Computing and Informatics, Vol. 43, 2024, 369–392, doi: 10.31577/cai_2024_2_369

DIFBFSR: BLIND FACE SUPER-RESOLUTION VIA CONDITIONAL DIFFUSION CONTRACTION

Wei Yu, Zonglin Li, Qinglin Liu, Yufan Chen, Shengping Zhang

School of Computer Science and Technology
Harbin Institute of Technology
Weihai, China
e-mail: 20b903014@stu.hit.edu.cn

Jingbo Lin*

Yantai Institute of Materia Medica
Yantai, China
e-mail: jblin@yimm.ac.cn

Abstract. Blind Face Super-Resolution (BFSR) has recently gained widespread attention, which aims to super-resolve Low-Resolution (LR) face images with complex unknown degradation to High-Resolution (HR) face images. However, existing BFSR methods suffer from two major limitations. First, most of them are trained on synthetic degradation data pairs with pre-defined degradation models, which leads to poor performance due to the degradation mismatch between other unknown complex degradations in real-world scenarios. Second, some methods rely on hand-crafted face priors as constraints, such as facial landmarks and parsing maps, which require additional callouts and laborious hyperparameter tuning for real cases. To tackle these issues, we propose a simple and effective self-supervised cooperative learning framework via a conditional diffusion contraction method for BFSR, dubbed DifBFSR, which establishes the posterior distribution of HR images from degraded LR images with unknown degradation via a powerful diffusion model without expensive supervised training or additional constraint design. Specifically, we first transform the degraded LR face image to an intermediate HR face prediction with degradation-invariant by a simple Super-Resolution module (SRM), which only relies on self-supervised optimization. To enhance the face pre-

* Corresponding author

diction, we propose a Contraction Filter Module (CFM) to gradually contract the restoration error by adaptive dynamic filtering, which efficiently leverages rich nature face prior encapsulated in the pre-trained diffusion model through conditional posterior sampling. Finally, by combining the SRM, CFM, and diffusion model in a self-supervised cooperative learning framework, DifBFSR can robustly handle unknown complex degradations, which favorably avoids the cumbersome training and parameter tuning. Extensive qualitative and quantitative experiments on complex degraded synthetic and real-world datasets show that our method outperforms state-of-the-art BFSR methods.

Keywords: Blind face super-resolution, diffusion model, face restoration, image generation

Mathematics Subject Classification 2010: 68U10

1 INTRODUCTION

Blind Face Super-Resolution (BFSR) has great potential for practical applications in various fields including surveillance, biometrics, and entertainment [1, 2, 3], which aims to restore high-resolution (HR) face images from their low-resolution (LR) counterparts suffering from arbitrary unknown degradation, such as noise, blurring, compression artifacts, and their hybrid forms.

The arbitrary and unknown nature of these degradations renders the problem highly ill-posed, posing a significant challenge for researchers.

To address this challenge, deep learning-based approaches for BFSR have made progress [4, 5, 6], whose main idea is to collect a large number of synthetic LR/HR image pairs by assuming a pre-defined degradation model and employ parameterized deep neural networks to learn the mapping between LR and HR images.

And various constraints are designed to improve the recovery quality, such as L_1, L_2 pixel-level loss, and adversarial loss, ensuring the restoration quality is conducive to predicting realistic and reasonable results.

However, most existing methods fail to generalize to other complex degradation cases and are sensitive to the real unknown degradations, which leads to poor performance where artifacts are often observed in the output. The primary cause of this unadaptability is the mismatch between the synthetic degradation of the training data and the more complex actual degradation in real-world scenarios. This requires collecting real pairs of training data, including each degradation type to retrain the model from scratch to cover all degradation cases in the wild, which is expensive and infeasible in practice.

To mitigate the effect of this mismatch, some existing methods introduce and exploit various hand-crafted face-specific priors as constraints to achieve significant improvements in the quality of facial restoration. Such as facial landmarks [7],

parsing maps [8], and heatmaps [9] are pivotal to constraining the recovery of accurate facial shapes. Furthermore, high-quality face images as reference priors [10, 11] are also introduced to help improve the recovered face details. Nevertheless, these methods require additional callouts or laborious designing while considering so many constraints that make the training unnecessarily complicated [12, 13]. When facing unknown face cases, it is usually necessary to make laborious hyperparameter adjustments to balance these constraints, which is not suitable for practical applications in real-world scenarios. Recent approaches investigate the use of generative priors, such as the Generative Adversarial Network (GAN) priors [14, 15], to help generate realistic details and textures. By incorporating generative priors into the restoration process, these methods can reduce the need for hand-crafted priors and achieve high-quality restoration results even for unknown faces.

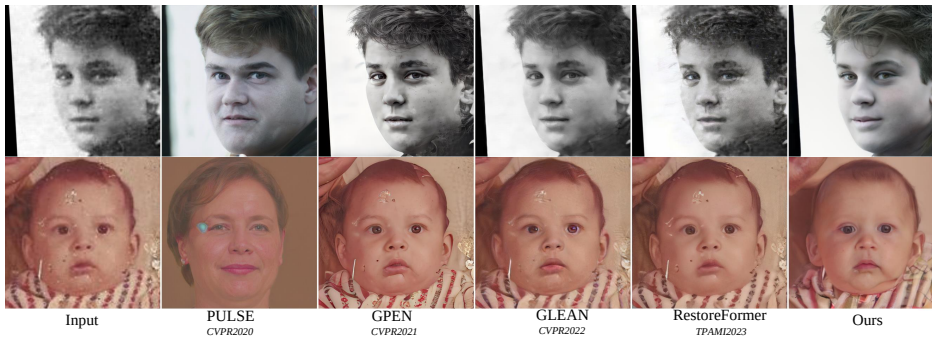

Figure 1. HR results from LR input with severe unknown degradations. Comparative results of recent state-of-the-art methods and our method on a severely degraded face and a real old photo face, which characterize the robust ability of our method to recover from unknown degradation. As a result, regardless of the different degradation types, DifBFSR achieves high-quality restoration results with fewer artifacts and more realistic results than previous techniques [16, 15, 14, 17].

However, due to the instability of the adversarial loss and the lack of diversity in GAN, it may lead to limited inversion performance and restricted performance, resulting in unnatural artifacts and lower fidelity in Figure 1. In summary, the urgent issue in the BFSR task is how to effectively incorporate sufficient facial priors while avoiding unnecessary complexity to achieve accurate and high-quality face restoration.

To overcome these limitations, inspired by the impressive successes of the diffusion model [18] in image reconstruction [19, 20, 21], we propose a self-supervised cooperative learning framework via a novel conditional diffusion contraction approach to explore its potentials for BFSR, which establishes the posterior distribution of HR face images from degraded LR face images with unknown degradation via a powerful diffusion model.

Different from existing methods, our method does not need expensive degradation supervised training from scratch or additional complex constraint design but sufficiently leverages the prior knowledge contained in the pre-trained diffusion model to mitigate bias in face restoration.

Specifically, we first transform the degraded LR face image to an intermediate HR face prediction with degradation-invariant by a simple Super-Resolution Module (SRM), which only relies on self-supervised optimization learning without additional training.

To enhance the face prediction and remove degradations, we propose a Contraction Filter Module (CFM) to gradually contract the restoration error by adaptive dynamic filtering, which leverages rich nature face prior information encapsulated in a pre-trained diffusion model through conditional posterior sampling.

Finally, by combining the SRM, CFM, and diffusion model in the self-supervised cooperative learning framework, DifBFSR can robustly handle unknown complex degradations in the iterative diffusion inverse process.

In conclusion, DifBFSR leverages the rich image priors and strong generative capability of a pre-trained diffusion model to deal with unknown and complex degradations in face images, which not only sufficiently leverages the powerful generation capability of the diffusion model, but also favorably avoids the cumbersome degradation training and parameter tuning.

Extensive qualitative and quantitative experiments in various settings show that our method outperforms state-of-the-art BFSR methods on complex degraded synthetic and real-world datasets. Moreover, DifBFSR also generalizes well for natural images or synthesized images with arbitrary degradations from various scenes out of the distribution of the FFHQ [22] training set.

In summary, the contributions of this work are as follows:

1. We propose a novel self-supervised cooperative learning framework for BFSR, which establishes the posterior distribution of HR face images as a Markov chain starting from LR face images and optimize them via conditional diffusion contraction strategy, which allows us to guarantee the fidelity and realness of the restoration.

2. We leverage face image prior encapsulated in a pre-trained diffusion model to robustly cope with complex and unknown degradations without expensive synthetic degradation supervised training or additional hand-crafted face constraint design.

3. Extensive qualitative and quantitative experiments on heavily degraded synthesis and real-world face datasets demonstrate that our DifBFSR outperforms state-of-the-art BFSR methods.

2 RELATED WORK

2.1 Blind Face Super Resolution

Blind face super-resolution aims to restore high-resolution face images from low-resolution face images suffering from unknown degradations, such as blurring, noise, compression, etc.

In the past few years, with the development of deep learning, the restoration quality of face images has significantly improved. Most existing algorithms [23, 24] directly learn a mapping from the low-resolution images to high-resolution images with a pixel loss constraint.

DFDNet [8] learns a deep dictionary network by the L_2 loss to guide the degraded restoration process, and adaptively fuses dictionary features into the input to adopt a multi-scale dictionary in a progressive manner to achieve coarse-to-fine face restoration.

VQFR [5] further introduces vector quantization technology to extract high-quality low-level feature libraries from high-quality faces, which can help restore realistic facial details.

CodeFormer [25] proposes a Transformer-based network to model the global composition and context of face information, which acquires rich expressiveness and enhances the adaptability to different degradations.

Despite remarkable results in terms of PSNR, training with only pixel constraints often leads to perceptually unconvincing output with severe over-smoothing artifacts.

To alleviate this problem, some existing methods also exploit face-specific priors to further constrain the restored solution, e.g., face landmarks [7, 26], facial components [8, 27], and generative priors [28, 14].

Besides, nature image priors in generative adversarial networks like pre-trained StyleGAN [22, 29] are used to produce more realistic textures and details. Recent BFSR methods also introduce the adversarial loss [30] and the perceptual loss [31] to achieve more realistic results.

PULSE [16] guarantees the authenticity of the output by guiding the exploration of the latent space of the generative model in a self-supervised manner. GPEN [14] embeds GAN as a priori decoder into a U-net network, and uses its deep and shallow features to control the global facial structure, local facial details and background of the reconstructed image.

Unlike the popular GAN inversion optimization, GLEAN [15] only requires one forward pass to generate upscaled images and utilizes multi-resolution skip connections to improve image fidelity and texture fidelity.

GFPGAN [28] utilizes rich and diverse prior knowledge encapsulated in pre-trained face GAN to simultaneously restore facial details and enhance color based on spatial feature transformation.

RestoreFormer [32, 17] utilize a fully-spatial attention to model face contextual information and introduce the interplay with the prior. However, due to the no-

torious instability of the adversarial training loss, unnatural artifacts can still be observed in the output.

As discussed above, most of the existing BFSR methods rely on the ideal face prior or specific degradation space and are not flexible enough to deal with volatile real degradation and application requirements.

To avoid cumbersome training and parameter tuning, our approach leverages the rich image priors from a pre-trained generative diffusion model to handle unknown complex degradation without complicated adversarial losses.

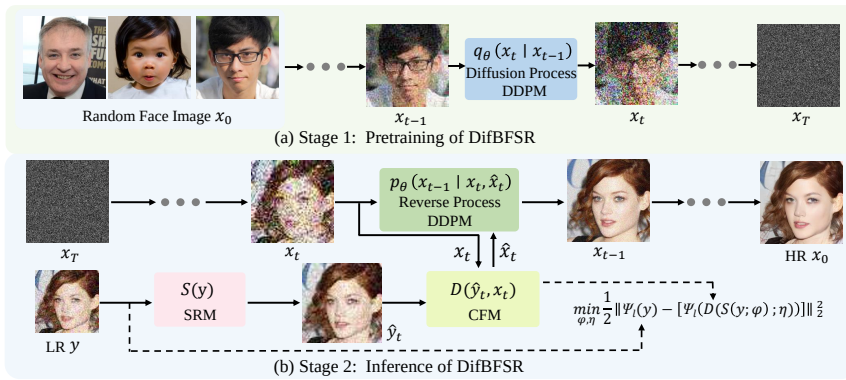

Figure 2. The overview of the proposed DifBFSR. a) The pre-training of DifBFSR is the forward process of the denoising diffusion probability model (DDPM), we take random face images x_0 in FFHQ [22] dataset as input and output the random noise x_T to learn the natural face detail texture information as prior. b) The inference of DifBFSR is the reverse process of the DDPM, we utilize the Contraction Filter Module (CFM) to leverage the texture prior encapsulated in the pre-trained DDPM to enhance the coarse face prediction $\hat{y}_t$ extracted from the Super-Resolution Module (SRM). It is worth noting that, our networks have not been trained on any degraded datasets and only utilize a single degraded image for self-supervised restoration during the inference stage.

2.2 Denoising Diffusion Probability Model

Inspired by non-equilibrium statistical physics, Sohl-Dickstein et al. [18] propose Denoising Diffusion Probability Model (DDPM) as a new generation method, which has achieved faster performance than Generative Adversarial Networks (GAN) in unconditional image generation.

Diffusion models have recently attracted significant interest in the community of low-level vision tasks due to their strong performance of generative models.

Attributed to its sound and perfect theoretical support, the diffusion model has achieved an impressive success in the image generation and reconstruction, such as super-resolution [21], inpainting [33], and colorization [34].

SR3 [21] modifies DDPM to be conditioned on low-resolution images through channel-wise concatenation, which first utilizes it for super-resolution. SRDiff [35] provides diverse and realistic predictions by converting Gaussian noise into LR-conditioned Markov chains and introduces residual predictions to speed up convergence.

ILVR [20] generates high-quality images and controls the generation by iterative latent variable refinement via a conditioning sampling method. DDRM [19] first proposes to solve the linear restoration inverse problem through SVD decomposition based on a pre-trained diffusion model.

DDNM [36] solves various linear restoration tasks well by the range-null space decomposition. LDM [37] performs super-resolution in a similar concatenation manner but in a low-dimensional latent space. However, these methods fix the degradation as a single fixed degradation, which is not suitable for other complex degradation cases.

3 PROPOSED METHOD

In this study, we aim to exploit a pre-trained Denoising Diffusion Probability Model (DDPM) as an effective prior for blind face super-resolution.

The proposed DifBFSR is shown in Figure 2, which combines a common U-Net DDPM [38], a simple Super-Resolution Module (SRM), and a novel Contraction Filter Module (CFM) in a unified self-supervised cooperative learning framework.

Given an input LR face image y with unknown degradation, we first transform its fidelity to a lost intermediate face prediction $\hat{y}_t$ with degradation-invariant by the SRM to refine the generative process.

Then, to enhance the face prediction $\hat{y}_t$, CFM leverages it and a diffusion sampling image x_t gradually contracts the restoration error by adaptive dynamic filtering and fuses them to obtain the refined HR face images $\hat{x}_t$ with rich nature face detail information.

As a result, we approximate this posterior distribution through conditional posterior sampling from the distribution $p(x_{t-1}|x_t, \hat{x}_t)$, where x_{t-1} is a diffused version of the desirable HR image $\hat{x}_t$, followed with a reverse Markov chain that estimates the final output $\boldsymbol{x}_0$ with higher resolution and high-quality details.

Our framework first infers an intermediate HR face image from the LR image that maintains coarse structural information, providing a suitable initial distribution and image resolution for subsequent diffusion sampling, and then uses the pre-trained face diffusion model to enrich face details to infer the ideal HR face image.

There are several advantages to our approach:

1. We pre-train the diffusion model on natural faces, which can handle real unknown degradations without designing and synthesizing complex degradation data pairs for additional training.
2. We take an advantage of the pre-trained diffusion model through self-supervised collaborative learning without complex hand-crafted face priors.

3. We adopt the diffusion conditional posterior sampling based on the transition prediction, which is more accurate and efficient than the Markov chain directly from y to x_0.

In the following, we detail the proposed self-supervised cooperative learning approach and conditional diffusion contraction strategy.

3.1 Self-Supervised Cooperative Learning

To guarantee that the image fidelity is maintained and the perceptual quality of the image is improved, we design the image Super-Resolution Module (SRM) and the Contraction Filter Module (CFM) to solve these problems.

In particular, we design a unified framework to alternately optimize these two modules via self-supervised cooperative learning, which first improve the resolution of the input y and obtaining $\hat{y}_t$ through SRM, and the CFM maps degradation-invariant $\hat{y}_t$ and natural face x_t to high-quality face image $\hat{x}_t$ and iteratively generate the final high-resolution high-quality x_0 face through DDPM.

Our target is to maximize the likelihood

$$\begin{aligned} p_{\phi,\eta}(\hat{\mathbf{x}}_t|\mathbf{y}) &= \int p_\eta(\hat{\mathbf{x}}_t|\mathbf{x_t}, \hat{\mathbf{y}}_\mathbf{t}) p_\phi(\hat{\mathbf{y}}_\mathbf{t}|\mathbf{y}) \mathrm{d}\hat{\mathbf{x}}_t \\ &= E_{\hat{\mathbf{x}}_0 \sim p_\phi(\hat{\mathbf{y}}_t)|\mathbf{y})} \left[p_\eta(\hat{\mathbf{x}}_t|\mathbf{x_t}, \hat{\mathbf{y}}_\mathbf{t}) \right], \end{aligned} \tag{1}$$

where $p_\phi(\cdot)$ corresponds to the first stage SRM, and $p_\eta(\cdot)$ corresponds to the second stage CFM, where $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ represent the learned parameters, respectively. $\hat{x}_t$ denotes the estimated x_0 at step t. Instead of directly learning the mapping from y to x_0, we prospose the conditional diffusion contraction approach. $\boldsymbol{\phi}, \boldsymbol{\eta}$ are the expected model parameters obtained by self-supervised learning without any hyper-parameters tuning.

The optimization objective is reformulated into a minimization problem as follows

$$\phi, \eta = \arg\min_{\phi,\eta} \mathcal{L}_s(\phi, \eta|\mathbf{y}). \tag{2}$$

The total optimization objective $\mathcal{L}_s$ is formulated by

$$\mathcal{L}_s = \frac{1}{2} \| \psi_l(\mathbf{y}) - [\psi_l(D(S(\mathbf{y}; \phi)); \eta)] \|_2^2, \tag{3}$$

where ψ_l denotes low-pass filter implemented by the CFM to obtain low-frequency information of a face image. $G(\cdot)$ and $D(\cdot)$ represent SRM and CFM modules, respectively, where the model parameters $\boldsymbol{\psi}, \boldsymbol{\eta}$ are updated as follows

$$\begin{aligned} \tilde{W_{\boldsymbol{\phi}}} &= W_{\boldsymbol{\phi}} - \alpha \frac{\partial}{\partial W_{\boldsymbol{\phi}}} \mathcal{L}_s, \\ \tilde{W_{\boldsymbol{\eta}}} &= W_{\boldsymbol{\eta}} - \beta \frac{\partial}{\partial W_{\boldsymbol{\eta}}} \mathcal{L}_s, \end{aligned} \tag{4}$$

where $\tilde{W}.$ denotes the updated parameters, α and β are the learning rates.

We alternately update these modules to transition degradation distribution by alternating back-propagation via the Adam algorithm.

The stepwise transforms the degenerate distribution to approach the unknown degradation of LR, which achieve the conditional diffusion contraction via iterative updates.

Super-Resolution Module. Our Super-Resolution Module (SRM) adopts a simple ResNet upsampling structure [39] that takes degraded LR face images as input and generates HR face predictions as output.

The main goal of SRM is to improve the image resolution to provide a suitable initialization distribution for the subsequent diffusion step.

In addition to improving the resolution, our SRM also aims to be degradation-invariant.

This means that it can generate an HR face prediction that is robust to arbitrary degradation, such as noise, blurriness, or compression artifacts.

To achieve this, our SRM relies on self-supervised optimization learning, which allows it to adapt to different types of degradation without an additional training.

This optimization process is performed without any additional training data or supervision, making it more efficient and flexible than other techniques that rely on supervised learning.

Overall, through adaptive iterative updates, our SRM provides a suitable initialization distribution and transforms degraded LR face images into desired HR resolution predictions, which are both robust to unknown complex degradation and avoid sampling directly from the Gaussian noise to effectively speed up the diffusion restoration step.

3.2 Conditional Diffusion Contraction

Simulating every possible degradation in the real world is obviously difficult and expensive.

To remove the reliance on synthetic degradation, we leverage the well-performing DDPM to remove the degradation from the input by the conditional diffusion contraction strategy in the iterative reverse restoration process, which consists of the contraction filter module and the conditional diffusion sampling.

Contraction Filter Module. To generate clean and natural face images while maintaining semantic information, we propose the Contraction Filter Module (CFM) to gradually contract the restoration error by adaptive dynamic filtering, which provides detailed guidance for the restoration process and robustly removes degradation.

The restoration process defined above involves an intermediate time step t that makes the distance between $\hat{y}_t$ and x_t very close, especially in the low-frequency part.

Among them, the high-frequency of the diffusion result x_t and the degenerate-invariant intermediate $\hat{y}_t$ contain high-quality details and blurred details due to degradation, respectively.

To adaptively fuse high- and low-frequency information to utilize the rich face details, we design an adaptive dynamic filtering strategy.

After each transition from x_{t+1} to x_t, we replace the low-frequency part of x_t with that of $\hat{y}_t$ because they are close in distribution, which is formulated as

$$\hat{\mathbf{x}}_t = \psi_l(\hat{\mathbf{y}}_t) + (I - \psi_l)(\mathbf{x}_t), \tag{5}$$

where $\psi_l(\cdot)$ denotes a low-pass filter implemented by downsampling and upsampling operations in CFM.

We remain the low-frequency part of $\hat{y}_t$ to ensure that the result $\hat{x}_t$ shares basic semantics with $\hat{y}_t$, and remove the high-frequency part of $\hat{y}_t$ because it contains little useful information due to the degradation-invariance, which guarantees the fidelity of the output image.

We remain the high-frequency part of x_t to leaverage high-frequency natural texture information in the diffusion model, ensuring the perceptual quality of the output.

Through iterative refinement of the network, the degradation error is progressively contracted, and the balanced fusion of low-frequency information of HR prediction images $\hat{y}_t$ and high-frequency information of natural face images x_t is dynamically realized.

Attributed to this diffusion contraction mechanism, the model has a greater error tolerance for degradation, thus we can handle complex and severe degradation by simply training on natural images, avoiding complex constraints designing.

Conditional Diffusion Sampling. With the output of CFM, to further enhance the image fidelity, we propose a conditional diffusion posterior sampling via degradation contraction by leveraging rich and diverse natural face priors encapsulated in a pre-trained DDPM [40].

As shown in Figure 2, DDPM consists of a forward process and a reverse process.

The forward process is a Markov chain that gradually corrupts the nature data by repeatedly adding Gaussian noise, which is equivalent to a degradation process with the following formulation

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I}\right), \tag{6}$$

where β_t is a pre-defined noise variance. t denotes the timestep.

We leverage this process to pre-train the DDPM model to learn rich natural face information.

The reverse process can iteratively sample real images $\mathbf{x}_t$ with rich realistic textures from random Gaussian noises $\mathbf{x}_T$, which is equivalent to a restoration process.

Benefiting from generative priors in the pre-trained DDPM, denoising within the DDPM manifold naturally normalizes the realism and fidelity of sampled images.

Therefore, to extract the natural face distribution from the DDPM $\epsilon_{\theta_s}(\mathbf{x}_t, t)$, we perform reverse diffusion from $\mathbf{x}_T$ by the following formulation

$$\mathbf{x}_t = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_{t+1} - \epsilon_{\theta_s}(\mathbf{x}_t, t) \sqrt{1 - \bar{\alpha}_t} \right), \tag{7}$$

where $\bar{\alpha}_t$ is diffusion hyper-parameter, $\mathbf{x}_t$ denote the estimated $\mathbf{x}_0$ at time step $t \in [0, T]$, which are equivalent to the original DDPM sampling [38]. $\mathbf{x}_t$ obeys the real face distribution with rich detail texture information.

Following the previous work [19, 36], we reparameterize the mean as

$$\mu_t(\mathbf{x_t}, \hat{\mathbf{x}}_\mathbf{t}) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\mathbf{x_t} + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\hat{\mathbf{x}}_\mathbf{t}, \tag{8}$$

where the mean $\mu_t(\cdot)$ is the target we want to estimate by the DDPM. β_t, $1 - \alpha_t$ are hyper-parameters in the reverse diffusion process, that satisfy the condition $\beta_t = 1 - \alpha_t$.

To maximize the preservation of face details and overall image quality, we model the desired high-frequency texture information during the reverse diffusion process via conditional posterior sampling.

We aim to learn a conditional probability from x_t to x_{t-1}, which is defined as the following Gaussian distribution

$$p_\theta(\mathbf{x_{t-1}}|\mathbf{x_t}, \hat{\mathbf{x}}_\mathbf{t}) = \mathcal{N}\left(\mathbf{x_{t-1}}; \boldsymbol{\mu}_\theta(\mathbf{x_t}, \hat{\mathbf{x}}_\mathbf{t}), \boldsymbol{\Sigma}_\theta(\mathbf{x_t}, \mathbf{t})\right), \tag{9}$$

where θ is the learnable parameter. $\boldsymbol{\Sigma}_\theta(\cdot)$ represents the same standard deviation as the original DDPM.

With such a learned probability distribution, we can approximate the data distribution $q(\boldsymbol{x}_0)$ via the following marginal distribution

$$p_\theta(\mathbf{x_0}) = \int \mathbf{p}(\mathbf{x_T}) \prod_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{p}_\theta(\mathbf{x_{t-1}}|\mathbf{x_t}) \mathrm{d}\mathbf{x_{1:T}}, \tag{10}$$

where $p(\mathbf{x_T}) = \mathcal{N}(\mathbf{x_T}; \mathbf{0}, \mathbf{I})$ denote the standard normal distribution.

We implement simple conditional sampling from the diffusion model to achieve blind degradation restoration.

Therefore, we adopt the conditional diffusion contraction strategy, where the model parameters of SRM and CFM are randomly initialized and optimized in the inverse process.

By combining the SRM, CFM, and DDPM in the self-supervised cooperative learning framework, DifBFSR can robustly handle unknown complex degradations in the iterative diffusion inverse process, which benefited from the natural face image texture prior in the diffusion model.

Figure 3. Comparative results of the state-of-the-art methods and the proposed method on three complex synthetic examples in the CelebA-Test dataset

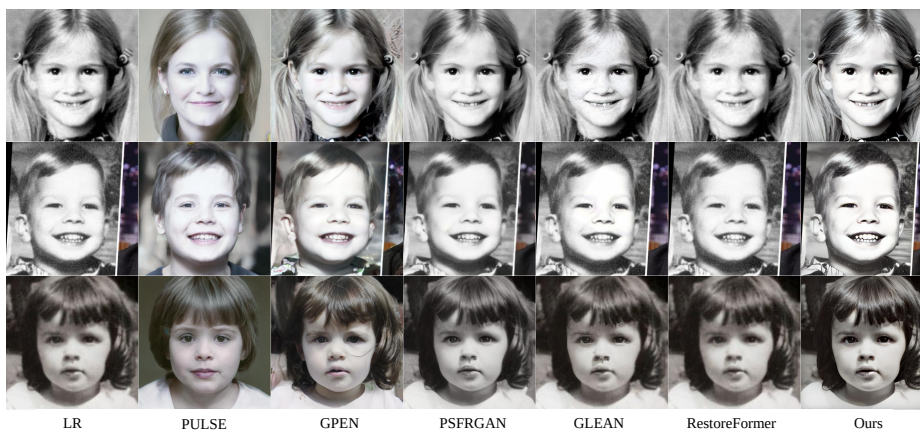

Figure 4. Comparative results of recent state-of-the-art methods and the proposed method on three real degraded face examples in CelebChild-Test dataset

Figure 5. Comparative results of recent state-of-the-art methods and the proposed method on three real old photo examples in Webphoto-Test dataset

4 EXPERIMENTS

Following existing BFSR methods, we conduct extensive experiments to verify the effectiveness of DifBFSR on both one synthetic and two real-world datasets at scale×4. To evaluate different methods comprehensively, we adopt seven quantitative metrics including: three reference evaluation metrics PSNR, SSIM [41] and LPIPS [31], four no-reference evaluation metrics NIQE [42], PI [43], NRQM [44], and FID [45].

Specifically, PSNR and SSIM are commonly used to evaluate image fidelity. LPIPS is a learned perceptual similarity metric calculated by the VGG [46] network.

NIQE, PI, and NRQM can assess the naturalness and perception quality of face images without reference images.

As for FID, it is the KL divergence between the feature distributions of the restored images and the ground truth nature images to quantitatively evaluate the overall quality of the restored image.

4.1 Compared Methods

We compare DifBFSR with recent state-of-the-art BFSR methods: PULSE [16], GPEN [14], PSFRGAN [6], and GLEAN [15].

Note that similar to these GAN-Based methods, our method only relies on a pre-trained network trained by FFHQ [22] dataset.

As for these methods, we adopt their official codes and pre-trained models for testing.

4.2 Implementation Details

To justly verify the ideas in this paper, we consider a simple and plain network structure without complex design.

For the super-resolution module, we follow the architecture in SRResNet [47] with five convolution layers of skip convolution and a upsampling layer.

To achieve adaptive dynamic filtering, the architecture of the contraction filter module consists of two iterations of convolutional layer and nonlinear LeakyRelu [48] layer.

DDPM with a common U-Net architecture obtained from [40] that are dependently trained on FFHQ [22] dataset, which contains 70 000 high-quality face image.

The learning rates for α and β are set as 2e-3 and 5e-3, respectively.

The number of iterations for the framework is 100 steps.

Metrics	Input	PULSE [16]	GPEN [14]	PSFRGAN [6]	GLEAN [15]	RestoreFormer [17]	Ours
PSNR↑	24.05	19.07	24.36	24.02	25.14	24.78	25.27
SSIM↑	0.6167	0.6216	0.6771	0.6723	0.6951	0.6841	0.7349
LPIPS↓	0.2854	0.2736	0.2467	0.2362	0.2361	0.2429	0.2276
NIQE↓	13.39	3.83	5.97	5.38	5.14	5.94	5.81
PI↓	10.04	2.89	5.46	4.73	4.74	5.54	5.49
NRQM↑	3.21	8.31	4.89	5.69	5.62	4.91	5.11
FID↓	252.55	107.15	102.92	108.10	98.01	98.33	77.46

Table 1. Comparison of different methods on the synthetic CelebA-Test dataset. The optimal and suboptimal results are highlighted in red and blue.

4.3 Experiments on the Synthesis Dataset

Following previous works [24, 3], we evaluate DifBFSR on one synthetic dataset, denoted as CelebA-Test, contains 100 HR images from CelebA-HQ [49], and the corresponding LR images are synthesized by the following formulation.

Mathematically, a LR image $\mathbf{y}$ is degraded from a HR image $\mathbf{x}$ by the designed degradation $D_{k,s,n,j}$ as

$$\mathbf{y} = D_{k,s,n,j}(\mathbf{x}) = [(\mathbf{x} \otimes k)\downarrow_s + n]_j, \tag{11}$$

where $\otimes$ represents the convolution operation, k denotes the Gaussian blur kernel, $\downarrow_s$ is the downscale operation, n is an additive noise and j denotes the JPEG compression.

To quantitatively evaluate the BFSR methods, k is randomly selected from 8 isotropic Gaussian kernels with a range of width $[1.8, 3.2]$, $\downarrow_s$ is bicubic downsampling with 4 scale and the noise is selected from white Gaussian with a range of noise level $[0, 30]$.

The JPEG compression is selected from a range of compression quality factor $[50, 100]$, which is a nonlinear operator due to the discrete cosine transform [50].

In particular, we set a total of 8 degradation cases from hard to easy to simulate a variety of complex degradations in real-world scenarios.

During testing, we take the average of the restoration results in all degeneration cases as the final result.

Quantitative experiment. The quantitative evaluation results of different SR methods on the synthesis datasets are shown in Table 1, and several conclusions can be drawn.

Note that PSNR and SSIM are used to measure fidelity, and higher values mean that the restored image is more similar to ground truth in detailed texture.

Firstly, as expected, PULSE and PSFRGAN have the highest perceptual quality metrics (i.e., NIQE, PI, and NRQM), and although it performs well visually, it does not perform well in fidelity due to the instability of GAN and the inaccuracy of latent space search.

Then, GPEN and GLEAN achieve promising results, as they are trained on similar degradation, but do not perform as well as DifBFSR in the case of unseen degradation.

The proposed DifBFSR achieves the best overall PSNR results and yields the best overall SSIM results, which benefits from a strong diffusion generation prior and an efficient trade-off of perception and fidelity.

We summarize the comparative results that the proposed DifBFSR achieves the best performance across all reference metrics, indicating its effectiveness and superiority in the task of BFSR.

Qualitative experiment. For easy visualization, three typical examples of CelebA dataset are shown in Figure 3.

In the first slightly degraded example, most methods except PULSE are able to recover realistic images. PULSE cannot guarantee fidelity because its optimization cannot find the correct latent code in GAN by inversion.

The second and third examples exhibit more severe degradation, most of the comparison methods produce noticeable artifacts.

It can be seen that GPEN produces blurry artifacts, GLEAN fails to remove the noise and recover sharp edges, and PSFRGAN fails to remove the compression artifacts.

In comparison, our DifBFSR produces better visual results than the other methods and makes a better compromise between artifact removal and detail preservation.

Experiments on synthetic datasets demonstrate that DifBFSR can perform robustly on images corrupted by challenging diverse degradations and achieves satisfactory results.

Metrics	PULSE	GPEN	PSFRGAN	GLEAN	RestoreFormer	Ours
NIQE↓	4.18	6.16	5.66	4.74	9.16	4.36
PI↓	3.42	5.75	5.28	4.50	7.31	4.17
NRQM↑	7.78	4.67	5.08	5.91	4.34	6.19
FID↓	108.45	121.78	106.16	114.50	122.06	98.92

Table 2. Comparison of different methods on the real-world CelebChild dataset. The optimal and suboptimal results are highlighted in red and blue.

4.4 Experiments on the Real-World Dataset

As for the real-world datasets, we consider two typical ones with different degrees of degradation, namely CelebChild-Test, WebPhoto-Test. CelebChild-Test consists of 100 degraded child face images in the wild. WebPhoto-Test is made up of 407 real old photo images crawled from the internet.

Some of them are old photos with real severe degradations, and are thus suitable to test the robustness of different methods under severe degradations.

Metrics	PULSE	GPEN	PSFRGAN	GLEAN	RestoreFormer	Ours
NIQE↓	3.92	6.54	5.58	5.38	12.75	4.46
PI↓	3.24	5.77	4.74	4.56	9.71	4.42
NRQM↑	7.94	4.93	6.16	6.65	3.13	6.06
FID↓	100.47	96.11	89.19	100.12	101.70	81.42

Table 3. Comparison of different methods on the real-world WebPhoto dataset. The optimal and suboptimal two results are highlighted in red and blue.

Quantitative experiment. In experiments on two real-world datasets, we mainly employ FID as a quantitative metric since its ground truth is not accessible.

We estimate the feature statistics of HQ images and the recovered images separately in the FFHQ [22] dataset, and then compute the KL divergence as the FID.

The ground truth of the real-world datasets is not available, we adopt the non-reference image quality assessment (IQA) metrics including NIQE, PI, and NRQM for perception quantitative evaluation.

Note that NIQE and PI are reference-free metrics, and lower values mean that the restored image has higher perceptual visual quality. The results of the comparison are summarized in Tables 2 and 3. We can observe that DifBFSR achieves the best performance on both CelebChild-Test and WebPhoto-Test datasets, it also outperforms most state-of-the-art BFR methods.

Although PULSE achieves the best non-reference metrics, the resulting images have completely lost their fidelity and are not worthy of comparison.

As can be seen, the proposed DifBFSR achieves the best perceptual metrics at all datasets, which indicates that our results are more in line with human visual senses and satisfy the semantic identity fidelity of the input.

Qualitative experiment. To further analyze the characteristics of our method, we show three representative examples of two datasets in Figure 4, and more visualizations are given in Figure 5.

It is again observed that DifBFSR gives better recovery results, especially in the second and third examples with severe unknown degeneracy in Figure 4.

We notice that our DifBFSR tends to produce realistic texture detail, e.g. hair and eyelashes, which cannot be restored by other methods.

The main reason is that DifBFSR reconstructs the complete structure and high-fidelity texture separately through conditional diffusion contractions, and takes advantage of more realistic face image priors in the diffusion model.

To further verify the robustness of DifBFSR, we compare it under real severe degradation in old photos shown in Figure 5.

As can be seen, DifBFSR repairs damaged old photos relying on the natural diffusion face prior and restores natural real faces while maintaining identity information. And other methods have lower fidelity, such as PULSE and GPEN.

These results confirm the effectiveness and robustness of our DifBFSR for unknown severe degradations in real-world scenarios.

4.5 Ablation Studies

Model ID	SRM	CFM	PSNR↑	SSIM↑	LPIPS↓
#1	✗	✗	20.38	0.5021	0.5052
#2	✗	✓	22.74	0.5964	0.4123
#3	✓	✗	24.01	0.6896	0.3561
#4	✓	✓	**25.27**	**0.7349**	**0.2276**

Table 4. The ablation study of the individual components

To verify the effectiveness of the proposed DifBFSR method, we conduct ablation experiments to study the impact of the Super-Resolution Module (SRM) and Contraction Filter Module (CFM) on the CelebA-Test dataset.

As shown in Table 4, the results of ID #2 and #3 indicate that our CFM and SRM achieve average improvements of 2.36 dB and 3.63 dB in terms of PSNR. The results of #4 show that the PSNR/SSIM values of 20.38/0.5021 from baseline are significantly increased to 25.27/0.7349 dB by the combination of the two modules.

According to the results, we can conclude that by relying on self-supervised optimization learning, our SRM can find intermediate faces with degradation invariance without complex parameter adjustment and pre-training. By mining face

prior information, our CFM can accurately eliminate complex unknown degradation to enhance the results through adaptive dynamic filtering.

5 CONCLUSIONS

In this paper, we propose a self-supervised cooperative learning framework via a conditional diffusion contraction method for Blind Face Super-Resolution, dubbed DifBFSR, which establishes the posterior distribution of HR face images from degraded LR face images with unknown degradation via a powerful diffusion model without expensive degradation supervised training or additional complex constraint design.

SRM transforms the degraded LR face image to an intermediate HR face prediction with degradation-invariant, which only relies on self-supervised optimization learning without additional training.

CFM gradually contract the restoration error by adaptive dynamic filtering to enhance the face prediction, which leverages rich nature face prior information encapsulated in a pre-trained diffusion model through conditional posterior sampling.

By combining the SRM, CFM, and DDPM in the self-supervised cooperative learning framework, DifBFSR can robustly handle unknown complex degradations, which favorably avoids the cumbersome training and parameter tuning.

Extensive qualitative and quantitative experiments in various settings show that our method outperforms state-of-the-art BFSR methods on complex degraded synthetic and real-world datasets.

As a result, DifBFSR provides a new way to solve BFSR for real-world applications, we hope that this work could inspire more robust diffusion-based restoration methods in the future.

REFERENCES

[1] Lin, J.—Zhou, T.—Chen, Z.: Multi-Scale Face Restoration with Sequential Gating Ensemble Network. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, No. 1, doi: 10.1609/aaai.v32i1.12296.

[2] Pan, X.—Zhan, X.—Dai, B.—Lin, D.—Loy, C. C.—Luo, P.: Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol. 44, 2022, No. 11, pp. 7474–7489, doi: 10.1109/TPAMI.2021.3115428.

[3] Liu, A.—Liu, Y.—Gu, J.—Qiao, Y.—Dong, C.: Blind Image Super-Resolution: A Survey and Beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 5, pp. 5461–5480, doi: 10.1109/TPAMI.2022.3203009.

[4] Feihong, L.—Hang, C.—Kang, L.—Qiliang, D.—Jian, Z.—Kaipeng, Z.—Hong, H.: Toward High-Quality Face-Mask Occluded Restoration. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Vol. 19, 2023, No. 1, Art. No. 24, doi: 10.1145/3524137.

[5] Gu, Y.—Wang, X.—Xie, L.—Dong, C.—Li, G.—Shan, Y.—Cheng, M. M.: VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., Hassner, T. (Eds.): Computer Vision – ECCV 2022. Springer, Cham, Lecture Notes in Computer Science, Vol. 13678, 2022, pp. 126–143, doi: 10.1007/978-3-031-19797-0_8.

[6] Chen, C.—Li, X.—Yang, L.—Lin, X.—Zhang, L.—Wong, K. Y. K.: Progressive Semantic-Aware Style Transformation for Blind Face Restoration. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11891–11900, doi: 10.1109/CVPR46437.2021.01172.

[7] Chen, Y.—Tai, Y.—Liu, X.—Shen, C.—Yang, J.: FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2492–2501, doi: 10.1109/CVPR.2018.00264.

[8] Li, X.—Chen, C.—Zhou, S.—Lin, X.—Zuo, W.—Zhang, L.: Blind Face Restoration via Deep Multi-Scale Component Dictionaries. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12354, 2020, pp. 399–415, doi: 10.1007/978-3-030-58545-7_23.

[9] Yu, X.—Fernando, B.—Ghanem, B.—Porikli, F.—Hartley, R.: Face Super-Resolution Guided by Facial Component Heatmaps. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11213, 2018, pp. 217–233, doi: 10.1007/978-3-030-01240-3_14.

[10] Li, X.—Li, W.—Ren, D.—Zhang, H.—Wang, M.—Zuo, W.: Enhanced Blind Face Restoration with Multi-Exemplar Images and Adaptive Spatial Feature Fusion. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2703–2712, doi: 10.1109/CVPR42600.2020.00278.

[11] Li, X.—Liu, M.—Ye, Y.—Zuo, W.—Lin, L.—Yang, R.: Learning Warped Guidance for Blind Face Restoration. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11217, 2018, pp. 278–296, doi: 10.1007/978-3-030-01261-8_17.

[12] Yue, Z.—Loy, C. C.: DifFace: Blind Face Restoration with Diffused Error Contraction. CoRR, 2022, doi: 10.48550/arXiv.2212.06512.

[13] Gao, G.—Xu, Z.—Li, J.—Yang, J.—Zeng, T.—Qi, G. J.: CTCNet: A CNN-Transformer Cooperation Network for Face Image Super-Resolution. IEEE Transactions on Image Processing, Vol. 32, 2023, pp. 1978–1991, doi: 10.1109/TIP.2023.3261747.

[14] Yang, T.—Ren, P.—Xie, X.—Zhang, L.: GAN Prior Embedded Network for Blind Face Restoration in the Wild. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 672–681, doi: 10.1109/CVPR46437.2021.00073.

[15] Chan, K. C. K.—Wang, X.—Xu, X.—Gu, J.—Loy, C. C.: GLEAN: Generative Latent Bank for Large-Factor Image Super-Resolution. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14240–14249, doi: 10.1109/CVPR46437.2021.01402.

[16] Menon, S.—Damian, A.—Hu, S.—Ravi, N.—Rudin, C.: PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2434–2442, doi: 10.1109/CVPR42600.2020.00251.

[17] Wang, Z.—Zhang, J.—Chen, T.—Wang, W.—Luo, P.: RestoreFormer++: Towards Real-World Blind Face Restoration from Undegraded Key-Value Pairs. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 12, pp. 15462–15476, doi: 10.1109/TPAMI.2023.3315753.

[18] Sohl-Dickstein, J.—Weiss, E.—Maheswaranathan, N.—Ganguli, S.: Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In: Bach, F., Blei, D. (Eds.): Proceedings of the 32nd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research (PMLR), Vol. 37, 2015, pp. 2256–2265, https://proceedings.mlr.press/v37/sohl-dickstein15.pdf.

[19] Kawar, B.—Elad, M.—Ermon, S.—Song, J.: Denoising Diffusion Restoration Models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.): Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Curran Associates, Inc., 2022, pp. 23593–23606, https://proceedings.neurips.cc/paper_files/paper/2022/file/95504595b6169131b6ed6cd72eb05616-Paper-Conference.pdf.

[20] Choi, J.—Kim, S.—Jeong, Y.—Gwon, Y.—Yoon, S.: ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14347–14356, doi: 10.1109/ICCV48922.2021.01410.

[21] Saharia, C.—Ho, J.—Chan, W.—Salimans, T.—Fleet, D. J.—Norouzi, M.: Image Super-Resolution via Iterative Refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 4, pp. 4713–4726, doi: 10.1109/TPAMI.2022.3204461.

[22] Karras, T.—Laine, S.—Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.

[23] Zhang, L.—Nie, J.—Wei, W.—Li, Y.—Zhang, Y.: Deep Blind Hyperspectral Image Super-Resolution. IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, 2021, No. 6, pp. 2388–2400, doi: 10.1109/TNNLS.2020.3005234.

[24] Zhu, F.—Zhu, J.—Chu, W.—Zhang, X.—Ji, X.—Wang, C.—Tai, Y.: Blind Face Restoration via Integrating Face Shape and Generative Priors. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7652–7661, doi: 10.1109/CVPR52688.2022.00751.

[25] Zhou, S.—Chan, K.—Li, C.—Loy, C. C.: Towards Robust Blind Face Restoration with Codebook Lookup Transformer. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.): Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Curran Associates, Inc., 2022, pp. 30599–30611, https://proceedings.neurips.cc/paper_files/paper/2022/file/c573258c38d0a3919d8c1364053c45df-Paper-Conference.pdf.

[26] Zhong, W.—Fang, C.—Cai, Y.—Wei, P.—Zhao, G.—Lin, L.—Li, G.: Identity-Preserving Talking Face Generation with Landmark and Appearance Priors.

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9729–9738, doi: 10.1109/CVPR52729.2023.00938.

[27] Wang, J.—Zhao, K.—Zhang, S.—Zhang, Y.—Shen, Y.—Zhao, D.—Zhou, J.: LipFormer: High-Fidelity and Generalizable Talking Face Generation with a Pre-Learned Facial Codebook. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13844–13853, doi: 10.1109/CVPR52729.2023.01330.

[28] Wang, X.—Li, Y.—Zhang, H.—Shan, Y.: Towards Real-World Blind Face Restoration with Generative Facial Prior. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9164–9174, doi: 10.1109/CVPR46437.2021.00905.

[29] Karras, T.—Laine, S.—Aittala, M.—Hellsten, J.—Lehtinen, J.—Aila, T.: Analyzing and Improving the Image Quality of StyleGAN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8107–8116, doi: 10.1109/CVPR42600.2020.00813.

[30] Goodfellow, I.—Pouget-Abadie, J.—Mirza, M.—Xu, B.—Warde-Farley, D.—Ozair, S.—Courville, A.—Bengio, Y.: Generative Adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. Q. (Eds.): Advances in Neural Information Processing Systems 27 (NIPS 2014). Curran Associates, Inc., 2014, pp. 2672–2680, https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[31] Zhang, R.—Isola, P.—Efros, A. A.—Shechtman, E.—Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595, doi: 10.1109/CVPR.2018.00068.

[32] Wang, Z.—Zhang, J.—Chen, R.—Wang, W.—Luo, P.: RestoreFormer: High-Quality Blind Face Restoration from Undegraded Key-Value Pairs. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17512–17521, doi: 10.1109/CVPR52688.2022.01699.

[33] Lugmayr, A.—Danelljan, M.—Romero, A.—Yu, F.—Timofte, R.—Van Gool, L.: RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11451–11461, doi: 10.1109/CVPR52688.2022.01117.

[34] Liu, H.—Xing, J.—Xie, M.—Li, C.—Wong, T. T.: Improved Diffusion-Based Image Colorization via Piggybacked Models. CoRR, 2023, doi: 10.48550/arXiv.2304.11105.

[35] Li, H.—Yang, Y.—Chang, M.—Chen, S.—Feng, H.—Xu, Z.—Li, Q.—Chen, Y.: SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models. Neurocomputing, Vol. 479, 2022, pp. 47–59, doi: 10.1016/j.neucom.2022.01.029.

[36] Wang, Y.—Yu, J.—Zhang, J.: Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. CoRR, 2022, doi: 10.48550/arXiv.2212.00490.

[37] Rombach, R.—Blattmann, A.—Lorenz, D.—Esser, P.—Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10674–10685,

doi: 10.1109/CVPR52688.2022.01042.

[38] Ho, J.—Jain, A.—Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., Lin, H. (Eds.): Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Curran Associates, Inc., 2020, pp. 6840–6851, https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[39] Ledig, C.—Theis, L.—Huszár, F.—Caballero, J.—Cunningham, A.—Acosta, A.—Aitken, A.—Tejani, A.—Totz, J.—Wang, Z.—Shi, W.: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4681–4690, doi: 10.1109/CVPR.2017.19.

[40] Dhariwal, P.—Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., Wortman Vaughan, J. (Eds.): Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Curran Associates, Inc., 2021, pp. 8780–8794, https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[41] Wang, Z.—Bovik, A. C.—Sheikh, H. R.—Simoncelli, E. P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing (TIP), Vol. 13, 2004, No. 4, pp. 600–612, doi: 10.1109/TIP.2003.819861.

[42] Chang, H.—Yeung, D. Y.—Xiong, Y.: Super-Resolution Through Neighbor Embedding. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), Vol. 1, 2004, doi: 10.1109/CVPR.2004.1315043.

[43] Xue, W.—Zhang, L.—Mou, X.—Bovik, A. C.: Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. IEEE Transactions on Image Processing, Vol. 23, 2013, No. 2, pp. 684–695, doi: 10.1109/TIP.2013.2293423.

[44] Ong, E.—Lin, W.—Lu, Z.—Yang, X.—Yao, S.—Pan, F.—Jiang, L.—Moschetti, F.: A No-Reference Quality Metric for Measuring Image Blur. Seventh International Symposium on Signal Processing and Its Applications, 2003, Proceedings, IEEE, Vol. 1, 2003, pp. 469–472, doi: 10.1109/ISSPA.2003.1224741.

[45] Heusel, M.—Ramsauer, H.—Unterthiner, T.—Nessler, B.—Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 30 (NIPS 2017. Curran Associates, Inc., 2017, pp. 6626–6637, https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[46] Simonyan, K.—Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Bengio, Y., LeCun, Y. (Eds.): Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2015, doi: 10.48550/arXiv.1409.1556.

[47] Liang, J.—Zhang, K.—Gu, S.—Van Gool, L.—Timofte, R.: Flow-Based Kernel Prior with Application to Blind Super-Resolution. 2021 IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10596–10605, doi: 10.1109/CVPR46437.2021.01046.

[48] Maas, A. L.—Hannun, A. Y.—Ng, A. Y.: Rectifier Nonlinearities Improve Neural Network Acoustic Models. Proceedings of the 30th International Conference on Machine Learning (ICML 2013), 2013.

[49] Karras, T.—Aila, T.—Laine, S.—Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. 6th International Conference on Learning Representations (ICLR 2018), 2018.

[50] Chen, Y.—Pock, T.: Trainable Nonlinear Reaction Diffusion: A Flexible Framework for Fast and Effective Image Restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, 2017, No. 6, pp. 1256–1272, doi: 10.1109/TPAMI.2016.2596743.

Wei Yu received his B.Sc. and M.Sc. degrees from the China University of Petroleum (East China), China, in 2017 and 2020, respectively, and is currently working toward his Ph.D. degree in computer science and technology at the Harbin Institute of Technology, China. His research interests include low-level vision, image super-resolution, and restoration.

Zonglin Li received his B.Sc. degree from the Harbin Institute of Technology, China, in 2017, and his M.Sc. degree from the University of Pittsburgh, USA, in 2019, and is currently working toward his Ph.D. degree with the Faculty of Computing, Harbin Institute of Technology, China. His research interests include computer vision, computer graphics, and 3D reconstruction.

Qinglin Liu received his B.Sc. degree from the Yanshan University, China, in 2014, and his M.Sc. degree from the Xidian University, China, in 2018, and is currently working toward his Ph.D. degree in computer science and technology at the Harbin Institute of Technology, China. His research interests include low-level vision, image segmentation, and image matting.

Yufan Chen received his B.Sc. degree from the Harbin Institute of Technology, China, in 2017, and his M.Sc. degree from the University of Melbourne, Australia, in 2020, and is currently Ph.D. student in computer science and technology at the Harbin Institute of Technology. He is dedicated to the research of using machine learning to solve computer vision and graphics problems, especially the reconstruction of 3D avatars from images and videos.

Shengping Zhang received his Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He is currently Professor with the School of Computer Science and Technology, Harbin Institute of Technology at Weihai. He has authored or co-authored over 50 research publications in refereed journals and conferences. His research interests include deep learning and its applications in computer vision.

Jingbo Lin graduated from the Yantai University in 2014 with Master's degree in electronics and communications engineering. He is currently working as Engineer at the Yantai Institute of Materia Medica. His main research direction is deep learning and its application in computer vision, as well as computer-assisted drug design.