality, identity preservation, and efficiency.

Effects of scan strategy in the network. We conduct ablation studies to investigate the impact of the region-adaptive scan strategy in GLASNet. As shown in Table 4, we first evaluate the global scan (Base) and then the region-adaptive scan (Variant 1) to assess their individual contributions. The results show that the region-adaptive scan significantly improves reconstruction quality by better capturing local spatial dependencies. Additionally, combining both global and region-adaptive scans further boosts performance, as the global scan provides overall context, while the region-adaptive

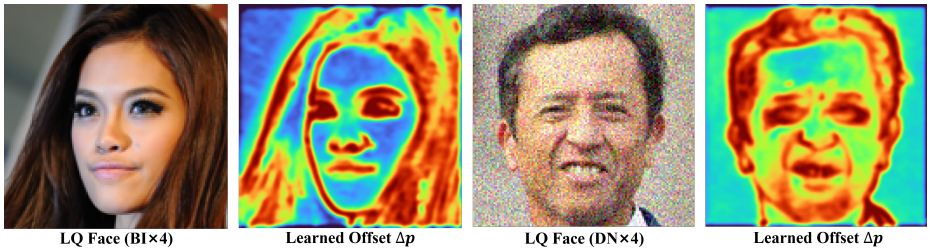

Fig. 7: Visualization of the learned offset Δp in Region-adaptive Scan.

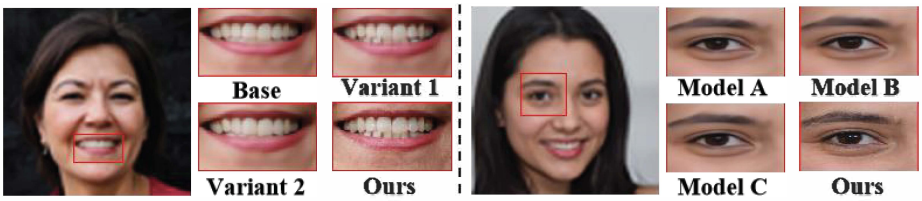

Fig. 8: Ablation comparison results.

scan refines local details. We also examine the effect of scan sequence. Variant 2 applies the region-adaptive scan before the global scan, while our final model applies the global scan first, followed by the region-adaptive scan. This sequence allows the network to capture global context before refining local regions, leading to higher PSNR and SSIM values. To further examine the effectiveness of Region-adaptive Scan, we visualize the learned offsets in Fig. 7. The results show that the offsets are clearly non-uniform and mainly concentrate on structure-rich and high-frequency regions, *e.g.,* the eyes, nose, mouth, facial contours, and background boundaries, while smoother skin regions receive weaker responses. This demonstrates that the Region-adaptive Scan learns content-adaptive paths rather than following a fixed raster-order traversal.

Importance of designs in the multi-domain refiner. We evaluate the contributions of channel attention (CA), spatial attention (SA), and frequency attention (FA) in our multi-domain refiner. As shown in Table 5, the baseline (Model A) with only convolutions achieves 28.30 dB PSNR and 0.8258 SSIM. Adding CA (Model B) improves PSNR to 29.92, showing its benefit for facial feature preservation. Further introducing SA (Model C) slightly increases PSNR to 30.20 but reduces SSIM, suggesting a trade-off in structural refinement. The full model with CA, SA, and FA (Model D) yields the best performance at 30.86 dB PSNR and 0.8451 SSIM, confirming that FA effectively restores high-frequency details. The ablation comparison results are shown in Fig. 8.

Table 5: Ablation study of multi-domain refiner. CA, SA, and FA refer to channel attention, spatial attention, and frequency attention used in the multi-domain refiner.

Models	Conv	CA	SA	FA	PSNR↑	SSIM↑
A	✓	✗	✗	✗	28.30	0.8258
B	✓	✓	✗	✗	29.92	0.8312
C	✓	✓	✓	✗	30.20	0.8214
D	✓	✓	✓	✓	**30.86**	**0.8451**

Effects of the number of HSSMG in the network. We conduct an ablation study to examine the impact of varying the number of HSSMG modules in the network. In Table 6, we evaluate models with 2, 4, 6, and 8 HSSMG modules. The results show that the model with 6 HSSMG modules (HSSMG-6) achieves the best performance, with 30.86 dB PSNR and 0.8451 SSIM, offering the best trade-off between performance and model complexity. While increasing the number of modules beyond 6 yields some performance improvement, it significantly increases the model's parameter count (to 18.34M for HSSMG-8), leading to diminishing returns in terms of efficiency. Thus, HSSMG-6 (ours) achieves a better trade-off between performance and complexity.

Table 6: Ablation study of the number of HSSMG in the network.

Models	HSSMG-2	HSSMG-4	**HSSMG-6**	HSSMG-8
PSNR↑	30.18	30.40	30.86	30.90
SSIM↑	0.8363	0.8440	0.8451	0.8459
Parameter↓	5.22M	9.60M	13.87M	18.34M

Efficiency, perceptual and identity-preservation comparison. We compare our method with recent 2024–2025 state-of-the-art (SOTA) techniques, including two Single Image Super-Resolution (SISR) methods, one Video Frame Reconstruction (VFR) method, and two Frame Super-Resolution (FSR) methods. These methods represent Transformer-based, Diffusion-based, and Mamba-based approaches. All methods are evaluated in the most challenging setting, DN $\times$4. As shown in Table 7, GLASNet outperforms other methods in terms of PSNR/SSIM, LPIPS, FID, IDS, and speed. Specifically, on the Helen DN$\times$4 benchmark, GLASNet achieves the highest PSNR (24.50), SSIM (0.7292), and the lowest LPIPS (0.4004), as well as the lowest FID (54.11) and the highest IDS (0.79). These results demonstrate that GLASNet excels in both perceptual quality and identity preservation, surpassing traditional PSNR/SSIM/LPIPS metrics. Regarding efficiency, GLASNet is the fastest among the methods tested, processing each image in 0.1426 seconds, while still maintaining the highest quality. This performance is driven by our innovative linear-complexity state-space modeling and global-local adaptive scanning technique.

Table 7: Comparison of recent SOTA methods on Helen (DN$\times$4). Speed is reported in sec/image.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	IDS↑	Speed ↓
KEEP [11]	23.45	0.6866	0.4312	89.12	0.32	0.1835
VmambaIR [34]	24.27	0.7193	0.4157	60.12	0.63	0.1674
PFT [30]	23.94	0.7112	0.4323	56.49	0.61	0.4313
DiT4SR [10]	23.48	0.6980	0.4219	55.89	0.41	54.2835
FaceMe [26]	24.02	0.7103	0.4211	55.32	0.69	68.9766
GLASNet	**24.50**	**0.7292**	**0.4004**	**54.11**	**0.79**	**0.1426**

6 Conclusion

In this work, we first build SFHQFSR, a new FSR dataset that is racially diverse and addresses critical challenges in reproducibility and fairness. With fixed training/testing splits and multiple degradation settings, SFHQFSR provides a practical benchmark for evaluating FSR methods under diverse restoration conditions. Then, we propose GLASNet, a hybrid state-space network designed to effectively capture both global context and fine local details for high-quality facial image restoration. Extensive experiments show that GLASNet achieves state-of-the-art performance. We hope our SFHQFSR with GLASNet will benefit the community.

References

1. Beniaguev, D.: Synthetic faces high quality (sfhq) dataset (2022)
2. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: CVPR. pp. 690–698 (2017)
3. Chakrabarti, A., Rajagopalan, A., Chellappa, R.: Super-resolution of face images using kernel pca-based prior. TMM **9**(4), 888–892 (2007)
4. Chen, C., Gong, D., Wang, H., Li, Z., Wong, K.Y.K.: Learning spatial attention for face super-resolution. TIP **30**, 1219–1231 (2020)
5. Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., Wong, K.Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: CVPR. pp. 11896–11905 (2021)
6. Chen, K., Chen, B., Liu, C., Li, W., Zou, Z., Shi, Z.: Rsmamba: Remote sensing image classification with state space model. GRSL **21**, 1–5 (2024)
7. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: CVPR. pp. 2492–2501 (2018)
8. Cui, Y., Knoll, A.: Dual-domain strip attention for image restoration. NN **171**, 429–439 (2024)
9. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: CVPR. pp. 5203–5212 (2020)
10. Duan, Z.P., Zhang, J., Jin, X., Zhang, Z., Xiong, Z., Zou, D., Ren, J.S., Guo, C., Li, C.: Dit4sr: Taming diffusion transformer for real-world image super-resolution. In: ICCV. pp. 18948–18958 (2025)
11. Feng, R., Li, C., Loy, C.C.: Kalman-inspired feature propagation for video face super-resolution. In: ECCV. pp. 202–218 (2024)
12. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
13. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. In: ECCV. pp. 222–241 (2024)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
15. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
16. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: ICCV. pp. 1689–1697 (2017)
17. Jia, K., Gong, S.: Generalized face super-resolution. TIP **17**(6), 873–886 (2008)
18. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: WACV. pp. 1548–1558 (2021)
19. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)
21. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: ECCV. pp. 679–692 (2012)
22. Lei, X., Zhang, W., Cao, W.: Dvmsr: Distillated vision mamba for efficient super-resolution. In: CVPR. pp. 6536–6546 (2024)
23. Li, T., Li, C., Lyu, J., Pei, H., Zhang, B., Jin, T., Ji, R.: Damamba: Vision state space model with dynamic adaptive scan. In: NeurIPS. pp. 39777–39799 (2026)
24. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCVW. pp. 1833–1844 (2021)

25. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW. pp. 136–144 (2017)
26. Liu, S., Duan, Z.P., OuYang, J., Fu, J., Park, H., Liu, Z., Guo, C.L., Li, C.: Faceme: Robust blind face restoration with personal identification. In: AAAI. pp. 5567–5575 (2025)
27. Liu, W., Lin, D., Tang, X.: Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In: CVPR. pp. 478–484 (2005)
28. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. In: NeurIPS. pp. 103031–103063 (2024)
29. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV. pp. 3730–3738 (2015)
30. Long, W., Zhou, X., Zhang, L., Gu, S.: Progressive focused transformer for single image super-resolution. In: CVPR. pp. 2279–2288 (2025)
31. Lu, T., Wang, Y., Zhang, Y., Wang, Y., Wei, L., Wang, Z., Jiang, J.: Face hallucination via split-attention in split-attention network. In: ACM MM. pp. 5501–5509 (2021)
32. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J.: Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In: CVPR. pp. 5569–5578 (2020)
33. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
34. Shi, Y., Xia, B., Jin, X., Wang, X., Zhao, T., Xia, X., Xiao, X., Yang, W.: Vmambair: Visual state space model for image restoration. TCSVT **35**(6), 5560–5574 (2025)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
36. Wang, C., Jiang, J., Zhong, Z., Liu, X.: Spatial-frequency mutual learning for face super-resolution. In: CVPR. pp. 22356–22366 (2023)
37. Wang, C., Jiang, J., Zhong, Z., Zhai, D., Liu, X.: Super-resolving face image by facial parsing information. TBBIS **5**(4), 435–448 (2023)
38. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., Qiao, Y.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: CVPR. pp. 14408–14419 (2023)
39. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: CVPR. pp. 9168–9178 (2021)
40. Wang, Y., Lu, T., Zhang, Y., Wang, Z., Jiang, J., Xiong, Z.: Faceformer: Aggregating global and local representation for face hallucination. TCSVT **33**(6), 2533–2545 (2022)
41. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV. pp. 3–19 (2018)
42. Xiao, Y., Yuan, Q., Jiang, K., Chen, Y., Zhang, Q., Lin, C.W.: Frequency-assisted mamba for remote sensing image super-resolution. TMM (2024)
43. Yang, C., Chen, Z., Espinosa, M., Ericsson, L., Wang, Z., Liu, J., Crowley, E.J.: Plainmamba: Improving non-hierarchical mamba in visual recognition. arXiv preprint arXiv:2403.17695 (2024)
44. Yang, L., Wang, S., Ma, S., Gao, W., Liu, C., Wang, P., Ren, P.: Hifacegan: Face renovation via collaborative suppression and replenishment. In: ACM MM. pp. 1551–1560 (2020)
45. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: ECCV. pp. 318–333 (2016)
46. Yu, X., Porikli, F.: Face hallucination with tiny unaligned images by transformative discriminative neural networks. In: AAAI (2017)
47. Zhang, K., Zuo, W., Gu, S., Zhang, L.: Learning deep cnn denoiser prior for image restoration. In: CVPR. pp. 3929–3938 (2017)
48. Zhang, K., Li, D., Luo, W., Liu, J., Deng, J., Liu, W., Zafeiriou, S.: Edface-celeb-1 m: Benchmarking face hallucination with a million-scale dataset. TPAMI **45**(3), 3968–3978 (2022)

49. Zhang, K., Li, D., Luo, W., Ren, W., Stenger, B., Liu, W., Li, H., Yang, M.H.: Benchmarking ultra-high-definition image super-resolution. In: ICCV. pp. 14769–14778 (2021)
50. Zhang, K., Zhang, Z., Cheng, C.W., Hsu, W.H., Qiao, Y., Liu, W., Zhang, T.: Super-identity convolutional neural network for face hallucination. In: ECCV. pp. 183–198 (2018)
51. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
52. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV. pp. 286–301 (2018)
53. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR. pp. 2472–2481 (2018)
54. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. CSUR **35**(4), 399–458 (2003)
55. Zheng, X., Guo, Y., Huang, H., Li, Y., He, R.: A survey of deep facial attribute analysis. IJCV **128**, 2002–2034 (2020)
56. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Learning face hallucination in the wild. In: AAAI (2015)
57. Zhou, S., Chan, K., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: NeurIPS. pp. 30599–30611 (2022)
58. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: efficient visual representation learning with bidirectional state space model. In: ICML. pp. 62429–62442 (2024)
59. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: ECCV. pp. 614–630 (2016)