Pattern Recognition 155 (2024) 110659

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

AdaptBIR: Adaptive Blind Image Restoration with latent diffusion prior for higher fidelity

Yingqi Liu [a,b], Jingwen He [c], Yihao Liu [d], Xinqi Lin [a,b], Fanghua Yu [a], Jinfan Hu [a,b], Yu Qiao [a,d], Chao Dong [a,d,*]

[a] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] The Chinese University of Hong Kong, Hong Kong, China
[d] Shanghai Artificial Intelligence Laboratory, Shanghai, China

ARTICLE INFO

Keywords:
Image restoration
Diffusion model
Adaptive adjustment

ABSTRACT

This work aims to help diffusion models get their footing in the low-level vision field, solving the pain point of insufficient fidelity. Specifically, we propose an Adaptive Blind Image Restoration framework with latent diffusion prior — AdaptBIR, which can adaptively distinguish and address various ranges of degradations. First, we quantitatively categorize images through an Image Quality Assessment (IQA) method. Then, a dual-encoder degradation removal module is employed with the guidance of IQA scores to reach better information preservation. Lastly, we utilize a two-phase controller to handle the reconstruction process in an organized manner. Extensive experiments show that applying such an adaptive framework achieves better performance on both fidelity and perceptual metrics. In this way, AdaptBIR represents more than just a novel framework, it paves the way for a broader application of the diffusion model in blind image restoration tasks.

1. Introduction

Blind image restoration (IR) aims to recover a natural image from a degraded one without explicit knowledge of the degradation process. Unlike purely generative applications in high-level vision, such as text-to-image [1,2] or image-to-image [3,4] transformations, blind IR requires a significantly higher level of fidelity in the restored images. This is particularly crucial in domains where accuracy and detail are paramount, like medical imaging, surveillance imaging, satellite remote sensing, and traffic hub monitoring. As an inherently ill-posed problem, blind IR presents unique challenges, particularly in the careful and precise handling of input low-quality (LQ) images to ensure accurate restoration.

The range of existing methods for blind IR spans from traditional Convolutional Neural Networks (CNNs) to advanced diffusion-based models. Each of these approaches has its limitations in effectively tackling the complexities of blind IR, underscoring the need for continued research and development to enhance accuracy and reliability. Traditional CNN-based methods [5,6], while effective in sticking to fidelity metrics like PSNR, often produce overly smooth outputs that lack details. Likelihood-based models such as Normalizing Flows (NFs) [7,8] and Variational Autoencoders (VAEs) [9,10] show improved visual results but still do not fully meet practical demands in terms of realism and diversity. Recently, Generative Adversarial Networks (GANs) [11,12] and Diffusion Models (DMs) [1,13] have emerged as leading approaches in this field. GANs have demonstrated promising performance in generating high-quality images. However, they often suffer from difficult optimization issues due to their adversarial training strategy. This drawback limits their practical utility and effectiveness in leveraging extensive prior knowledge from large datasets.

In contrast, diffusion models, known for their stable optimization and strong generalization capabilities, excel in producing visually appealing results. Besides benefiting from the iterative denoising process, they also establish powerful priors through large-scale training datasets. However, the stochastic nature of DMs can sometimes lead to outputs that fail to achieve the desired level of fidelity, especially in scenarios where the input images only contain mild degradation. The diffusion model behaves just like a brilliant but arrogant student. Despite being full of creativity, it may disregard important guidance and produce something deviating from reality. This tendency towards over-creativity can be a shortcoming in practical applications. Therefore, there is a need for models that are not only capable of handling a wide range

* Corresponding author at: Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
E-mail addresses: yq.liu3@siat.ac.cn (Y. Liu), hejingwen@pjlab.org.cn (J. He), liuyihao14@mails.ucas.ac.cn (Y. Liu), linxinqi23@mails.ucas.ac.cn (X. Lin), fanghuayu96@gmail.com (F. Yu), jf.hu1@siat.ac.cn (J. Hu), yu.qiao@siat.ac.cn (Y. Qiao), chao.dong@siat.ac.cn (C. Dong).

https://doi.org/10.1016/j.patcog.2024.110659
Received 7 March 2024; Received in revised form 15 May 2024; Accepted 3 June 2024
Available online 7 June 2024

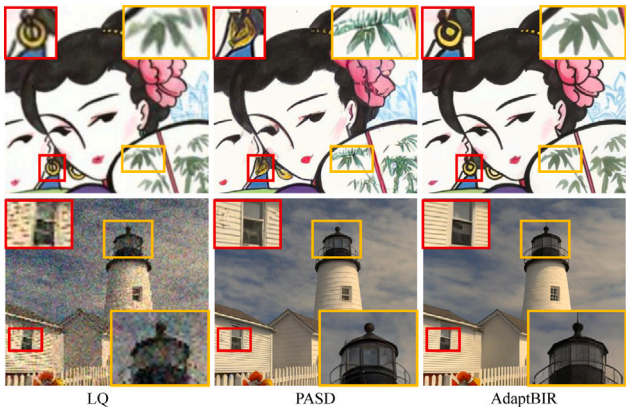

Fig. 1. Visual comparison. In mild degradation, AdaptBIR offers higher fidelity; in severe cases, it produces more realistic details compared to PASD.

of degraded inputs but also maintain a balance between creativity and realism.

Previous works applying the diffusion model to blind IR, have attempted to address the vexing inherent randomness but yielded limited success. StableSR [14] stores all intermediate outputs from the diffusion process and utilizes them to generate refined results through cumbersome post-processing. DiffBIR [15] employs latent image guidance [16] to constrain the denoising process, but it inadvertently limits the model's generative capabilities. FreeU [17] focuses on controlling the inference process, for instance, re-weighting features in the UNet's skip connections and backbone. Although it offers an innovative way to influence the output, a significant amount of trial and error is required to achieve optimal settings. The cost of trying poses challenges for practical applications.

To address this issue, this paper introduces AdaptBIR, an adaptive blind image restoration framework leveraging powerful diffusion priors. AdaptBIR stands out with its ability to assess the quality of input images and perform adaptive processing accordingly. Adopting a dual-encoder architecture as a pre-processing module, AdaptBIR can handle diverse degrees of degradation commonly encountered in real-world images. Referring to an image's IQA score, AdaptBIR employs adaptive linear interpolation to ensure that crucial information in the image is preserved effectively. Furthermore, AdaptBIR incorporates a novel two-phase control mechanism to provide in-depth denoising guidance. The first phrase is borrowed from DiffBIR, which can achieve a coarse control based on the input image and guarantee a visually appealing outcome. The second phase introduces Controllable Spatial Feature Transformation (CSFT) layers, enhancing the fidelity of the restored images. Notably, this two-phase control operates seamlessly in a single inference step, with the CSFT layers' influence being dynamically adjusted based on the IQA scores. This integrated framework ensures that AdaptBIR not only excels in restoring images but also maintains a superior balance between generative quality and fidelity.

We test our approach on a diverse set of synthetic and real-world datasets. The results show that AdaptBIR not only retains information in high-quality images but also effectively manages severely degraded inputs (see Fig. 1). This balance between fidelity and realism is a standout feature of our model. To further assess AdaptBIR's generalization capabilities, we establish three distinct degradation scenarios, ranging from mild to severe. These tests demonstrate the model's adaptability across varying levels of image quality. Additionally, our ablation studies confirm the efficacy of each component in AdaptBIR, while comparative analysis with other control mechanisms further validates our model's superiority.

The main contributions of our work can be summarized as follows:

- We propose AdaptBIR, an innovative DM-based framework utilizing IQA scores to guide adaptive image restoration, showcasing robust performance in real-world scenarios.
- We design a dual-encoder structure that achieves excellent degradation removal while maintaining self-information well. We also introduce CSFT, a fine-grained control layer, to significantly enhance fidelity.
- Extensive experimental results confirm our AdaptBIR's exceptional ability to generalize across multiple scenarios, showcasing its practicality and effectiveness in a wide range of applications.

2. Related work

2.1. CNN-based image restoration

Image restoration is an active field in computer vision that aims to recover high-quality (HQ) images from degraded low-quality (LQ) observations. The early pioneer works typically employ pixel-wise loss functions, which tend to produce overly smooth results. Another noteworthy issue with these methods is the lack of sufficient generalizability. Although they can demonstrate impressive performance when applied to synthetic data, their effectiveness diminishes when confronted with real-world scenarios. To address the challenges, blind IR has been proposed.

While numerous methodologies for blind IR have currently been explored, the key to overcoming this ill-posed problem lies in comprehensively modeling the complex distributions of the real world. Among traditional CNN-based approaches, some early methods employ explicit priors for modeling distributions, such as estimating the pre-parameterized degradation kernels with given examples [18,19]. However, they heavily rely on predefined assumptions, which are often insufficient for the complexity of real-world cases. Shifting away from reliance on explicit priors, more recent works turn their attention to exploring implicit priors, such as semantic segmentation probability maps [20] and pre-trained GAN priors [21,22]. Although the implicit priors are more proficient in capturing the inherent structures of the image, most of these methods are confined to specific scenarios, and therefore lack generalizability for real-world BIR tasks.

Beyond these CNN-based methods, other generative models can also serve as priors. Our work primarily focuses on the robust and extensive generative priors from pre-trained diffusion models, which are more powerful in handling diverse degradations and enable controllable generation.

2.2. DM-based image restoration

Denoising Diffusion probabilistic models (DDPM) [23] are a kind of generative models using a *Markov* chain to transform latent variables in simple distributions (e.g., Gaussian) to data in complex distributions. Since LDM [1] extends DDPM [23] to latent space, it substantially achieves remarkable results with less computational resources. This not only sparks a series of developments in diffusion models for image synthesis tasks, such as SD and Imagen [24], but also triggers a significant surge of interest in the low-level vision field. SRDiff [25] pioneers the utilization of diffusion-based models for single image super-resolution; Dif-Fusion [26] capitalizes on diffusion models to effectively fuse infrared and visible images; DiffIR [27] demonstrates the potential of diffusion models in image completion; Ren [28] introduced a multiscale structure guidance mechanism to assist image-conditioned diffusion models in the task of image deblurring. Although these methods showcase the capabilities of diffusion models in real-world applications, they have traditionally been trained from scratch for specific tasks. However, with the advancement of large model technology, there is an increasingly prominent trend towards fine-tuning based on subtasks.

ControlNet [29] constructs a parallel structure of UNet for conditional inputs. As the sole module updated during the fine-tuning

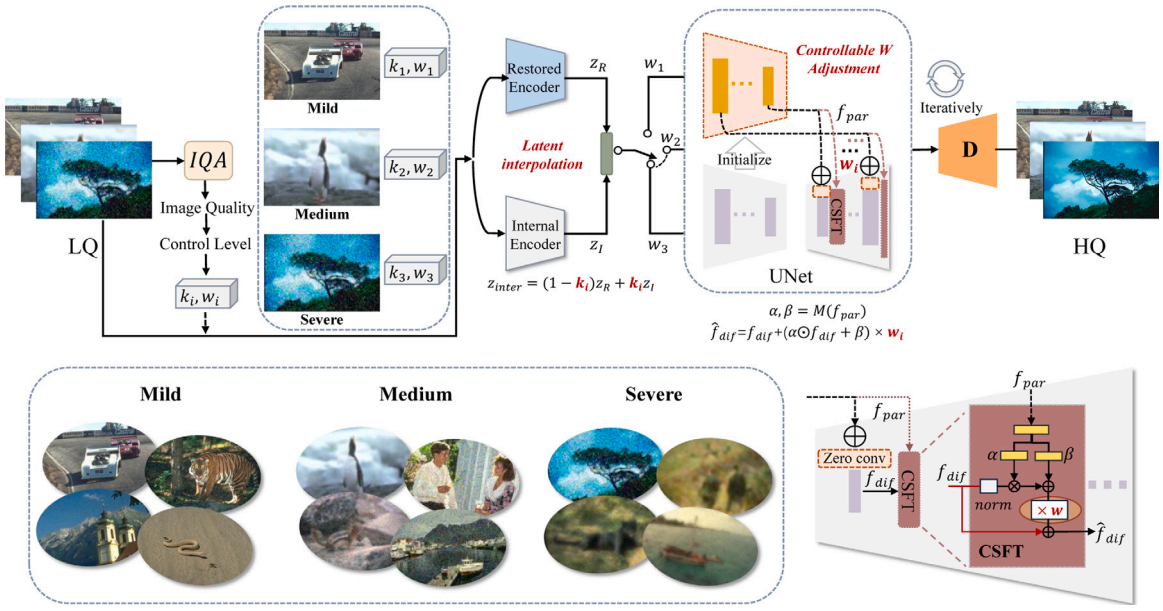

Fig. 2. Framework of AdaptBIR. Firstly, we categorize the input images into three classes based on their image quality. Each class corresponds to a specific control level along with associated control parameters k and w. We introduce a dual-encoder architecture to selectively extract useful information z_{inter} from the input images, with the latent interpolation ratio determined by the coefficient k. Subsequently, we perform a two-phase control on the images using z_{inter}, leveraging the coefficient w. AdaptBIR achieves a win–win situation by enhancing both the quality and fidelity of the reconstruction results.

process, the parallel module enables further guidance to the generation process of large pre-trained diffusion models. Building upon ControlNet, some recent works attempt to use LQ images as specific conditions to produce HQ outputs. DiffBIR [15] proposes a two-stage pipeline to handle complex degraded input images. It adopts SwinIR [30] as a pre-restoration module. StableSR [14] injects degradation information directly into the UNet with inserted Spatial Feature Transformations (SFT) layers. PASD [17] introduces a pixel-aware cross-attention module into the UNet, enabling the network to perceive image local structures.

Although these methods have notable advancements in visual quality, there are still some deficiencies which can be summarized into three aspects: 1. The restoration module easily erases useful details when encountering images with mild degradation. 2. Over-injecting control information into the UNet may disrupt the internal denoising process. 3. With heavy attention modules, computational costs significantly increase. In contrast, AdaptBIR constructs a dual-encoder restoration module and introduces a lightweight two-phase controller. This approach can effectively handle degradations with varying degrees, while imposing minimal burden on the training process.

2.3. Controllable image restoration

Controllable image restoration allows users to tailor the restoration process according to specific user preferences. Codeformer [18] introduces a Controllable Feature Transformation (CFT) module to control information flow from the LQ encoder to the HQ decoder. StableSR [14] constructs a similar Controllable Feature Wrapping (CFW) module to refine the results of the diffusion model in an extra training stage. DiffBIR [15] proposes latent image guidance to force spatial alignment and color consistency between reconstructed images and LQ images. However, these methods still exhibit limitations. StableSR requires retaining all intermediate products during the denoising period, such as image latent and sample results, leading to massive computational costs. The modulation design of DiffBIR would suppress the model's generative capacity when enhancing fidelity, resulting in a significant fidelity-realism trade-off. Unlike the above methods, we merge modulation into the denoising process, so there is no need for additional data processing. With a better controlling mechanism, AdaptBIR can achieve high realism while maintaining fidelity.

3. Methodology

In this section, we introduce AdaptBIR. It is a unified pipeline that first recognizes the image quality of the LQ images and then adaptively decides the controlling mechanism for handling the diverse and unknown degradations. The overall framework is illustrated in Fig. 2. AdaptBIR mainly consists of two key components:

(1) A dual-encoder for degradation removal with VAE Prior. This step utilizes the IQA method to guide an adaptive linear latent interpolation.

(2) A two-phase controller for image reconstruction with latent diffusion prior. The interpolated latent is applied to the feature maps of UNet's skip connections and backbone, with an IQA-related intensity.

3.1. Dual-encoder for degradation removal with VAE prior

Previous DM-based blind IR methods often rely on a pre-restoration model as a preprocessing step. For example, DiffBIR [15] utilizes a pre-trained SwinIR [30] to remove the input LQ images' degradations. However, since the pre-restoration model is trained from scratch, these methods face challenges when the actual degradation and the training degradation are different. Additionally, those SR models tend to remove excessive details, resulting in the loss of valuable information.

To address these issues, we propose a dual-encoder structure with the VAE prior. Instead of using an external pre-restoration model, we directly train a parallel restoration VAE ($\mathcal{V}_R$), which naturally aligns better with the internal VAE ($\mathcal{V}_I$) in Stable Diffusion. The restoration process is optimized by L_2 pixel loss, which can be represented as follows:

$$I_{IR} = \mathcal{V}_R(I_{LQ}), L_{IR} = \|I_{IR} - I_{HQ}\|_2^2. \quad (1)$$

where I_{LQ}, and I_{IR} represent the LQ input and restored image, respectively. $\mathcal{V}_R$ learns to remove the degradations and generate high-quality outputs with realistic textures. When encountering severely distorted images, it can execute restoration and produce visually pleasing results. As a companion, $\mathcal{V}_I$ is not focused on the restoration task. It captures the underlying distribution of input images without considering the degradation. It serves as a reference for the restoration process,

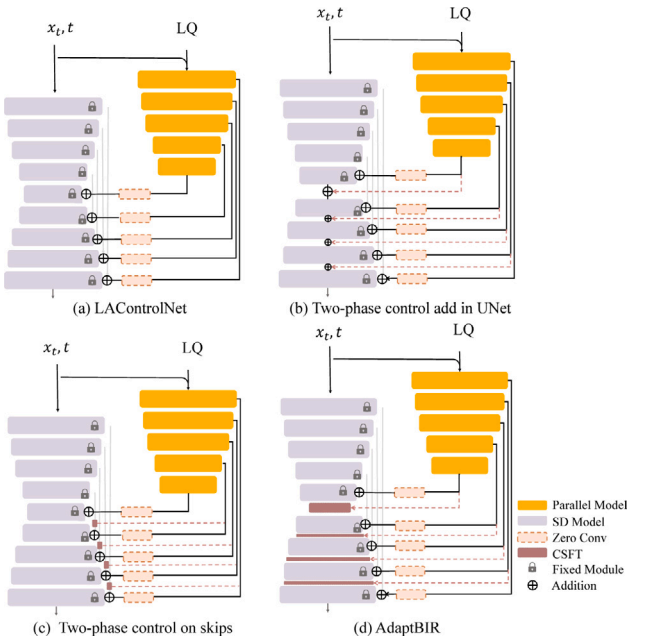

Fig. 3. Comparison among four different control modes. (a) LAcontrolNet [15] adds conditions to the skip connections of the UNet, which can be regarded as a one-phase control. (b) Imitating the addition operation of LAControlNet, one can introduce a second phase control into the UNet backbone. (c) Applying CSFT layers for the second-phase control, one can position them on the skip connections of the UNet. (d) Based on LAControlNet, the proposed AdaptBIR conducts a second-phase control into the UNet backbone through CSFT layers.

providing prior knowledge about the true underlying structure of the image.

During the inference stage, we leverage the dual-encoder design to utilize the advantages of both $\mathcal{V}_R$ and $\mathcal{V}_I$. Firstly, an IQA method is used to evaluate the quality of the test image and obtain the corresponding IQA score. Based on this score, we adaptively produce the diffusion latent using a linear interpolation:

$$z_{inter} = (1-k)\cdot z_R + k\cdot z_I, z_R = \varepsilon_R(x), z_I = \varepsilon_I(x). \tag{2}$$

Given an input x, z_R and z_I denote the latent features produced by encoders ε_R and ε_I of two VAEs, respectively. z_{inter} is the interpolated latent using the adaptive coefficient k, which is derived from the IQA score.

By combining the outputs of two encoders, the restoration process can benefit from the complementary information provided by each encoder. This dual-encoder design is particularly beneficial for cases with mild distortion, where the available information can be effectively utilized. A more detailed explanation of this adaptive interpolation process is provided in Section 3.3.

3.2. Two-phase controller for image reconstruction with diffusion prior

Preliminary: LAControlNet. Leveraging a pre-trained T2I diffusion prior to the image restoration task is challenging. Because the conditional image generation is under strict constraints, such as semantic content and texture details. To address this problem, DiffBIR [15] proposes LAControlNet (Fig. 3(a)) that projects the condition image to the latent space using a pre-trained VAE encoder, thus achieving more effective image-level control compared with the original ControlNet. Specifically, given a skipped feature map $\mathbf{x}_{skip}$ from the UNet denoiser and the control signals $\mathbf{c}$ with the same spatial size from LAControlNet, the control mechanism is illustrated as follows:

$$\hat{\mathbf{x}}_{skip} = \mathbf{x}_{skip} + zero\_conv(\mathbf{c}), \tag{3}$$

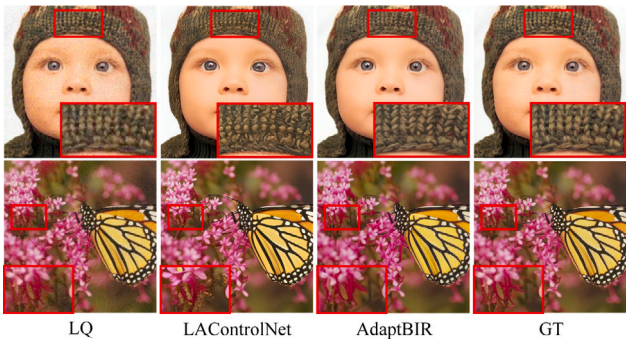

Fig. 4. Visual comparison of AdaptBIR and LAControlNet. LAControlNet produces textures that do not match the original image, while AdaptBIR achieves higher fidelity during restoration. See the Appendix for more results.

where $zero\_conv$ is a convolution layer that initialized with zero, and $\hat{\mathbf{x}}_{skip}$ is the obtained output skipped feature map. After this addition operation, LAControlNet concatenates the modified skipped feature map $\hat{\mathbf{x}}_{skip}$ with the features from the UNet backbone ($\mathbf{x}_{fea}$):

$$\hat{\mathbf{x}}_{fea} = concat(\hat{\mathbf{x}}_{skip}, \mathbf{x}_{fea}), \tag{4}$$

where $\hat{\mathbf{x}}_{fea}$ is the output feature map.

Is there a better controller? LAControlNet has been demonstrated to be very effective in several conditional image generation tasks. But for image restoration, we observe that this simple addition operation may bring some inconsistent texture details with the LQ input, as shown in Fig. 4. This indicates that this simple addition control mechanism might be too weak for the image restoration task, as image restoration requires the generated results to be both realistic and faithful.

Inspired by FreeU [31], we assume that the UNet backbone plays the main role in image generation, where the skip connections are more related to high-frequency information. Thus, DiffBIR maintains the generation ability well by not touching the features of the UNet backbone, but producing stochastic, or even wrong texture details occasionally. To remedy this defect, we explore a more potent control mechanism.

Two-phase Controller. Building upon existing research foundations, we propose a two-phase controller to regulate the generation process. Specifically, the first phase inherits LAControlNet and is responsible for generating realistic results (see Fig. 2). While the second phase imposes control on the features of the UNet backbone. Spatial Feature Transform (SFT) is a fine-grained control method, proposed by Wang [20]. The transformation is realized by scaling and shifting feature maps in spatial dimensions, using specific parameters related to conditions. Leveraging its effective and lightweight characteristics, we propose a Controllable Spatial Feature Transform (CSFT) layer. Based on the primary SFT layer, an adaptive coefficient w is introduced to control the strength of the second phase's impact, which depends on the quality of the actual input image. Let F_{par} and F_{dif} be the parallel module and UNet backbone features, respectively:

$$\hat{F}_{dif} = F_{dif} + (\alpha \odot F_{dif} + \beta)\times w; \alpha, \beta = M_\theta(F_{par}). \tag{5}$$

where α, β denote the affine parameters in CSFT and M_θ denotes a small network consisting of several convolution and normalization layers. Note the second phase control is nullified when w is set to zero.

During finetuning, we freeze the weights of the stable diffusion model and its parallel module, and only train the CSFT layers. This strategy enables us to make a slight modulation on preliminary results, simultaneously retaining the generative prior of LAControlNet. Extensive experiments indicate that this two-phase control can significantly enhance fidelity. To provide more meticulous confirmation of our hypothesis, we apply CSFT to the UNet skip connections' features (Fig. 3(c)). Not surprisingly, it yields limited improvements compared

to its application within the UNet backbone, since its control also acts on the surface features of the network. More specific comparisons can be observed in the ablation study.

3.3. Adaptive image restoration framework

In this section, we introduce our adaptive pipeline, which consists of three periods — IQA Classification, Degradation Removal, and Image Reconstruction.

The goal of the IQA Classification is to tell "What is the extent of degradation on the input image?" Some traditional IQA metrics, including NIQE [32], have been proven to be inaccurate and have consequently been discarded in recent research [33]. Other approaches, like MANIQA [34], are unsuitable for incorporation into the training process due to their slow computational speed. Therefore, we choose MUSIQ [35] to evaluate the quality of input images, which boasts both prevalence and remarkable computational speed. As more advanced IQA approaches emerge, such as CICI [36] and DGQA [37], they are expected to bring further enhancements to our methodology.

With the help of IQA, We classify the input images into three categories based on their degradation levels: mild, medium, and severe. Each class is associated with different modulation parameters k and w, designed to address the unique characteristics within each category. Specifically, a mild degree (MUSIQ $\in [70, \infty]$) corresponds to a more dominant internal encoder ε_I and fully effective CSFT layers ($k = 0.7, w = 1$). While the severe degree (MUSIQ $\in [0, 40]$) corresponds to an original LAControlNet setting ($k = 0, w = 0$). In medium degree (MUSIQ $\in [40, 70]$), the network adaptively chooses k and w within a specified range ($k \in [0, 0.7], w \in [0, 1]$), thereby the whole restoration process is tailored to individual needs.

During the Degradation Removal process, the input image will first go through a dual-encoder structure and be transformed into an interpolated restored latent z_{inter}. The interpolation ratio is determined by the coefficient k. In the third Image Reconstruction period, the interpolated latent will be divided into two branches. One branch emulates LAControlNet to influence the features of UNet skip connections, while the other adaptively controls the features in the UNet backbone, guided by the coefficient w.

If an image is assigned to the mild partition, it means that quite a lot of useful information is still contained within the image. So the model will adopt a high ratio towards the internal encoder and maximize the utilization of CSFT layers to supplement information from the inputs. On the contrary, the severe degree corresponds to poor quality with heavy degradation. At this point, the baseline model is already sufficient, so we directly turn off the modulation module to avoid unpleasant disturbance to the internal features of UNet from input LQ images. When the evaluation metric falls in the medium partition, the model automatically chooses the appropriate parameters to balance the quality and fidelity.

4. Experiments

4.1. Datasets, settings, metrics

Datasets. We train the proposed AdaptBIR on the ImageNet1k dataset, adopting two distinct degradation pipelines during training. This allows the Network to recognize various degradation types.

We evaluate our method on both synthetic and real-world datasets. For synthetic data, to assess our method thoroughly, we merge six commonly used image restoration datasets into an extensive collection, including Set5, Set14, Live1, CBSD68, Urban100 and Manga109. We name this collection CommonIR dataset, comprising 323 images. We also conduct testing on DIV2K validation dataset following the previous work. Both of the LQ images in these two datasets are synthesized under the degradation pipeline of Real-ESRGAN [38]. Further, to affirm the effectiveness of our approach, we conduct scenario-specific tests with varying degradation levels on CommonIR dataset. For real-world datasets, we follow previous works to conduct comparisons on Real47 and RealSRSet.

Settings. AdaptBIR is built based on Stable Diffusion 2.1-base.[1] We train the model with 512 × 512 resolution with 8 NVIDIA 48G-A6000 GPUs. In stage I, we finetune a pre-trained VAE with a batch size of 48 for 100k iterations, following the same Real-ESRGAN pipeline. After we obtain a reliable restoration encoder from the fine-tuned VAE, we combine it with the inherent encoder inside Stable Diffusion. The combined dual-encoder is positioned at the beginning of the parallel module, which is employed to remove degradations. We use the AdamW optimizer during this stage, with the initial learning rate setting to 4.5×10^{-6}.

In stage II, we finetune the LAControlNet for 60k iterations with a batch size of 128 and adopt it as our baseline model. The CSFT layers are introduced in stage III, which are inserted after each residual block of Stable Diffusion. At this stage, we adopt a milder version of Codeformer pipeline [18] to train the CSFT layers. Equipped with the IQA Classification and dual-encoder, the network learns how to choose appropriate values for k and w at this stage. An additional 20k finetuning iterations should be more than ample since we only need the CSFT layers to grasp an effective utilization of higher-quality latent. In these two stages, we follow Stable Diffusion to use Adam optimizer and the learning rate is set to 10^{-4}.

For inference, we adopt spaced DDPM [39] sampling with 50 timesteps. Our model is able to handle images with arbitrary sizes beyond 512 × 512. For images under 512 × 512, we first upsample them with the short side enlarged to 512, and then resize them back.

Metrics. Regarding the evaluation with ground truth, we adopt the traditional metrics: PSNR, SSIM, and LPIPS. Since ground-truth images are unavailable in the real scenes, we employ the widely used non-reference metrics: CLIP-IQA[2] and MUSIQ for perceptual quality evaluation.

4.2. Comparison with existing methods

We compare our AdaptBIR with state-of-the-art methods, including BSRGAN [40], Real-ESRGAN+ [38], DiffBIR [15], StableSR [14], PASD [17].

Evaluation on the general datasets. We first show the quantitative comparison on the two synthetic datasets and two real-world benchmarks in Table 1. For the synthetic CommonIR dataset and DIV2K validation dataset, on both the image fidelity metrics PSNR and SSIM, our AdaptBIR achieves the best scores among the comparative DM-based methods. Simultaneously, its performance on image perceptual metrics also ranks among the top. This demonstrates that AdaptBIR has stronger generalization capability across diverse datasets. Although the GAN-based methods can obtain higher numeric results on the two fidelity metrics, i.e., PSNR, and SSIM, they fail to restore the detailed textures and tend to generate blurry results, which can be reflected in their inferior performance on the other perceptual metrics. Additionally, we present the qualitative visual comparison in Fig. 5. From the comparison, it can be observed that our AdaptBIR is capable of generating vivid images with faithful details.

For real-world benchmarks, our AdaptBIR also outperforms state-of-the-art methods in terms of both perceptual metrics. The specific visual results can be found in Fig. 6. It can be seen that our AdaptBIR can generate more realistic details with better visual quality (see the restored textures in clothes, skins, jewelry, etc.). More visual comparison results can be found in the Appendix.

Evaluation on different degrees of degradation. To comprehensively assess our model's performance across diverse scenarios, we

[1] https://github.com/Stability-AI/stablediffusion
[2] https://github.com/IceClear/CLIP-IQA

Table 1
Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. Red and blue colors represent the best and second best performance. The third-place achievements are marked as Underline.

Datasets	Metrics	BSRGAN	Real-ESRGAN+	DiffBIR	StableSR	PASD	**AdaptBIR**
CommonIR	PSNR(dB)↑	21.70	**22.05**	21.20	20.71	21.10	21.71
	SSIM↑	0.7371	**0.7762**	0.7298	0.7143	0.7340	0.7447
	LPIPS↓	0.3497	0.3149	0.2998	0.3146	0.2960	**0.2763**
	CLIP-IQA↑	0.7416	0.6894	0.8604	0.7227	0.8137	**0.8853**
	MUSIQ↑	61.9082	59.3755	67.9199	61.4855	**71.3656**	69.9369
DIV2K Valid	PSNR(dB)↑	**22.55**	22.00	20.80	18.73	20.87	21.00
	SSIM↑	**0.7309**	0.7180	0.6572	0.6242	0.6578	0.6707
	LPIPS↓	0.3583	0.3591	0.3381	0.3551	0.3175	**0.3167**
	CLIP-IQA↑	0.6308	0.6131	0.8884	0.8253	0.8951	**0.8985**
	MUSIQ↑	58.1763	58.4103	71.1641	69.2836	**72.0061**	71.1978
Real47	CLIP-IQA↑	0.7968	0.7844	0.9188	0.8283	0.8868	**0.9234**
	MUSIQ↑	69.4703	68.2936	70.3881	68.3422	70.5479	**70.9915**
RealSRSet	CLIP-IQA↑	0.7914	0.6829	0.8241	0.7444	0.7661	**0.9085**
	MUSIQ↑	67.6637	63.2737	62.6910	64.8372	63.0822	**68.4932**

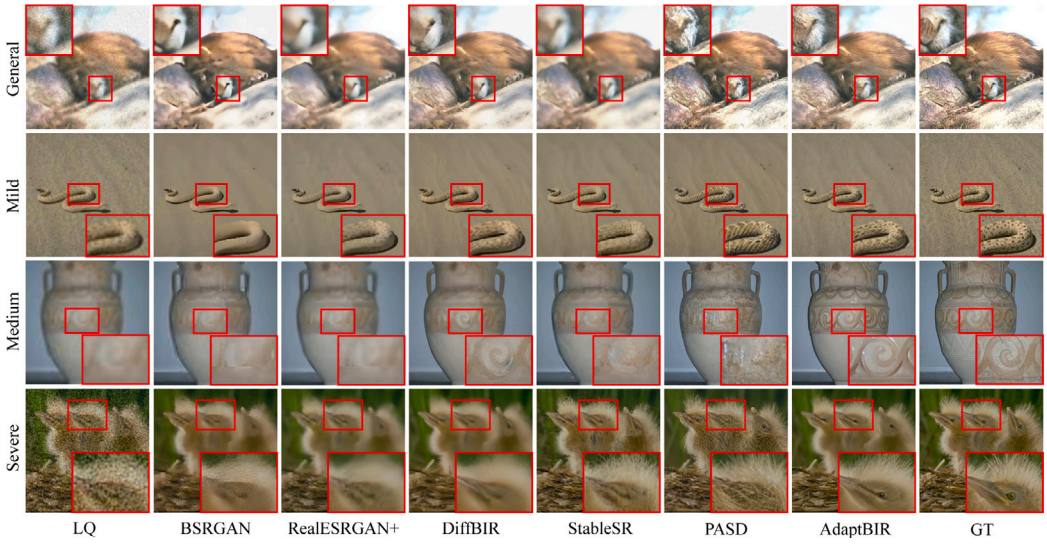

Fig. 5. Visual comparison of AdaptBIR with the state-of-the-art methods on synthetic datasets with four degradation degrees. The first column, the second column, the third column, and the fourth column represent the degradation levels of General (RealESRGAN), Mild, Medium, and Severe, respectively. Our AdaptBIR significantly improves fidelity while also achieving enhanced visual effects. See the Appendix for more results.

design three test sets with varying degrees of degradation on the CommonIR dataset. To avoid the unfair comparison caused by training priors, we use a different degradation pipeline [18] to synthesize the LQ images. They are created in three versions, each corresponding to a different degree of degradation, i.e., mild, medium, and severe.

In addition to the visual comparison results consolidated in Fig. 5, we particularly showcase the remarkable fidelity capability of AdaptBIR under mild degradation in Fig. 7. The results indicate that AdaptBIR excels in retaining high-quality texture details, supported by the evidence that only it successfully restores the texts and architectural contours. We further conduct a more detailed quantitative analysis to compare AdaptBIR's performance with the other DM-based models. As results shown in Table 2, AdaptBIR consistently demonstrates superior performance across three degradation levels, achieving a win–win situation of fidelity and realism. Notably, its PSNR scores surpass the second-best performance by 1.82 dB, 0.76 dB, and 0.27 dB in mild, medium, and severe scenarios, respectively. While StableSR gains a slight advantage in CLIP-IQA and MUSIQ under mild scenarios, it introduces numerous unrealistic artifacts, as reflected in the lowest fidelity metric. Incidentally, GAN-based methods may also achieve satisfactory fidelity in mild scenarios, but they always fail to handle medium and severe degradation. Therefore, a diffusion model with high fidelity is the superior choice in practical applications.

4.3. Ablation studies

Reliability of IQA Classification. To evaluate the rationality of IQA metric classification, we analyze the statistical characteristics of CommonIR dataset under medium degradation and general degradation (Real-ESRGAN) scenarios. Using the mentioned classification method, we separate them into three classes and compute quality metrics for each subclass. As depicted in Fig. 8, we present three kinds of representative images with corresponding MUSIQ and PSNR ranges. It is observed that high IQA values are related to mildly degraded images, while images with low IQA scores contain complex degradations. Their respective PSNR values support the assessment of image quality from an alternative perspective. Detailed values are shown in Table 3, where Scenario $\mathcal{A}$ denotes medium degradation and Scenario $\mathcal{B}$ represents general degradation. Since the second-order degradation of

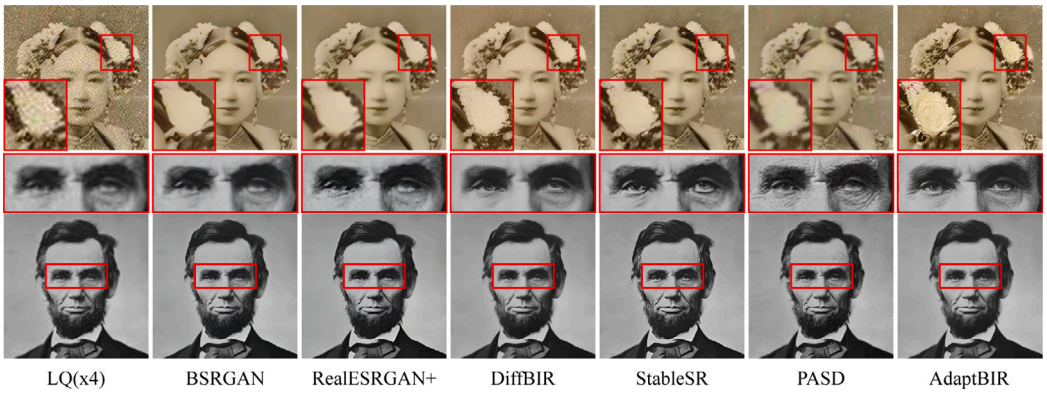

Fig. 6. Visual comparison of AdaptBIR with the state-of-the-art methods on several representative Real47 and RealSRSet samples (×4 SR). Our AdaptBIR can generate more realistic details. See the Appendix for more results.

Fig. 7. Visual comparison of AdaptBIR with the state-of-the-art methods on mild degradation. Our AdaptBIR can significantly enhance the fidelity of the DM-based methods.

Table 2
Quantitative comparison on CommonIR dataset with different degrees of degradation. Red and blue colors represent the best and second best performance, respectively.

Degradation	Metrics	DiffBIR	StableSR	PASD	**AdaptBIR**
Mild	PSNR(dB)↑	22.56	21.59	22.82	**24.64**
	SSIM↑	0.8008	0.8126	0.8223	**0.8512**
	LPIPS↓	0.2417	0.2225	0.2177	**0.1770**
	CLIP-IQA↑	0.8960	**0.9111**	0.8728	0.9069
	MUSIQ↑	71.3228	**71.8352**	71.4488	71.4505
Medium	PSNR(dB)↑	21.49	21.17	21.44	**22.25**
	SSIM↑	0.7504	0.7524	0.7477	**0.7679**
	LPIPS↓	0.2815	0.2704	0.2793	**0.2556**
	CLIP-IQA↑	0.8778	0.7980	0.8491	**0.8780**
	MUSIQ↑	69.7173	67.1958	69.5581	**70.0037**
Severe	PSNR(dB)↑	20.09	20.25	20.10	**20.52**
	SSIM↑	0.6479	0.6562	0.6607	**0.6697**
	LPIPS↓	0.3572	0.3609	0.3506	**0.3315**
	CLIP-IQA↑	0.8108	0.5602	0.7521	**0.8342**
	MUSIQ↑	62.9190	52.2118	62.8792	**66.9381**

Fig. 8. Visualization of three classes and their corresponding MUSIQ and PSNR ranges.

Table 3
Quality metrics of three divided subclasses under two scenarios. Scenario $\mathcal{A}$ represents the medium degradation, Scenario $\mathcal{B}$ represents the general degree degradation.

Subclass	Scenario $\mathcal{A}/\mathcal{B}$		
	PSNR(dB)↑	SSIM↑	LPIPS↓
Mild	25.73/-	0.9474/-	0.1675/-
Medium	25.20/21.18	0.8875/0.8386	0.2409/0.3124
Severe	21.86/20.88	0.6920/0.6370	0.3931/0.4323

Table 4
Comparison of SwinIR and Dual-encoder on CommonIR dataset with mild degradation.

Models	PSNR(dB)↑	LPIPS↓	CLIP-IQA↑	MUSIQ↑
SwinIR	24.14	0.2622	0.7419	56.6101
Dual Encoder	**25.05**	**0.1921**	**0.7568**	**61.6535**

Table 5
Ablation studies of different control positions on CommonIR dataset and DIV2K validation dataset.

Control Position	CommonIR/DIV2K Valid		
	PSNR(dB)↑	SSIM↑	LPIPS↓
One-phase	21.10/20.62	0.7336/0.6560	0.2809/0.3273
Skip Connections	21.69/20.83	0.7455/0.6650	0.2769/0.3233
UNet Backbone	**21.71/21.00**	**0.7457/0.6707**	**0.2763/0.3167**

Real-ESRGAN is relatively heavy, no image in Scenario $\mathcal{B}$ is categorized to a mild degree.

Superiority of Dual-Encoder. To verify that our dual-encoder can retain more useful information compared to the traditional CNN network SwinIR, we compare the performance of the two modules in mild degradation. As shown in Table 4, the experimental results indicate that the dual-encoder outperforms SwinIR in both quality and perceptual metrics. This implies the dual-encoder can make better preservation of original image information. Therefore, more valuable details are left for subsequent reconstruction. The visual comparisons in Fig. 9 also confirm this conclusion.

Importance of Controlling Position. In order to assess the impact of controlling position on the results, we compare the differences between applying the CSFT layers to the features of the UNet backbone and UNet skip connections (See Fig. 3). We find that even though applying CSFT to the features of UNet skip connections can gain some benefits, they are negligible compared to the performance improvement by directly controlling the features of the UNet backbone. A more concrete comparison result is displayed in Table 5.

Suitability of Controlling mode. In addition to using scale and bias to affect the features of the UNet backbone, we further compare it with ordinary addition operations inherited from LAControlNet. Note

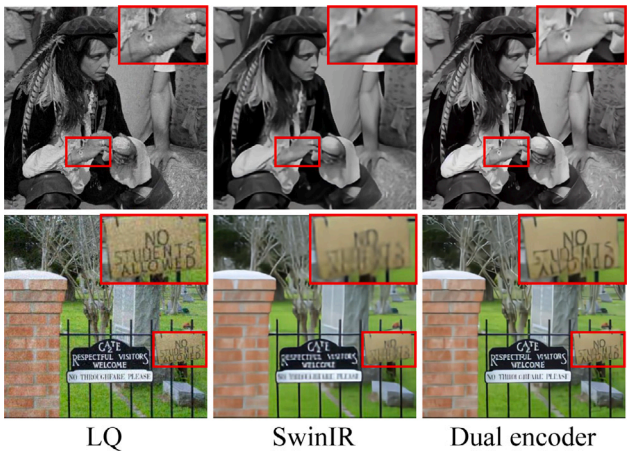

Fig. 9. Visual comparisons of SwinIR and Dual-encoder on CommonIR dataset with mild degradation. SwinIR erases valuable information which crucial for authentic image reconstruction, such as the bracelet on the man's hand and the text on the wooden board. In contrast, VAE preserves these details. See the Appendix for more results.

Table 6
Ablation studies of different control modes on CommonIR dataset and DIV2K validation dataset.

Control Mode	CommonIR/DIV2K Valid		
	PSNR(dB)↑	SSIM↑	LPIPS↓
One-phase	21.10/20.62	0.7336/0.6560	0.2809/0.3273
CAdd	21.41/20.77	0.7396/0.6620	0.2806/0.3251
CSFT	**21.71/21.00**	**0.7457/0.6707**	**0.2763/0.3167**

we also use an adaptive coefficient w to control the addition injection guided by IQA scores. As shown in Table 6, the addition operation shows little performance improvement. We speculate that this brute-force second injection does not genuinely benefit the network, as it disregards the distribution discrepancy between conditional information and features of the UNet backbone. In contrast, CSFT is based on a condition-mapped data distribution. Thus, its employment appears to have a more natural and gentle effect, achieving the best performance as expected.

5. Conclusion

In this study, we introduce an adaptive image restoration framework with latent diffusion prior — AdaptBIR. It can adaptively handle inputs from various scenarios without the need for any manual regulation. During the implementation process of AdaptBIR, we design a dual-encoder based on pre-trained VAE. It achieves excellent degradation removal while maintaining self-information well. Additionally, we propose CSFT to adaptively scale and shift the features of the UNet backbone. Such fine-grained control enhances fidelity significantly. Moreover, we utilize IQA to assess LQ images and classify them into three intervals. By associating each interval with distinct settings, AdaptBIR achieves robust performance in real-world scenarios. Extensive experiments show that the proposed AdaptBIR can achieve higher perceptual metrics in blind IR tasks, while its fidelity is far superior to other DM-based methods.

Even though these innovative components help the model outperform existing methods, there are some limitations. Firstly, while the whole pipeline is robust to image quality classification results, its performance may be affected when the IQA module misidentifies some special cases. With better IQA methods, our approach can be further enhanced. Secondly, the proposed method requires a staged training process. Despite its advantages in terms of flexibility and generalizability, it remains more complex than typical end-to-end approaches. Lastly, for more specialized computer vision tasks, the proposed method may need further fine-tuning with corresponding datasets.

In the future, we plan to apply our proposed framework to larger diffusion models, such as SDXL. Since these models exhibit stronger generative priors due to the benefits from model scaling, We believe such controllable guidance will facilitate their optimal expressive capabilities in image restoration tasks. Furthermore, we intend to incorporate more low-level vision tasks into our model training, enabling its application to a broader range of high-fidelity-demanding scenarios, such as computational photography, surveillance video, and old film restoration. We hope that AdaptBIR will serve as a strong framework for future research and benefit the open-world community.

CRediT authorship contribution statement

Yingqi Liu: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Jingwen He:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Yihao Liu:** Writing – review & editing, Methodology. **Xinqi Lin:** Methodology, Formal analysis, Conceptualization. **Fanghua Yu:** Methodology, Formal analysis. **Jinfan Hu:** Visualization, Validation. **Yu Qiao:** Supervision. **Chao Dong:** Writing – review & editing, Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62276251, 62272450), the Joint Lab of CAS-HK, and the Shanghai AI Laboratory.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2024.110659.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.
[2] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, F. Wei, Textdiffuser: Diffusion models as text painters, Adv. Neural Inf. Process. Syst. 36 (2024).
[3] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, F. Wen, Pretraining is all you need for image-to-image translation, 2022, arXiv preprint arXiv: 2205.12952.
[4] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, J.-Y. Zhu, Zero-shot image-to-image translation, in: ACM SIGGRAPH 2023 Conference Proceedings, 2023, pp. 1–11.
[5] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2015) 295–307.
[6] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.
[7] Z. Tu, W. Xie, J. Cao, C. Van Gemeren, R. Poppe, R.C. Veltkamp, Variational method for joint optical flow estimation and edge-aware image restoration, Pattern Recognit. 65 (2017) 11–25.

[8] A. Lugmayr, M. Danelljan, L. Van Gool, R. Timofte, Srflow: Learning the super-resolution space with normalizing flow, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 715–732.
[9] A. Vahdat, J. Kautz, NVAE: A deep hierarchical variational autoencoder, Adv. Neural Inf. Process. Syst. 33 (2020) 19667–19679.
[10] X. Liu, Z. Ma, Z. Chen, F. Li, M. Jiang, G. Schaefer, H. Fang, Hiding multiple images into a single image via joint compressive autoencoders, Pattern Recognit. 131 (2022) 108842.
[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, Adv. Neural Inf. Process. Syst. 30 (2017).
[12] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.
[13] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, 2020, arXiv preprint arXiv:2010.02502.
[14] J. Wang, Z. Yue, S. Zhou, K.C. Chan, C.C. Loy, Exploiting diffusion prior for real-world image super-resolution, 2023, arXiv preprint arXiv:2305.07015.
[15] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, C. Dong, DiffBIR: Towards blind image restoration with generative diffusion prior, 2023, arXiv preprint arXiv:2308.15070.
[16] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.
[17] T. Yang, P. Ren, X. Xie, L. Zhang, Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization, 2023, arXiv preprint arXiv:2308.14469.
[18] S. Zhou, K. Chan, C. Li, C.C. Loy, Towards robust blind face restoration with codebook lookup transformer, Adv. Neural Inf. Process. Syst. 35 (2022) 30599–30611.
[19] Q. Yan, A. Niu, C. Wang, W. Dong, M. Woźniak, Y. Zhang, KGSR: A kernel guided network for real-world blind super-resolution, Pattern Recognit. 147 (2024) 110095.
[20] X. Wang, K. Yu, C. Dong, C.C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 606–615.
[21] K.C. Chan, X. Wang, X. Xu, J. Gu, C.C. Loy, Glean: Generative latent bank for large-factor image super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14245–14254.
[22] S. Gonzalez-Sabbagh, A. Robles-Kelly, S. Gao, DGD-cGAN: A dual generator for image dewatering and restoration, Pattern Recognit. 148 (2024) 110159.
[23] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[24] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 36479–36494.
[25] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, Y. Chen, Srdiff: Single image super-resolution with diffusion probabilistic models, Neurocomputing 479 (2022) 47–59.
[26] J. Yue, L. Fang, S. Xia, Y. Deng, J. Ma, Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models, IEEE Trans. Image Process. (2023).
[27] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, L. Van Gool, Diffir: Efficient diffusion model for image restoration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13095–13105.
[28] M. Ren, M. Delbracio, H. Talebi, G. Gerig, P. Milanfar, Multiscale structure guided diffusion for image deblurring, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10721–10733.
[29] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
[30] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1833–1844.
[31] C. Si, Z. Huang, Y. Jiang, Z. Liu, Freeu: Free lunch in diffusion U-net, 2023, arXiv preprint arXiv:2309.11497.
[32] A. Mittal, R. Soundararajan, A.C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2012) 209–212.
[33] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, C. Dong, Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild, 2024, arXiv preprint arXiv:2401.13627.
[34] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, Y. Yang, Maniqa: Multi-dimension attention network for no-reference image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1191–1200.
[35] J. Ke, Q. Wang, Y. Wang, P. Milanfar, F. Yang, Musiq: Multi-scale image quality transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5148–5157.
[36] B. Hu, G. Zhu, L. Li, J. Gan, W. Li, X. Gao, Blind image quality index with cross-domain interaction and cross-scale integration, IEEE Trans. Multimed. (2023).
[37] A. Li, J. Wu, Y. Liu, L. Li, Bridging the synthetic-to-authentic gap: Distortion-guided unsupervised domain adaptation for blind image quality assessment, 2024, arXiv preprint arXiv:2405.04167v1.
[38] X. Wang, L. Xie, C. Dong, Y. Shan, Real-esrgan: Training real-world blind super-resolution with pure synthetic data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1905–1914.
[39] A.Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR, 2021, pp. 8162–8171.
[40] K. Zhang, J. Liang, L. Van Gool, R. Timofte, Designing a practical degradation model for deep blind image super-resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4791–4800.

Yingqi Liu is currently a second-year master student at Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science. She received the B.Eng. degree from Hunan University, Changsha, China, in 2022. Her research interests focus on blind image restoration and generative model.

Jingwen He is currently a first-year Ph.D. student at The Chinese University of Hong Kong. She received the B.Eng. degree in computer science and technology from Sichuan University, China, in 2016, and the M.Phil. degree in electronic and information engineering from the University of Sydney, Australia, in 2019. Her research interests focus on multi-modal understanding and face Restoration.

Yihao Liu currently holds a research position at the Shanghai Artificial Intelligence Laboratory. He received the B.S. and Ph.D. degrees from the University of Chinese Academy of Sciences in 2018 and 2023, respectively. His research interests focus on computer vision and image processing, with a distinct focus on image and video restoration and enhancement.

Xinqi Lin is currently a first-year master student at Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science. He received the B.Eng. degree from Tianjin University, Tianjin, China, in 2023. His research interests focus on image restoration.

Fanghua Yu currently works as a research intern in the Multimedia Laboratory at Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science. He received the B.Eng. degree from Huazhong University of Science and Technology, Wuhan, China, in 2019, and the M.S. degree from Peking University, Beijing, China, in 2022. His research interests focus on image restoration.

Jinfan Hu is currently a second-year Ph.D. student at Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science. He obtained his B.Sc. and M.S. degrees in Mathematics from the University of Electronic Science and Technology of China (UESTC), Chengdu, China in 2019 and 2022. His research interests include image restoration and low-level computer vision interpretability.

Yu Qiao (Senior Member, IEEE) is a professor at Shanghai Artificial Intelligence Laboratory and Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science. He has published more than 600 articles in top-tier conferences and journals and conferences in computer science. He was awarded the Wangxuan Distinguished Youth Prize, the First Prize of the Guangdong Technological Invention Award, and the Jiaxi Lu Young Researcher Award from the Chinese Academy of Sciences. He received an Honorable Mention in Computer Vision for the AI 2000 Most Influential Scholar Award in 2022, 2023, and 2024. His research interests include computer vision, video understanding and generation, and multimodal large models.

Chao Dong is a professor at Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Science and Shanghai Artificial Intelligence Laboratory. In 2014, he first introduced deep learning method – SRCNN into the super-resolution field. This seminal work was chosen as one of the top ten "Most Popular Articles" of TPAMI in 2016. In 2021, he was chosen as one of the World's Top 2% Scientists. In 2022, he was recognized as the AI 2000 Most Influential Scholar Honorable Mention in computer vision. His current research interest focuses on low-level vision problems, such as image/video super-resolution, denoising and enhancement.