PortraitSR: Artist-Inspired Prior Learning for Progressive Face Super-Resolution

Miaoqing Wang[1,2], Jiaxu Leng[1,2*], Shuang Li[1,2], Changjiang Kuang[1,2], Long Sun[3]

[1]School of Computer Science and Technology, Chongqing University of Post and Telecommunication
[2]Chongqing Institute for Brain and Intelligence, Guangyang Bay Laboratory, Chongqing, China
[3]School of Computer Science and Engineering, Nanjing University of Science and Technology
{s230201110, s230201049}@stu.cqupt.edu.cn, lengjx@cqupt.edu.cn, shuangli936@gmail.com, cs.longsun@njust.edu.cn

Abstract

Face super-resolution (FSR) aims to reconstruct high-resolution (HR) face images from low-resolution (LR) inputs. While recent methods have advanced this task through architectural innovations and generative modeling, but they often leads to semantically inconsistent structures and unrealistic textures, particularly under high magnification. To mitigate these limitations, we draw inspiration from the human artistic process of "structuring before detailing" and propose a progressive prior-guided restoration strategy. Specifically, we first introduce a Sketching Structure Prior (SSP) module that embeds global semantics and refines local geometry through implicit parsing guidance and explicit spatial modulation. Then, an Associative Texture Prior (ATP) module leverages a High-Quality Dictionary (HD) learned from high-quality reconstruction to guide fine-grained detail recovery. Finally, to unify structure and detail features, we design a Holistic Prior Fusion (HPF) module that adaptively integrates them within semantically consistent facial regions. Our method surpasses state-of-the-art on CelebA and Helen in both structural fidelity and texture realism.

Code — https://github.com/amazingwmq/PortraitSR.

Introduction

Face super-resolution (FSR) aims to reconstruct high-resolution (HR) facial images from low-resolution (LR) inputs by recovering global structures and fine-grained textures. In real-world scenarios such as surveillance and social media (Leng et al. 2025; Li et al. 2025b,c; Yang et al. 2025b; Yan et al. 2023; Han et al. 2025), image quality is often degraded due to hardware or environmental limitations, which impairs downstream tasks like face recognition that rely on detailed and consistent facial representations. Thus, recovering high-fidelity textures and structures from degraded inputs remains challenging.

Recent advances in FSR have primarily focused on enhancing model capacity through architectural and generative innovations. These methods can be broadly categorized into two paradigms: (1) Architectural Modeling, which includes CNN-based and Transformer-based architectures (Gao et al. 2023; Bao et al. 2023; Sun et al. 2023; Sun, Pan, and Tang 2022; Zheng et al. 2025) designed to improve feature representation and contextual modeling; and (2) Generative Modeling, which leverages techniques like GAN-based and diffusion-based models (Wang et al. 2021; Yue and Loy 2024; Wang et al. 2024c, 2025) to generate visually realistic face images by modeling high-frequency details and complex textures. However, as illustrated in Fig. 1(a), these methods often produce distorted geometric structures (e.g., deformed eyes) or hallucinate unrealistic textures (e.g., overly smooth or unnatural teeth), especially under high magnification or severely degraded inputs. This highlights the intrinsic limitations of learning directly from LR inputs, which impedes the model's ability to recover coherent structures and realistic textures. To address these challenges, prior works have incorporated explicit priors, from 2D structures (Wang et al. 2022a; Leng and Wang 2022; Leng et al. 2022; Li et al. 2025d) (e.g., landmarks, parsing maps) to high-level semantics (Xie et al. 2023; Leng et al. 2024; Li et al. 2024b) and 3D geometry (Chen et al. 2024), to guide the reconstruction process. While these priors aid global structure modeling, they tend to overlook fine textures and structural-detail consistency. This leaves **the challenge of jointly preserving geometric integrity and restoring fine-grained realism** largely unresolved. This motivates us to rethink the modeling process from a higher-level cognitive perspective.

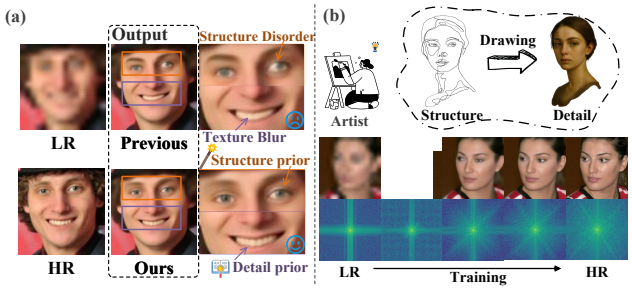

Figure 1: Motivation. (a) Visual comparison with previous methods. Our prior-guided coarse-to-fine strategy produces clearer structures and finer textures. (b) Like sketching before detailing, models progress from simple to complex patterns, as illustrated in spatial and frequency visualizations.

*Corresponding authors.

Inspired by the artistic process of sketching, detailing, and refining, as illustrated in Fig. 1(b), we draw an analogy between human prior-driven learning and model training, both of which follow a progression from simple to complex patterns. While humans rely on inherent priors to guide their learning and refinement, models lacking such priors are prone to hallucinations when encountering unseen patterns. To address this limitation, we propose a **prior-guided coarse-to-fine framework** that mimics the internal knowledge humans use to infer and refine. where structural and textural priors are introduced to guide global layout preservation and fine-detail enhancement.

Emulating the staged workflow of portrait artists, we observe that in the early stage of reconstruction, artists typically sketch structural outlines to anchor key facial components before refining textures. Inspired by this, we introduce facial parsing maps as structural priors to provide semantically meaningful and spatially aligned guidance. After recovering coarse structures, models often struggle to reconstruct high-frequency facial details. In contrast, humans benefit from powerful internal priors formed through accumulated experience, which allow them to imagine and refine plausible textures. To mimic this capability, we introduce detail priors in the later stages to progressively guide the synthesis of fine-grained facial cues. Finally, to address the spatial misalignment between structure and texture, we propose a region-aware fusion module, echoing the artist's final refinement stage, where local adjustments are made to harmonize fine brushwork with the overall composition, ensuring both perceptual consistency. Notably, this coarse-to-fine progression also reflects the essence of the generalized curriculum learning paradigm (Wang et al. 2024b), wherein models are encouraged to first master easier tasks, such as global structure recovery, before tackling more complex objectives like fine texture restoration.

Motivated by these insights, we propose PortraitSR, **an artist-inspired prior learning for progressive face super-resolution**. We first introduce the Sketching Structure Prior (SSP), inspired by the human strategy of first outlining coarse structures before adding details. It incorporates structural priors through two complementary mechanisms: implicit supervision encourages global semantic learning via facial parsing, while explicit feature modulation injects spatially aligned cues to enhance local geometric consistency. Building upon this structural foundation, we further incorporate an Associative Texture Prior (ATP), which integrates a high-quality dictionary learned from a reconstruction task, which mimics the human ability to recall plausible textures from visual memory. Finally, to reconcile and unify the coarse structural layout and refined details, we propose a Holistic Prior Fusion (HPF) that adaptively fuses the two branches, it utilizes facial component masks from the parsing branch to perform region-aware feature integration. Our main contributions are as follows:

- We propose a cognitively inspired FSR paradigm with a hierarchical coarse-to-fine architecture that emulates the human sketching process, progressively reconstructing facial structures and textures.
- We introduce SSP and ATP to form a dual-prior learning scheme that disentangles and enhances structural and textural representations.
- To further integrate spatial priors, we introduce an HPF module that performs region-aware fusion guided by facial component masks, strengthening local consistency.
- Experiments on CelebA and Helen demonstrate the effectiveness of our architecture and prior-guided design in preserving structure and recovering high-fidelity details.

Related Work

Face Super-Resolution

Face super-resolution has evolved from early shallow models (Baker and Kanade 2000; Wang and Tang 2005; Chakrabarti, Rajagopalan, and Chellappa 2007) to deep CNN-based and Transformer-based architectures (Yang, Ma, and Yang 2014; Gao et al. 2023; Li et al. 2024a, 2025a), with increasing emphasis on incorporating facial priors to address the ill-posed nature of the task. Explicit priors, such as facial landmarks, parsing maps, and alignment information have been widely used to guide structural consistency and identity preservation (Chen et al. 2018; Yin et al. 2020). Additionally, cross-modal priors, including textual cues, have emerged to further enhance semantic guidance (Xie et al. 2023). These prior-guided methods markedly enhance restoration under extreme degradation. However, existing methods rely on fixed, low-frequency priors, limiting the recovery of fine details. To overcome this, we introduce a novel prior framework inspired by the progressive refinement process used by human artists, facilitating robust, perceptually consistent reconstruction.

Dictionary Learning-Based Methods

Dictionary learning plays a crucial role in enhancing detail fidelity and restoration quality in face restoration tasks. VQFR (Gu et al. 2022) restores facial details and maintains identity consistency using a VQ codebook. DMDNet (Li et al. 2022) employs a dual-dictionary design that separately stores general facial and identity-specific features, using a dictionary transformation module and multi-scale recovery method to handle diverse restoration scenarios. RestoreFormer (Wang et al. 2022b) utilizes a full-space attention mechanism and a high-quality facial feature dictionary, significantly improving detail recovery. Overall, dictionary learning has proven to be a key method for enhancing detail recovery in face restoration. Inspired by these advances, we introduce a novel dictionary-based prior for facial detail restoration, which progressively refines both structural and textural features to further enhance restoration quality.

Methodology

Overview

We propose PortraitSR, an artist-inspired prior learning for progressive face super-resolution, as illustrated in Fig. 2. PortraitSR adopts a Transformer-based encoder–decoder backbone for its strong representational power. Specifically, the encoder $E(\cdot)$ encodes the low-resolution input

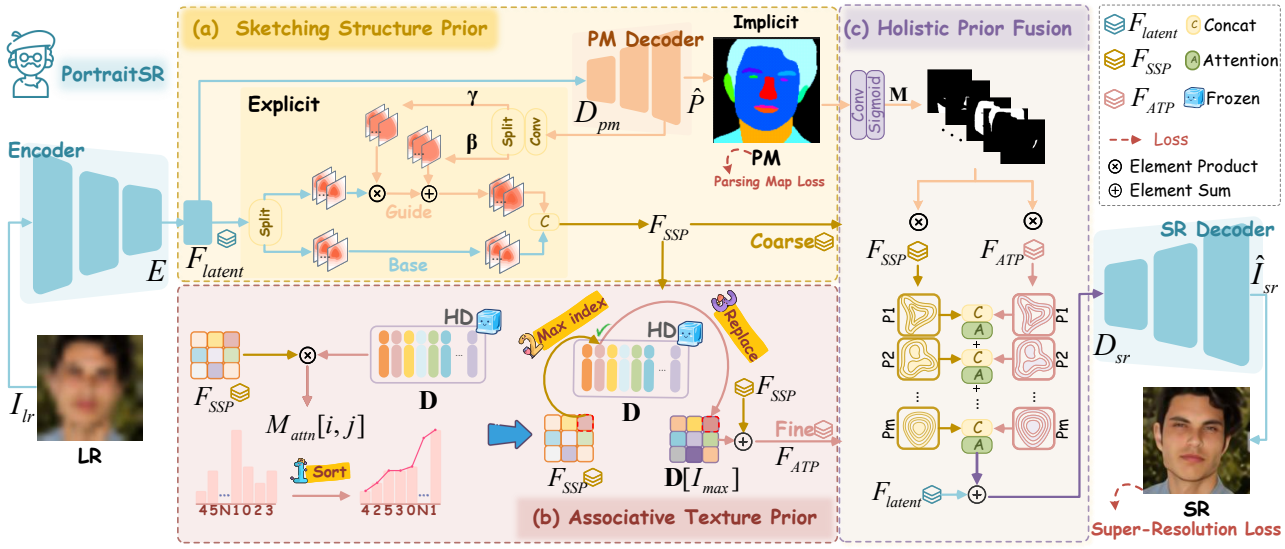

Figure 2: Overview of the proposed PortraitSR framework. It comprises three key modules: (1) Sketching Structure Prior (SSP) enhances structural representation by injecting parsing-guided geometry; (2) Associative Texture Prior (ATP) retrieves fine-grained detail features from a learnable high-resolution dictionary; (3) Holistic Prior Fusion (HPF) performs region-aware integration of structural and textural features guided by parsing-based masks.

$I_{lr} \in \mathbb{R}^{H\times W\times 3}$ into latent representations $F_{latent} \in \mathbb{R}^{H/8\times W/8\times 8C}$, and the decoder $D_{sr}(\cdot)$ takes the enhanced features as input and generates the super-resolved output $\hat{I}_{sr}$.

While the backbone captures high-level features, it lacks dedicated mechanisms for structure and texture modeling, which often leads to unstable and suboptimal restoration results. To address this, PortraitSR introduces three specialized modules: (1) **Sketching Structure Prior (SSP)**: Compensates for geometric degradation by injecting implicit supervision and explicit feature modulation into F_{latent}, producing structure-aware features F_{SSP}. (2) **Associative Texture Prior (ATP)**: Complements detail modeling by injecting texture priors from a learnable dictionary, yielding refined features F_{ATP}. (3) **Holistic Prior Fusion (HPF)**: Aligns F_{SSP} and F_{ATP} via mask-guided region-aware integration, mitigating spatial and semantic mismatches to produce a unified decoding representation. Together, these modules enable progressive reconstruction with structural fidelity and fine-detail recovery.

Sketching Structure Prior (SSP)

Low-resolution images under severe degradation often lack reliable geometry, making structural recovery difficult. Parsing maps (PM) encode rich structural priors, such as part locations and spatial relationships. which are inherently resolution-invariant, thereby furnishing low-resolution images with a resolution-agnostic "shape sketch" to guide reliable reconstruction. However, previous methods often utilize them in limited ways (e.g., pre-processing or shallow guidance), hindering effective alignment during restoration. To fully leverage parsing priors to guide structural consistency, we adopt a complementary strategy of implicit supervision and explicit feature modulation: the former promotes global semantic alignment, the latter injects spatially precise cues to resolve geometric ambiguities.

Implicit Supervision. Relying solely on degraded inputs, the encoder often struggles to capture consistent and accurate structural semantics. To address this, we introduce facial parsing supervision to enhance the encoder's ability to model geometry-aware representations. Unlike conventional multi-task frameworks that treat parsing as an auxiliary task, our implicit supervision directly guides the shared encoder via parsing prediction supervision, embedding structural priors into the SR pathway.

Specifically, the shared latent features F_{latent} serve both the super-resolution reconstruction and the auxiliary facial parsing prediction, with the parsing map $\hat{P}$ generated as:

$$\hat{P} = D_{pm}(F_{latent}), \tag{1}$$

where $D_{pm}(\cdot)$ denotes the parsing map decoder, $\hat{P}$ is the predicted parsing map. This implicit supervision effectively injects structural semantics into the shared representation, enabling the encoder $E(\cdot)$ to retain global facial geometry.

Explicit Feature Modulation. Although implicit supervision enhances global structural understanding, it lacks the spatial precision needed for fine-grained reconstruction. To address this, we introduce an explicit feature modulation that injects parsing priors in a structured and controllable way, aligning latent features F_{latent} with spatial semantics to improve structural consistency.

To inject fine-grained structural cues into latent representations, we derive modulation parameters by applying a 3×3 convolution to parsing features F_{pm}, and split the output into spatially aligned scaling and shifting parameters:

$$\boldsymbol{\gamma}, \boldsymbol{\beta} = Split(Conv^{mod}_{3\times3}(F_{pm})), \tag{2}$$

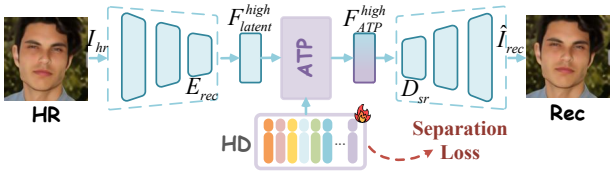

Figure 3: Detail prior learning strategy of the Associative Texture Prior (ATP).

where $Split(\cdot)$ represent channel-wise split operations. $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ represent spatial structural guidance derived from F_{pm}, used to modulate intermediate latent features.

To inject structural guidance without overwhelming the original representation, we split F_{latent} into two channel-wise parts: F_1 is modulated by $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ for geometric alignment, while F_2 retains unaltered semantics. The concatenated output forms the structure-aware representation F_{SSP}:

$$F_{SSP} = [(\boldsymbol{\gamma} \odot Split(F_{latent})_1 + \boldsymbol{\beta}), Split(F_{latent})_2], \quad (3)$$

where $\odot$ and $[,]$ denotes element-wise multiplication and channel-wise concatenation. F_{SSP} is the output of SSP.

Associative Texture Prior (ATP)

Although structural priors help maintain global consistency, they cannot reconstruct fine textures that are often absent and irrecoverable from low-resolution inputs. To close this gap, we turn to high-resolution (HR) images, which are rich in reusable high-frequency patterns. If these patterns can be memorized during training, the model can later "recall" them to replenish missing details at inference time. Building on this intuition, we introduce ATP constructs a learnable high-quality dictionary from HR images and retrieves semantically relevant priors via associative matching. Specifically, it comprises two stages: a detail prior learning strategy and an application of detail priors, which collaboratively learn to preserve and effectively utilize texture cues for fine-grained reconstruction. Emulating human memory, the design retrieves high-frequency cues for realistic detail recovery under severe degradation.

Detail Prior Learning Strategy. Unlike prior works using fixed dictionaries (Li et al. 2020; Wang et al. 2023b) or external references (Lu et al. 2021), ATP adopts a learn-as-you-use strategy that each input dynamically constructs task-aware guidance for fine-grained detail recovery, ensuring generalizability and efficiency.

We adopt a reconstruction framework that shares the encoder–decoder structure (with shared decoder weights) of the super-resolution model (see Fig. 3), which ensures alignment in feature space and facilitates seamless transfer of learned texture priors. Within this framework, a learnable dictionary $\mathbf{D} \in \mathbb{R}^{N \times C}$ is initialized from $\mathcal{N}(\mu, \delta^2)$ to serve as a memory bank for storing discriminative texture representations. Given a high-resolution input $I_{hr} \in \mathbb{R}^{H \times W \times 3}$, the encoder extracts latent features F_{latent}^{high}, which are refined by ATP using $\mathbf{D}$ to produce texture-enhanced features F_{ATP}^{high}. The decoder $D_{sr}(\cdot)$ then reconstructs the image $\hat{I}_{rec}$:

$$\hat{I}_{rec} = D_{sr}(ATP(E_{rec}(I_{hr}), \mathbf{D})), \quad (4)$$

where E_{rec} denote the reconstruction encoder, the reconstruction decoder D_{sr} is shared with the SR task. This high-resolution reconstruction process drives the dictionary to encode high-fidelity texture priors, which are later retrieved to guide super-resolution inference.

Application of Detail Priors. To inject high-fidelity textures during inference, we retrieve the most relevant priors from the learned dictionary based on input features. Specifically, we compute the cosine similarity between each spatial location in $F_{SSP} \in \mathbb{R}^{H \times W \times C}$ and dictionary entries, enabling selective incorporation of semantically aligned details:

$$M_{attn}[i,j] = \frac{F_{SSP}[i] \cdot \mathbf{D}[j]}{\|F_{SSP}[i]\|_2 \cdot \|\mathbf{D}[j]\|_2}, \quad (5)$$

where $M_{attn}[i,j]$ indicates the similarity between the i-th spatial feature and the j-th dictionary entry.

We then identify the index of the dictionary entry with the highest similarity for each spatial location. These selected detail priors are integrated into the original features $F_{SSP}[i]$ via residual addition to enhance the representation:

$$F_{ATP}[i] = F_{SSP}[i] + \mathbf{D}[\arg\max_j M_{attn}[i,j]], \quad (6)$$

where $\arg\max_j M_{attn}[i,j]$ denotes the dictionary entry most similar to the feature at location i. The resulting F_{ATP} contains the corresponding texture-enhanced features.

Holistic Prior Fusion (HPF)

Despite the complementary nature of structural and texture priors, feature-level inconsistencies arise due to modality heterogeneity. HPF mitigates this by performing spatially aligned fusion under the guidance of parsing maps, which performs region-aware alignment and integration of structure and texture features, analogous to the final refinement phase in artistic creation, where shading and linework are harmonized for visual coherence.

Leveraging the regional cues in parsing maps, we derive m spatial attention masks $\mathbf{M}$ from the parsing map $\hat{P}$:

$$\mathbf{M} = \sigma(Conv_{3\times3}^{pre}(\hat{P})), \quad \mathbf{M} \in \mathbb{R}^{H \times W \times m}, \quad (7)$$

where we use a convolutional layer $Conv_{3\times3}^{pre}(\cdot)$ and sigmoid activation $\sigma(\cdot)$ to obtain $\mathbf{M}$, each mask $\mathbf{M}^{(k)}$ modulates F_{SSP} and F_{ATP} to produce region-specific features for semantic component:

$$F_{SSP}^{(k)} = \mathbf{M}^{(k)} \odot F_{SSP}, \quad F_{ATP}^{(k)} = \mathbf{M}^{(k)} \odot F_{ATP}, \quad (8)$$

where $k = 1, 2, \ldots, m$ and $F_{SSP}^{(k)}$ and $F_{ATP}^{(k)}$ represent region-specific structural and texture features, which are concatenated and refined via an attention module:

$$F_{attn}^{(k)} = Attn([F_{SSP}^{(k)}, F_{ATP}^{(k)}]), \quad (9)$$

where $Attn(\cdot)$ sequentially applies channel and spatial attention to jointly model global and local dependencies, yielding the enhanced regional feature $F_{attn}^{(k)}$.

To integrate complementary cues from all regions, the attentive features $F_{attn}^{(k)}$ are aggregated and added residually to

Method	Venue	CelebA (Liu et al. 2015)						Helen (Le et al. 2012)					
		×4			×8			×4			×8		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Bicubic	-	27.48	0.817	0.184	23.58	0.269	0.269	28.22	0.663	0.177	23.88	0.663	0.256
DIC	20'CVPR	31.53	0.911	0.053	27.37	0.802	0.092	31.98	0.930	0.058	26.94	0.803	0.114
SPARNet	20'TIP	31.71	0.913	0.048	27.42	0.804	0.089	31.98	0.930	0.059	26.95	0.083	0.117
NLSN	21'CVPR	32.08	0.919	0.087	27.45	0.804	0.091	31.91	0.916	0.061	26.78	0.793	0.119
KDFSR	22'TCSVT	31.41	0.908	0.080	26.94	0.804	0.112	31.28	0.910	0.089	26.75	0.793	0.126
CTCNet	23'TIP	32.02	0.917	0.098	27.65	0.808	0.182	32.62	0.926	0.094	26.69	0.801	0.194
MOHA†	23'PR	–	–	–	27.11	0.786	–	–	–	–	26.82	0.795	–
SFMNet	23'CVPR	32.01	0.918	0.044	27.56	0.807	0.087	32.51	0.936	0.050	27.22	0.814	0.106
ECDP	24'AAAI	28.08	0.824	0.163	25.25	0.725	0.142	28.10	0.825	0.183	24.65	0.742	0.151
IC-DENet	24'AAAI	29.93	0.868	0.163	25.85	0.795	0.194	29.10	0.850	0.200	25.16	0.745	0.222
DPI†	25'AAAI	–	–	–	24.97	–	0.140	–	–	–	–	–	–
Ours	—	**32.22**	**0.920**	**0.042**	**27.90**	**0.819**	**0.079**	**32.79**	**0.939**	**0.042**	**27.70**	**0.832**	**0.089**

Table 1: Quantitative comparison on the CelebA and Helen datasets for ×4 and ×8 face hallucination. Bold and underlined represent the best and second best. (†: Results are quoted from the original paper due to the absence of publicly available code.)

the original latent feature F_{latent}, preserving global semantics. The result is then decoded by $D_{sr}(\cdot)$ to generate the final super-resolved image:

$$\hat{I}_{sr} = D_{sr}(F_{latent} + \sum_{k=1}^{m} F_{attn}^{(k)}), \tag{10}$$

where $\sum_{k=1}^{m} F_{attn}^{(k)}$ represents the sum of locally enhanced detail features obtained from HPF, and $\hat{I}_{sr}$ represents the final output of the super-resolution process.

Model Objectives

In order to effectively train the proposed network we designed the following multiple loss functions:

Super-Resolution Loss: The super-resolution branch is jointly optimized with reconstruction and perceptual losses:

$$\mathcal{L}_1 = \left\| \hat{I}_{sr} - I_{hr} \right\|_1, \quad \mathcal{L}_{per} = \left\| \phi(\hat{I}_{sr}) - \phi(I_{hr}) \right\|_1, \tag{11}$$

where $\hat{I}_{sr}$ denotes the super-resolved output, I_{hr} is the ground-truth high-resolution image, and $\phi(\cdot)$ represents the feature extractor based on a pre-trained VGG (Simonyan and Zisserman 2014) network.

Parsing Map Loss: To enforce accurate structural guidance, we apply a supervision loss on the predicted PM:

$$\mathcal{L}_{pm} = \left\| \hat{P} - P_{gt} \right\|_1, \tag{12}$$

where $\hat{P}$ and P_{gt} denote the predicted and ground-truth parsing maps(Wang et al. 2022a).

Dictionary Learning Loss: To enhance detail modeling, the dictionary learning branch is jointly optimized with a reconstruction loss between the generated image $\hat{I}_{rec}$ and I_{hr}, and a dictionary constraint that suppresses off-diagonal correlations in the self-similarity matrix to promote diversity and reduce redundancy:

$$\mathcal{L}_{rec} = \left\| \hat{I}_{rec} - I_{hr} \right\|_1, \tag{13}$$

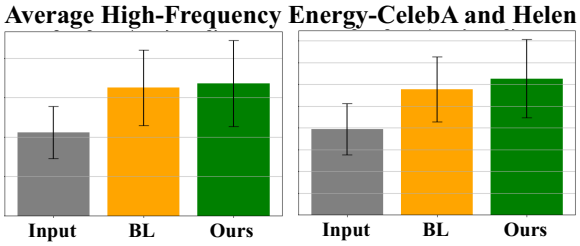

Figure 4: Average high-frequency energy on CelebA and Helen: bars represent mean energy in the high-frequency band (0.5–1.0), with error bars indicating standard deviation.

$$\mathcal{L}_{dic} = \frac{1}{B \cdot N(N-1)/2} \sum_{i \neq j} sim(\mathbf{D}_i, \mathbf{D}_j), \tag{14}$$

where $\mathbf{D}_{i/j}$ is the dictionary matrix, N is the dictionary size, and B is the batch size.

Total Loss: The overall objective is formulated as a weighted sum of the above loss terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \lambda_1 \cdot \mathcal{L}_{per} + \lambda_2 \cdot \mathcal{L}_{\text{pm}} + \lambda_3 \cdot \mathcal{L}_{rec} + \lambda_4 \cdot \mathcal{L}_{dic}, \tag{15}$$

where λ_1, λ_2, λ_3, and λ_4 are weighting coefficients that control the relative importance of each loss term.

Experiments

Datasets

We train our model on CelebA (Liu et al. 2015) due to its large-scale diversity and evaluate it on both CelebA and Helen (Le et al. 2012), which are widely used benchmarks for facial image analysis. Specifically, we use 128×128 ground truth (GT) HR face images, generating 4× and 8× FSR inputs by downsampling HR images to 32×32 and 16×16 using bicubic interpolation. We use 168,854 images from CelebA for training and evaluate on 1,000 samples from CelebA and 50 samples from Helen, using PSNR,

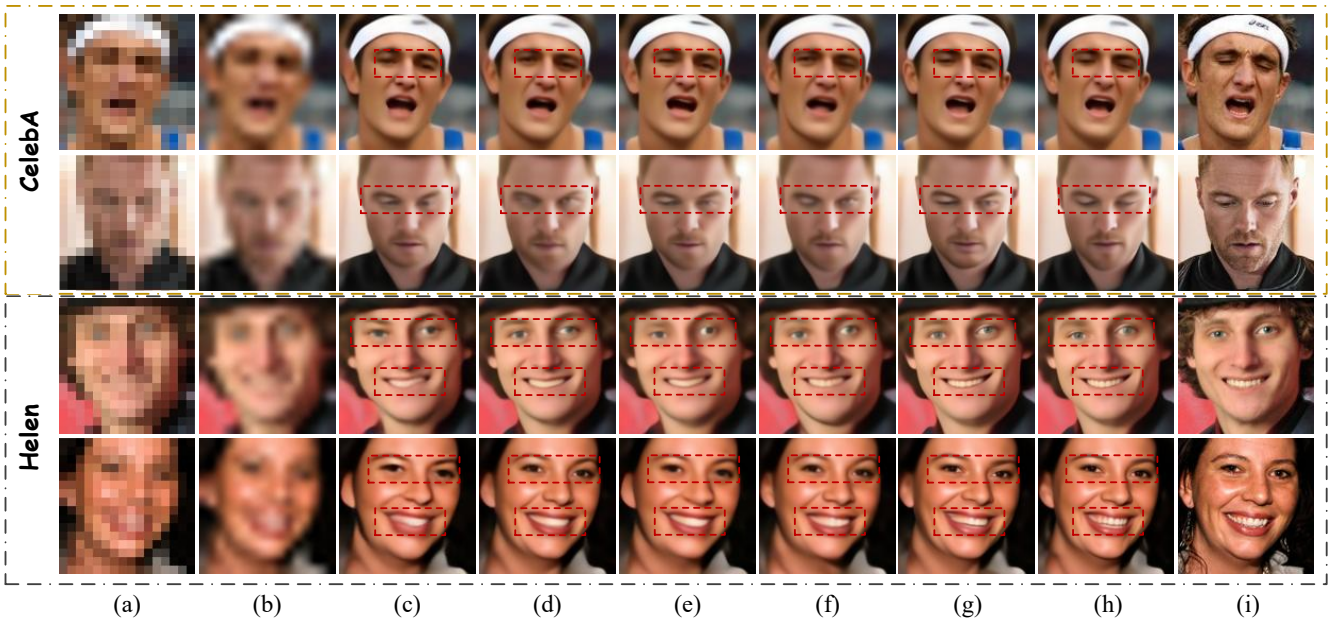

Figure 5: Visual comparison results of different methods on CelebA and Helen by ×8. Please zoom in for better view. (a) LR, (b) Bicubic, (c) DIC (Ma et al. 2020), (d) SPARNet (Chen et al. 2020), (e) KDFSR (Wang et al. 2022a), (f) CTCNet (Gao et al. 2023), (g) SFMNet (Wang et al. 2023a), (h) PortraitSR, (i) HR.

Components				Dataset	
Baseline	SSP	ATP	HPF	CelebA	Helen
✓	✗	✗	✗	27.57	27.25
✓	✓	✗	✗	27.72 (+0.15)	27.44 (+0.19)
✓	✗	✓	✗	27.66 (+0.09)	27.40 (+0.15)
✓	✓	✓	✗	27.74 (+0.17)	27.47 (+0.22)
✓	✓	✓	✓	**27.90** (+0.33)	**27.70** (+0.45)

Table 2: Ablation study on CelebA and Helen, using the PSNR to evaluate the effects of SSP, ATP, and HPF.

Settings		Fusion Strategy			Performance	
Base	PM	Att	Cat	SSP	PSNR/SSIM	PSNR/SSIM
✓	✗	✗	✗	✗	27.57 / 0.810	27.25 / 0.818
✓	✓	✗	✗	✗	27.65 / 0.812	27.39 / 0.821
✓	✓	✓	✗	✗	27.68 / 0.813	27.36 / 0.823
✓	✓	✗	✓	✗	27.65 / 0.813	27.35 / 0.822
✓	✓	✗	✗	✓	**27.72 / 0.813**	**27.44 / 0.823**

Table 3: Ablation study on CelebA and Helen (×8) to verify the effectiveness of SSP.

SSIM (Wang and Bovik 2002), and LPIPS (Zhang et al. 2018) for evaluation.

Implementation Details

We implement our model using the PyTorch framework. The network is optimized with the Adam optimizer, where $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and the initial learning rate is set to 2×10^{-4}. We set the dictionary size to $N = 512$ and feature dimension to $C = 384$. The dictionary $\mathbf{D} \in \mathbb{R}^{512\times384}$ is initialized with a normal distribution, $\mu = 0$ and $\delta = 0.02$. The loss function is composed of four components: $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, and $\mathcal{L}_4$, with weights 0.1, 0.005, 0.5, and 0.001 respectively.

Comparison With State-of-the-Art Methods

In this section, we compare our method with state-of-the-art approaches in both FSR (Ma et al. 2020; Chen et al. 2020; Wang et al. 2022a; Wei et al. 2023; Wang et al. 2023a, 2024a; Yang et al. 2025a) and general SR (Mei, Fan, and Zhou 2021; Gao et al. 2023; Yuan and Yuan 2024). All models are trained and tested under consistent settings. We evaluate PortraitSR on the CelebA and Helen datasets. As shown in Table 1, PortraitSR consistently outperforms state-of-the-art methods. We visualize the average high-frequency energy with standard deviation across models, and Fig. 4 shows that PortraitSR restores richer fine details with greater stability. The visual comparisons in Fig. 5 highlight its advantage in recovering high-frequency facial details such as the eyes and nose, where other methods often produce blurry results or hallucinated, structurally inconsistent artifacts.

Ablation Studies

We perform ablation studies to evaluate the impact of each component in PortraitSR by removing or modifying individual modules and measuring the resulting performance.

Component-Wise Ablations. We conduct ablation studies to assess the effectiveness of each component in PortraitSR. As shown in Table 2, individually adding SSP or ATP to the baseline yields consistent improvements, while

Figure 6: SSP ablation visualizations with low-frequency wavelet.

Detail Prior Learning				Dataset	
BL	HD-RF	HD-Retrain	Ours	CelebA	Helen
✓	✗	✗	✗	27.57	27.25
✓	✓	✗	✗	27.62 (+0.05)	27.37 (+0.12)
✓	✗	✓	✗	27.61 (+0.04)	27.38 (+0.13)
✓	✗	✗	✓	**27.66** (+0.09)	**27.40** (+0.15)

Table 4: Ablation study on CelebA and Helen ($\times 8$), using the PSNR to verify the effectiveness of ATP.

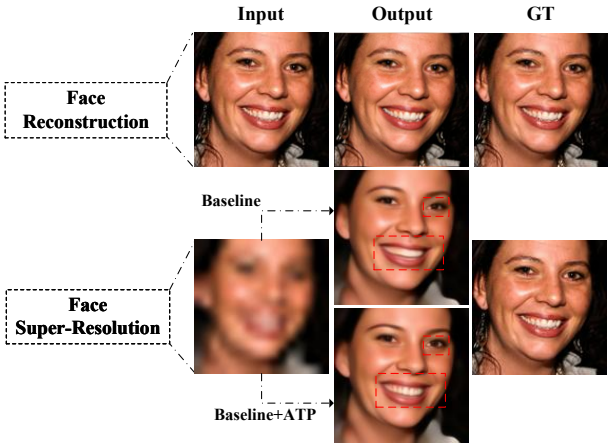

Figure 7: Visualizations of input, output, and GT: the first row shows face reconstruction, and the second and third rows show baseline and ATP-enhanced results.

combining both leads to further gains, validating the coarse-to-fine design. However, the improvement remains limited due to insufficient interaction. Introducing HPF results in a significant performance boost, highlighting its role in enhancing the fusion of structural and detail features.

Effectiveness of Sketching Structure Prior (SSP). We evaluate five model variants to analyze the impact of structural priors. As shown in Table 3, PM supervision enhances performance, but naive fusion strategies (Attention, Concat) degrade results due to semantic misalignment. In contrast, SSP achieves the best performance by selectively modulating SR features with structure-aware guidance. Visual results about low-frequency wavelet in Fig. 6 further support.

Layers	PSNR↑	SSIM↑	LPIPS↓	VIF↑
None	**27.90/ 27.70**	**0.819**/0.832	0.079/0.089	**0.418/0.425**
1-layer	27.86/27.68	0.818/**0.833**	0.079/**0.085**	0.416/0.425
2-layer	27.85/27.64	0.819/0.832	**0.078**/0.086	0.416/0.424
3-layer	27.82/27.72	0.819/0.833	0.078/0.086	0.414/0.423

Table 5: Ablation study on the number of decoder fusion layers on CelebA and Helen ($\times 8$).

Effectiveness of Associative Texture Prior (ATP). We compare three dictionary integration strategies in Table 4: (1) HD-RF: a pretrained HD learned from an external dataset, (2) HD-Retrain: a retrained HD using a different model architecture on the target dataset, (3) ours is jointly trained within the same framework and dataset. While the first two settings suffer from domain or architectural mismatch, ours achieves superior performance via aligned, task-aware feature transfer. Fig. 7 shows the dictionary task significantly improves texture and structure recovery in FSR.

Effectiveness of the "Structuring Before Detailing". To examine the importance of maintaining a coarse-to-fine learning order, we study the effect of injecting structural cues at different decoder stages. As shown in Table 5, performance consistently degrades with deeper-stage injection. This indicates that late-stage structural guidance disrupts already refined representations. These results support our design rationale: structural cues are most effective when introduced early to shape global layouts, followed by localized refinement. Preserving this top-down flow is essential for stable and semantically aligned reconstruction.

Conclusion

We propose PortraitSR, a cognitively inspired face super-resolution framework that mimics the human sketching-to-refinement process through a progressive, prior-guided architecture. By integrating a Sketching Structure Prior (SSP) for semantic-aligned structure modeling, an Associative Texture Prior (ATP) for dictionary-based texture enhancement, and a Holistic Prior Fusion (HPF) for region-aware feature alignment, PortraitSR effectively addresses the challenges of structural distortion and detail hallucination under severe degradation. Extensive experiments demonstrate that our method achieves state-of-the-art performance in both quantitative metrics and perceptual quality, validating the effectiveness of our coarse-to-fine, dual-prior learning paradigm.

Limitations. While effective, our method still has room for improvement. Its parsing-based guidance may vary under extreme poses or occlusions. Future work may explore more flexible structural estimation and adaptive priors for better scalability.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2022YFA1004100, in part by the Science and Technology Innovation Key R&D Program of Chongqing under Grant No. CSTB2023TIAD-STX0016, in part by the National Natural Science Foundation of China under Grants No. 62472060 and 62221005, in part by the Natural Science Foundation of Chongqing under Grants No. CSTB2024NSCQ-QCXMX0060, CSTB2023NSCQ-LZX0061 and CSTB2023NSCQ-LZX0085, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant No. KJZD-K202300604.

References

Baker, S.; and Kanade, T. 2000. Hallucinating faces. In *Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, 83–88. IEEE.

Bao, Q.; Liu, Y.; Gang, B.; Yang, W.; and Liao, Q. 2023. SCTANet: A spatial attention-guided CNN-transformer aggregation network for deep face image super-resolution. *IEEE Transactions on Multimedia*, 25: 8554–8565.

Chakrabarti, A.; Rajagopalan, A.; and Chellappa, R. 2007. Super-resolution of face images using kernel PCA-based prior. *IEEE Transactions on Multimedia*, 9(4): 888–892.

Chen, C.; Gong, D.; Wang, H.; Li, Z.; and Wong, K.-Y. K. 2020. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30: 1219–1231.

Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2492–2501.

Chen, Z.; Lu, L.; Yuan, Z.; Zhu, Y.; Li, Y.; Yuan, C.; and Deng, W. 2024. Blind face restoration under extreme conditions: Leveraging 3d-2d prior fusion for superior structural and texture recovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1263–1271.

Gao, G.; Xu, Z.; Li, J.; Yang, J.; Zeng, T.; and Qi, G.-J. 2023. CTCNet: A CNN-transformer cooperation network for face image super-resolution. *IEEE Transactions on Image Processing*, 32: 1978–1991.

Gu, Y.; Wang, X.; Xie, L.; Dong, C.; Li, G.; Shan, Y.; and Cheng, M.-M. 2022. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, 126–143. Springer.

Han, S.; Li, S.; Wang, S.; Yuan, L.; Zhang, Y.; and Gao, X. 2025. Deepfake Detection Leveraging Self-Blended Artifacts Guided by Facial Embedding Discrepancy. *IEEE Transactions on Circuits and Systems for Video Technology*.

Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; and Huang, T. S. 2012. Interactive facial feature localization. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, 679–692. Springer.

Leng, J.; Kuang, C.; Li, S.; Gan, J.; Chen, H.; and Gao, X. 2025. Dual-Space Video Person Re-identification. *International Journal of Computer Vision*, 133(6): 3667–3688.

Leng, J.; Wang, J.; Gao, X.; Hu, B.; Gan, J.; and Gao, C. 2022. Icnet: Joint alignment and reconstruction via iterative collaboration for video super-resolution. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6675–6684.

Leng, J.; Wang, J.; Mo, M.; Gan, J.; Lu, W.; and Gao, X. 2024. Difficulty-Guided Variant Degradation Learning for Blind Image Super-Resolution. *IEEE Transactions on Neural Networks and Learning Systems*.

Leng, J.; and Wang, Y. 2022. RCNet: Recurrent collaboration network guided by facial priors for face super-resolution. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06. IEEE.

Li, L.; Zhang, Y.; Yuan, L.; and Gao, X. 2024a. PLGNet: prior-guided local and global interactive hybrid network for face super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 10166–10181.

Li, L.; Zhang, Y.; Yuan, L.; and Gao, X. 2025a. SANet: Face super-resolution based on self-similarity prior and attention integration. *Pattern Recognition*, 157: 110854.

Li, S.; Leng, J.; Gan, J.; Mo, M.; and Gao, X. 2025b. Shape-centered representation learning for visible-infrared person re-identification. *Pattern Recognition*, 111756.

Li, S.; Leng, J.; Kuang, C.; Tan, M.; and Gao, X. 2025c. Video-Level Language-Driven Video-Based Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.

Li, X.; Chen, C.; Zhou, S.; Lin, X.; Zuo, W.; and Zhang, L. 2020. Blind face restoration via deep multi-scale component dictionaries. In *European conference on computer vision*, 399–415. Springer.

Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; and Liu, R. 2024b. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, 270–288. Springer.

Li, X.; Wang, Z.; Zou, Y.; Chen, Z.; Ma, J.; Jiang, Z.; Ma, L.; and Liu, J. 2025d. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7534–7544.

Li, X.; Zhang, S.; Zhou, S.; Zhang, L.; and Zuo, W. 2022. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5904–5917.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Lu, L.; Li, W.; Tao, X.; Lu, J.; and Jia, J. 2021. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6368–6377.

Ma, C.; Jiang, Z.; Rao, Y.; Lu, J.; and Zhou, J. 2020. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5569–5578.

Mei, Y.; Fan, Y.; and Zhou, Y. 2021. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3517–3526.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, L.; Dong, J.; Tang, J.; and Pan, J. 2023. Spatially-Adaptive Feature Modulation for Efficient Image Super-Resolution. In *ICCV*.

Sun, L.; Pan, J.; and Tang, J. 2022. ShuffleMixer: An Efficient ConvNet for Image Super-Resolution. In *Advances in Neural Information Processing Systems*.

Wang, C.; Jiang, J.; Jiang, K.; and Liu, X. 2024a. Low-light face super-resolution via illumination, structure, and texture associated representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5318–5326.

Wang, C.; Jiang, J.; Zhong, Z.; and Liu, X. 2022a. Propagating facial prior knowledge for multitask learning in face super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7317–7331.

Wang, C.; Jiang, J.; Zhong, Z.; and Liu, X. 2023a. Spatial-frequency mutual learning for face super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22356–22366.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9168–9178.

Wang, X.; and Tang, X. 2005. Hallucinating face by eigentransformation. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, 35(3): 425–434.

Wang, Y.; Yue, Y.; Lu, R.; Han, Y.; Song, S.; and Huang, G. 2024b. Efficienttrain++: Generalized curriculum learning for efficient visual backbone training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Z.; and Bovik, A. C. 2002. A universal image quality index. *IEEE signal processing letters*, 9(3): 81–84.

Wang, Z.; Hu, B.; Zhang, M.; Li, J.; Li, L.; Gong, M.; and Gao, X. 2025. Diffusion model-based visual compensation guidance and visual difference analysis for no-reference image quality assessment. *IEEE Transactions on Image Processing*.

Wang, Z.; Li, D.; Zhang, M.; Luo, H.; and Gong, M. 2024c. Enhancing hyperspectral images via diffusion model and group-autoencoder super-resolution network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5794–5804.

Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; and Luo, P. 2022b. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17512–17521.

Wang, Z.; Zhang, J.; Chen, T.; Wang, W.; and Luo, P. 2023b. RestoreFormer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15462–15476.

Wei, F.; Wang, S.; Yang, J.; Sun, X.; Wang, Y.; and Chen, Y. 2023. A composite network model for face super-resolution with multi-order head attention facial priors. *Pattern Recognition*, 139: 109503.

Xie, C.; Ning, Q.; Dong, W.; and Shi, G. 2023. Tfrgan: Leveraging text information for blind face restoration with extreme degradation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2535–2545.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32: 6032–6046.

Yang, C.-Y.; Ma, C.; and Yang, M.-H. 2014. Single-image super-resolution: A benchmark. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 372–386. Springer.

Yang, J.; Dai, T.; Zhu, Y.; Li, N.; Li, J.; and Xia, S.-T. 2025a. Diffusion Prior Interpolation for Flexibility Real-World Face Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9211–9219.

Yang, Y.; Li, S.; Ye, J.; Dong, N.; Li, F.; and Li, H. 2025b. DINOv2 Driven Gait Representation Learning for Video-Based Visible-Infrared Person Re-identification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 8283–8292.

Yin, Y.; Robinson, J.; Zhang, Y.; and Fu, Y. 2020. Joint super-resolution and alignment of tiny faces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12693–12700.

Yuan, Y.; and Yuan, C. 2024. Efficient conditional diffusion model with probability flow sampling for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6862–6870.

Yue, Z.; and Loy, C. C. 2024. Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, M.; Sun, L.; Dong, J.; and Pan, J. 2025. Efficient Video Super-Resolution for Real-time Rendering with Decoupled G-buffer Guidance. In *CVPR*.