ReFIR: Grounding Large Restoration Models with Retrieval Augmentation

Hang Guo[1] **Tao Dai**[*2] **Zhihao Ouyang**[3] **Taolin Zhang**[1]
Yaohua Zha[1] **Bin Chen**[4] **Shu-tao Xia**[1,5]
[1]Tsinghua University [2]Shenzhen University [3]Aitist.ai
[4]Harbin Institute of Technology [5]Peng Cheng Laboratory
https://github.com/csguoh/ReFIR

Abstract

Recent advances in diffusion-based Large Restoration Models (LRMs) have significantly improved photo-realistic image restoration by leveraging the *internal knowledge* embedded within model weights. However, existing LRMs often suffer from the *hallucination* dilemma, *i.e.*, producing incorrect contents or textures when dealing with severe degradations, due to their heavy reliance on limited internal knowledge. In this paper, we propose an orthogonal solution called the **Re**trieval-augmented **F**ramework for **I**mage **R**estoration (ReFIR), which incorporates retrieved images as *external knowledge* to extend the knowledge boundary of existing LRMs in generating details faithful to the original scene. Specifically, we first introduce the nearest neighbor lookup to retrieve content-relevant high-quality images as reference, after which we propose the cross-image injection to modify existing LRMs to utilize high-quality textures from retrieved images. Thanks to the additional external knowledge, our ReFIR can well handle the hallucination challenge and facilitate faithfully results. Extensive experiments demonstrate that ReFIR can achieve not only high-fidelity but also realistic restoration results. Importantly, our ReFIR requires no training and is adaptable to various LRMs.

1 Introduction

Restoring a high-quality image (HQ) from its low-quality counterpart (LQ) is a well-known ill-posed problem and has been studied over the years [1, 2, 3, 4, 5, 6, 7, 8, 9]. Previous efforts attempt to handle this problem through employing various neural network architectures, including CNNs, GANs and Transformers. Recently, diffusion models [10, 11] have emerged as a promising alternative, delivering noteworthy results in real-world image restoration [12, 13, 14]. In particular, some works [15, 16, 17, 18, 19, 20] have successfully leveraged the powerful generative prior of pre-trained text-to-image (T2I) diffusion models for scaling up, to obtain the Large Restoration Model (LRM) with billions of parameters, bringing significant progress in restoring photo-realistic images.

Although scaling up restoration models has achieved remarkable success, existing LRMs may not always produce results that are faithful to the original scene, particularly when faced with heavily degraded images that surpass the LRMs' capabilities (see Fig. 1). This issue is similar to the hallucination problem observed in large language models (LLMs) [21, 22], *e.g.* ChatGPT might generate nonsense responses when highly specialized questions exceed its knowledge boundary. Similarly, if one LRM has never seen a specific scene, it will struggle to restore corresponding images faithfully. By analogizing LLM to LRM, we define the phenomenon where LRMs generate textures inconsistent with the original scene when facing hard samples as the hallucination of LRMs.

To address the hallucination problem in LRMs, simply expanding the *internal knowledge* through additional training data and parameters might seem straightforward, but it can significantly increase

[*]Corresponding author: Tao Dai (daitao.edu@gmail.com).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1478

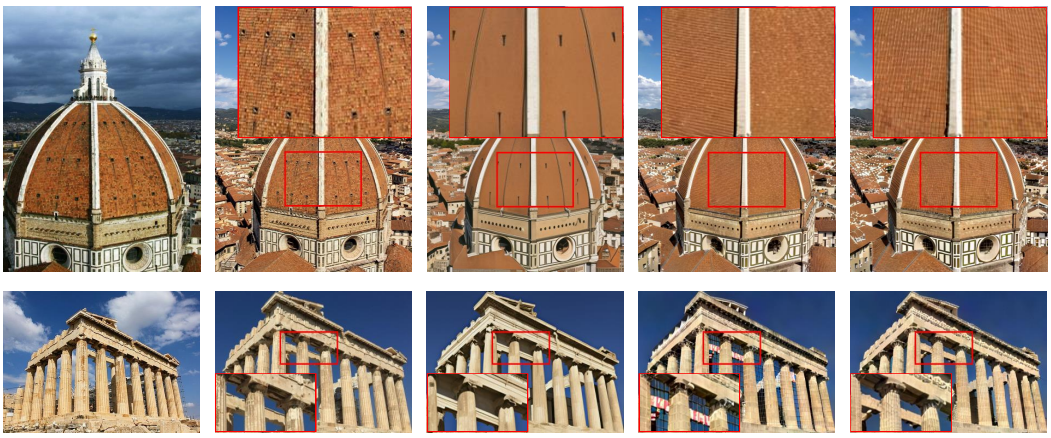


Figure 1: Existing LRMs encounter hallucination issues, *i.e.*, generating contents or details that deviate from the original scene, when dealing with challenging degradations. By incorporating the proposed ReFIR to existing LRMs [19] without any training, the additional external knowledge facilitates producing more faithful results. Please zoom in for better visualization.

computational and storage costs. Instead, this work considers another orthogonal strategy that enhances the *external knowledge* of LRMs without adding parameter counts. Drawing inspiration from the retrieval-augmented generation (RAG) used in LLMs [23, 24, 25], we aim to use the retrieved high-quality content-relevant images as external knowledge to alleviate the hallucination of LRMs. However, applying RAG to image restoration poses specific challenges. Specifically, in natural language, simply feeding the retrieved documents along with the original user query to LLMs can allow it to produce grounded responses. However, in the context of image restoration, allowing low-quality images to attend to retrieved images during their restoration process is non-trivial, which motivates us to develop novel techniques to enable LRMs to utilize external knowledge in restoration.

To this end, we delve deep into the working mechanisms of LRMs for insightful observations. Details of the experimental setup are described in Sec. 3. Our key findings indicate that the workflow of LRMs can be divided into two distinct stages: the **Denoising Structure Reconstruction** stage, during which the self-attention in the ControlNet [26] reconstructs a clear overall structure from the noised representation. After that, in the **Detail Texture Restoration** stage, the self-attention in the UNet [27] decoder fills scene-specific textures based on the denoised structure map. Based on these findings, a natural solution emerged: we can transfer high-quality, scene-specific textures from the retrieved images to the low-quality images during the detail texture restoration stage. In this way, the restored image is allowed a consistent texture with the retrieved image, thus mitigating the hallucination.

Inspired by the above observation, in this work, we propose the Retrieval-augmented Framework for Image Restoration, dubbed ReFIR, to offer a simple but effective way to expand the knowledge boundary of LRMs using the external knowledge from the retrieved images. Specifically, we first construct the retriever which employs the nearest neighbor lookup in the semantic embedding space to retrieve content-relevant reference images in the high-quality image database. After that, we develop the cross image injection which modifies the self-attention layer of original LRMs to enable the queries from the low-quality denoising chain to attend to the keys and values from the denoising chain of retrieved reference. To avoid the domain preference problem during injection, we propose separate attention to perform intra-chain and inter-chain attention, respectively. Given the spatial misalignment between the LQ and the retrieved HQ, we further adopt spatial adaptive gating to mask meaningless pixels during injection. At last, we employ the distribution alignment to narrow the domain gap between LQ and retrieved images. Thanks to the proposed ReFIR, the restoration of the LQ image can make full use of the external knowledge from the reference to generate high-fidelity images. Notably, the proposed pipeline is training-free and can be applied to multiple LRMs.

The contribution of this paper can be summarized as: **(i)** We introduce retrieval-augmented restoration, a novel concept to mitigate the hallucination problems in existing LRMs. **(ii)** We conduct an in-depth analysis of the working mechanisms of LRMs, based on which we propose a training-free framework to utilize the retrieved images. **(iii)** Extensive experiments validate that our proposed method effectively mitigates hallucination and is applicable to a broad spectrum of existing LRMs.

https://doi.org/10.52202/079017-1478

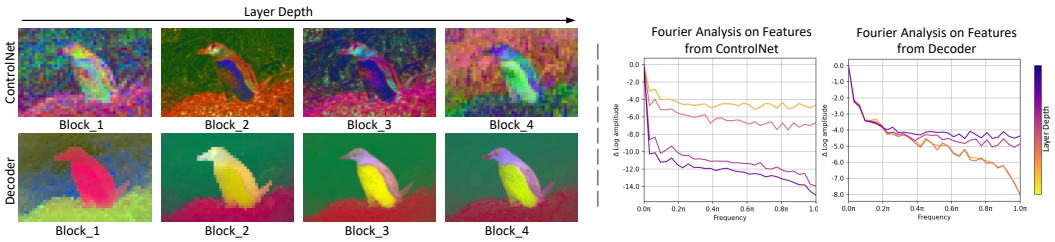


Figure 2: In-depth visualization about the working mechanism of LRM. **Left**: we use PCA to visualize the top three principal components of latent extracted from the self-attention layer of the ControlNet and UNet decoder. **Right**: quantitative power spectrum of the corresponding latent using Fourier analysis. More visualization can be found in Appendix H.

2 Related Works

2.1 Diffusion Model for Image Restoration

Diffusion models have recently achieved significant advancements across various computer vision tasks [28, 29, 30, 31, 32, 33, 34, 35, 36, 37]. In the realm of image restoration, early explorations often involved training diffusion models from scratch to obtain the restoration tailored models [38, 39, 40, 41]. While these models are capable of producing high-fidelity results, they usually fall short of generating perceptually pleasing images. To leverage the powerful generative capabilities of large pre-trained text-to-image diffusion models like Stable Diffusion [38], recent attempts [15, 16, 17, 42, 18, 19] have focused on using the ControlNet [43] with a LQ image as the condition to generate HQ images. Benefiting from the scaling law [44], these large restoration models with billions of parameters have shown impressive restoration results with photo-realistic textures and details. However, similar to the large language models, when the user query, *i.e.*, the LQ image in this setting, exceeds the knowledge boundary of the large models, the models often fail to generate meaningful or correct responses, which is unacceptable for image restoration tasks that pursue high-fidelity.

2.2 Reference-based Image Super-resolution

Compared with single image super-resolution [1, 4, 2, 3], Reference-based Image Super-Resolution (RefSR) can achieve enhanced performance by employing content-similar reference images as the additional input, and has attracted great research interests in the past few years [45, 46, 47]. For instance, C2-Matching [47] introduces a teacher-student correlation distillation and a dynamic DCN aggregation module for more precise alignment between low-quality and reference images. Following this, DATSR [48] employs reciprocal learning and SwinTransformer to further boost performance. Additionally, MRefSR [49] introduces a simple baseline to facilitate RefSR with multiple reference images. It is worth mentioning that despite both using additional images as references, our proposed retrieval augmented restoration pipeline differs from previous RefSR methods in several key aspects. Firstly, current RefSR models are typically small-scale due to limited training data, leading to performance degradation under challenging real-world conditions. Secondly, most RefSR methods can only use one single reference image and even fail to work in the absence of reference images. Thirdly, different from RefSR models that require training, our method can inject image-specific external knowledge into LRMs in a training-free manner. We give a detailed discussion about the difference in Appendix B.

2.3 Retrieval Augmented Generation

In the domain of natural language processing, Retrieval-Augmented Generation (RAG) leverages the strengths of pre-trained Large Language Models (LLMs) combined with knowledge retrieved from an external document database to enhance the quality of generated content [21, 22]. Typically, a RAG system initially retrieves documents relevant to the user's query from the knowledge base and then integrates the retrieved document along with the original user query into the LLMs without any tuning to generate a response. Even when no relevant document is available, this system can still operate by using the internal knowledge embedded in the LLMs' parameters. The integration of RAG allows LLMs to produce outputs that are not only contextually rich but also factually accurate, effectively mitigating the hallucination problem in knowledge-intensive tasks [23, 24, 25]. In this

https://doi.org/10.52202/079017-1478

Figure 3: Our ReFIR consists of two stages: the **Reference Image Retrieval** stage employs the retriever $\mathcal{R}$ to search content-relevant images from high-quality image database $\mathcal{D}$, and then the **High-fidelity Image Restoration** stage restores HQ image with reference images $\mathbf{I_R}$ as condition. The proposed framework is highly generic and can be applied to multiple existing LRMs without any training or fine-tuning.

work, we extend the concept of RAG to image processing and propose retrieval-augmented restoration to alleviate the hallucination issues in LRMs. By utilizing external textures embedded in the retrieved reference images, our tuning-free framework significantly facilitates faithful restoration results.

3 Probing Large Restoration Models

In order to manipulate the LRM so that it can utilize the retrieved reference images as external knowledge, we first delve into the underlying mechanism of existing LRMs to find useful insights. We choose the current popular LRM method SUPIR [19] as a representative. Inspired by previous image editing efforts [50, 31, 30, 29], which show that the self-attention layer of diffusion models contains important spatial correlation of an image, we thus follow this clue and employ the PCA to visualize the principal components of the latent from self-attention layers of SUPIR. We further utilize the Fourier analysis [51] to allow for quantitative results. The results are shown in Fig. 2.

It can be seen that the ControlNet of the LRM can denoise the latent as the layers deepen, facilitating the reconstruction of a clear overall structure. However, this process is accompanied by a reduction in the high-frequency meaningful texture of the original image. This qualitative visualization can be also verified by the frequency characteristic plots, with high-frequency components decaying as layer number increases. On the other hand, the role of the UNet decoder is significantly different. Based on the previous clear structural map, the decoder restores the high-frequency details and textures with the help of skip connections, which is also shown through the strengthening high-frequency component in the decoder's frequency curve.

Considering the above observations, we can divide the image restoration process of the LRM into two phases: the Denosing Structure Reconstruction phase in the ControlNet, and the Detail Texture Restoration phase in the UNet decoder. Inspired by these probing experiments, in this work, we employ the detail texture restoration nature in the self-attention layer of the decoder to inject the high-fidelity textures of retrieved images into the restoration process of the low-quality image.

4 Methodology

This work considers using retrieved reference images as an explicit part of the model. In contrast to the existing restoration pipeline, our ReFIR is parameterized by not only the internal knowledge from the network weights but also the external knowledge retrieved from suitable data representations. Fig. 3 gives an overview of our ReFIR. In the following part, we will first give the technical details of the retriever for reference image retrieval in Sec. 4.1, followed by the cross image injection to inject the external data knowledge into the restoration process of LRMs in Sec. 4.2.

4.1 Nearest Neighbor Lookup for Reference Image Retrieval

Our reference image retrieval system can be represented as a binary set $\{\mathcal{D}, \mathcal{R}\}$, where $\mathcal{D}$ is a fixed database containing a large number of HQ images, and $\mathcal{R}$ denotes a non-parametric retriever to obtain the retrieved image set $\mathbf{I_R}$ which consists of k elements and is a subset of $\mathcal{D}$ given a query LQ image $I_{LQ} \in \mathbb{R}^{3\times H\times W}$, *i.e.*, $\mathcal{R} : I_{LQ}, \mathcal{D} \mapsto \mathbf{I_R}$, where $\mathbf{I_R} \subseteq \mathcal{D}$ and $|\mathbf{I_R}| = k$. Ideally, $\mathcal{R}$ has to be

https://doi.org/10.52202/079017-1478

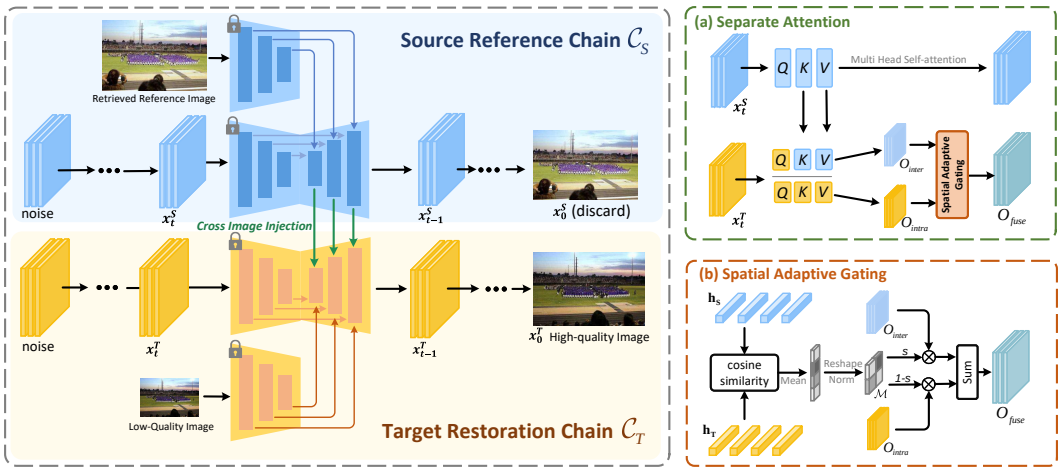


Figure 4: An illustration of cross image injection. Both $\mathcal{C}_T$ and $\mathcal{C}_S$ share the same model weights.

designed such that it provides the model with beneficial data representations from $\mathcal{D}$ to help restore images containing details faithful to the original scenes.

In this work, we implement a conceptually simple solution of $\mathcal{R}$, which uses the query image I_{LQ} to retrieve its k nearest neighbor in $\mathcal{D}$ using cosine similarity in the compact feature space derived from any feature extractors, such as VGG [52], ResNet [53] or CLIP [54]. Since the $\mathcal{D}$ is fixed, in practice, we can pre-extract and store the compact feature before training. Given a sufficiently large database $\mathcal{D}$, this strategy ensures that the set of neighbors $\mathbf{I_R}$ shares sufficient semantic consistency with I_{LQ} and thus provides useful visual information for the restoration. Although this scheme seems simple, we show that it is efficient and effective, please see Sec. 5.3 for discussion.

4.2 Cross Image Injection for High-fidelity Image Restoration

Given the retrieved reference images $\mathbf{I_R} = \mathcal{R}(I_{LQ}, \mathcal{D})$, we further propose the cross image injection to allow the original LRMs to use the external knowledge from $\mathbf{I_R}$. As shown in Fig. 4, we first construct two parallel denoising chains: the target restoration chain $\mathcal{C}_T$ which is used to restore I_{LQ}, and the source reference chain $\mathcal{C}_S$ which unfolds $\mathbf{I_R}$ into denoising time steps. After that, we introduce separate attention to separately perform attention within and between chains, followed by spatial adaptive gating to filter out irrelevant pixels. At last, we use the distribution alignment to mitigate the domain gap between chains. More details are given below.

Separate attention. To allow the $\mathcal{C}_T$ to learn the knowledge from the $\mathcal{C}_S$, an effective interaction between the latents is crucial. Inspired by the observation in Sec. 3, we aim to transfer the knowledge embedded in the self-attention layer of $\mathcal{C}_S$'s decoder to the counterpart of $\mathcal{C}_T$. To this end, we modify the original self-attention in $\mathcal{C}_T$ to our proposed separate attention. The core idea of our separate attention is to add "inter-chain cross-attention" to the original "intra-chain self-attention" so that $\mathcal{C}_T$ can attend high-quality texture knowledge from $\mathcal{C}_S$ while preserving its original features. As shown in Fig. 4(a), formally, denote Q_T, K_T, V_T as the query, key and value from the $\mathcal{C}_T$, and K_S, V_S as the key and value from the $\mathcal{C}_S$, the intra-chain self-attention preserves the original attention of $\mathcal{C}_T$ to obtain the output O_{intra}, and the inter-chain cross-attention uses the Q_T to query the K_S and V_S to facilitate $\mathcal{C}_T$ utilizing the knowledge from $\mathcal{C}_S$ to get the result O_{inter}. In short, the proposed separate attention can be formalized as follows:

$$O_{intra} = \texttt{Attention}(Q_T, K_T, V_T), \quad O_{inter} = \texttt{Attention}(Q_T, K_S, V_S). \tag{1}$$

It is worth mentioning that directly using Q_T to query the concatenate results of K_T and K_S can only yield sub-optimal results due to the domain preference issue, *i.e.*, Q_T will prefer latent from the same domain $\mathcal{C}_T$ even though $\mathcal{C}_S$ is more helpful for reconstruction. By using the proposed separate attention, the Q_T is separated to attend K_T and K_S, thus effectively mitigating this problem. We give more discussion in Sec. 5.3.

Spatial adaptive gating. We then consider fusing the separate attention results O_{intra} and O_{inter}. The main challenge is the spatial misalignment between I_{LQ} and $\mathbf{I_R}$. For instance, the same objects

https://doi.org/10.52202/079017-1478

may appear in different locations in I_{LQ} and $\mathbf{I_R}$, or some objects in I_{LQ} may not present in $\mathbf{I_R}$ and vice versa. As a result, some pixels in Q_T may not find the corresponding reference in K_S, resulting in some pixels in O_{inter} meaningless.

To address this spatial misalignment, we propose the spatial adaptive gating to selectively fuse O_{intra} and O_{inter} without introducing additional parameters (Fig. 4(b)). Specifically, given latents at specific denoising blocks from $\mathcal{C}_T$ and $\mathcal{C}_S$, respectively, we first flatten them along the spatial dimension to obtain $\mathbf{h_T}, \mathbf{h_S} \in \mathbb{R}^{C\times HW}$. Next, we compute their pixel-wise cosine similarity to obtain the similarity matrix $\mathrm{sim} \in \mathbb{R}^{HW\times HW}$. Since the i-th row of sim represents the similarity of the i-th pixel in $\mathbf{h_T}$ to all the pixels in $\mathbf{h_S}$, therefore, a large sum of the i-th row indicates a large impact of $\mathbf{h_S}$ in restoring the i-th pixel of $\mathbf{h_T}$. Following this idea, we summation over the i-th row of the sim to approximate the utility of $\mathbf{h_S}$ to the i-th pixel of $\mathbf{h_T}$. Finally, we reshape this summation results back to 2D shape and use min-max normalization to restrict the range to $[0, 1]$, to get the pixel-wise mask $\mathcal{M}$ for adaptive gated fusion:

$$O_{fuse} = (\mathbf{1} - s\mathcal{M}) \otimes O_{intra} + s\mathcal{M} \otimes O_{inter}, \tag{2}$$

where s is a user-defined scalar to control the degree to which the restored image attends the retrieved images, $\otimes$ denotes the Hardamard product, and $\mathbf{1}$ is an all one tensor with the same shape as $\mathcal{M}$.

Distribution alignment. Using the O_{fuse} to replace the original intra-chain self-attention results O_{intra} seems to be a promising way to integrate useful external knowledge from $\mathcal{C}_S$. However, it should be noticed that there is a domain gap between $\mathcal{C}_T$ and $\mathcal{C}_S$ due to the image quality and content differences, and thus a direct insertion of O_{intra} into $\mathcal{C}_T$ will result in a distribution shift of the original denoising chain in $\mathcal{C}_T$.

To this end, we propose the distribution alignment as a complementary to calibrate the distribution shift. Specifically, considering the latent in the diffusion chain is a Gaussian, we propose to use the Adaptive Instance Normalization (AdaIN) [55] to align the mean and variance of O_{fuse} to the original statistics of O_{intra}:

$$O'_{fuse} = \texttt{AdaIN}(O_{fuse}, O_{intra}), \tag{3}$$

where $\texttt{AdaIN}(u, v)$ denotes replacing the mean and variance of u with the corresponding part of v. Finally, we replace the original self-attention result in $\mathcal{C}_T$ with the well-aligned O'_{fuse} to finish the cross image injection process.

5 Experiments

5.1 Experiments Setup

Datasets and metrics. In this work, we include experiments with two difficulty levels for performance evaluation. The first setup considers restoration with manually provided ideal reference images, which share a high content similarity with the LQ image, to evaluate the ability to utilize the reference knowledge. The datasets for this setting employ the widely used RefSR dataset including CUFED5 [56, 57] and WR-SR [47], in which the reference images are already provided. Since these datasets only contain HQ images, we thus use the second-order degradation model from Real-ESRGAN [8] with $\times 4$ down-sampling scale to generate the real-world degraded images. The second setup turns to more challenging practice where the reference images have to be retrieved using the retriever, and we use the RealPhoto60 [19] which contains 60 real-world degraded images without ground truth for evaluation. And we use DIV2K [58] as the high-quality image database for retrieval and employ the image encoder of VGG16 [52] as the feature extractor. As for the evaluation metrics, we use both the fidelity metrics containing PSNR and SSIM, as well as the perceptual metrics including LPIPS [59], NIQE [60], FID [61], MUISQ [62], and CLIPIQA [63], to assess the performance of the different methods.

Implementation details. For a fair comparison, we use one reference image if not specified. Experiments with multiple reference images are given in Appendix A. Following the common practice of existing LRMs [19, 17, 18, 16], the I_{LQ} is up-sampled to the desired size using Bicubic before going through the LRMs. We use reflective padding to ensure the input size of $\mathcal{C}_T$ and $\mathcal{C}_S$ are the same. We use fixed random seeds for results reproducibility in all experiments. The hyperparameters of different baselines follow their original settings. We apply the proposed retrieval augmented restoration framework to two popular LRMs, namely SeeSR [18] and SUPIR [19], and denoted the models augmented with our ReFIR as "SeeSR+ReFIR" and "SUPIR+ReFIR", respectively.

Table 1: Quantitative comparison with state-of-the-art RefSR methods, GAN-based methods, and Diffusion-based methods on real-world image super-resolution. Our ReFIR achieves consistent performance improvements in both fidelity and perceptual quality.

Method	CUFED5					WR-SR				
	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓
C2-Matching [47]	20.77	0.5169	0.7282	8.4438	282.43	22.63	0.5627	0.7177	8.3238	157.61
DATSR [48]	20.75	0.5130	0.7301	8.6765	282.19	22.62	0.5620	0.7210	8.4329	157.54
MrefSR [49]	20.84	0.5218	0.7853	9.6524	286.44	22.68	0.5703	0.7748	9.7742	156.57
BSRGAN [9]	20.22	0.5256	0.4135	4.2204	203.17	22.07	0.5735	0.4073	3.8703	133.50
Real-ESRGAN [8]	20.31	0.5543	0.3698	3.8832	175.91	22.14	0.5974	0.3631	3.7001	97.88
StableSR [16]	20.46	0.4480	0.6532	6.3433	292.69	21.22	0.4421	0.5899	5.2040	145.07
DiffBIR [15]	19.76	0.4886	0.3820	3.5629	154.75	21.30	0.5284	0.3938	3.8736	76.05
PASD [17]	20.22	0.4959	0.5252	5.4828	208.64	21.12	0.5254	0.4292	4.2505	98.16
SeeSR [18]	19.94	0.5195	0.3660	3.7912	142.92	21.73	0.5658	0.3501	4.0155	65.78
SeeSR+ReFIR	20.32	0.5289	0.3338	3.7831	134.62	21.86	0.5664	0.3460	3.9089	61.22
Δimprovement	**+0.38**	**+0.0094**	**+0.0322**	**+0.0081**	**+8.30**	**+0.13**	**+0.0006**	**+0.0041**	**+0.1066**	**+4.56**
SUPIR	18.97	0.4665	0.4807	4.5624	168.26	20.91	0.5426	0.3791	3.7587	75.85
SUPIR+ReFIR	19.00	0.4729	0.4341	4.2085	148.69	21.02	0.5497	0.3785	3.7478	71.82
Δimprovement	**+0.03**	**+0.0064**	**+0.0466**	**+0.3539**	**+19.57**	**+0.11**	**+0.0071**	**+0.0006**	**+0.0109**	**+4.03**

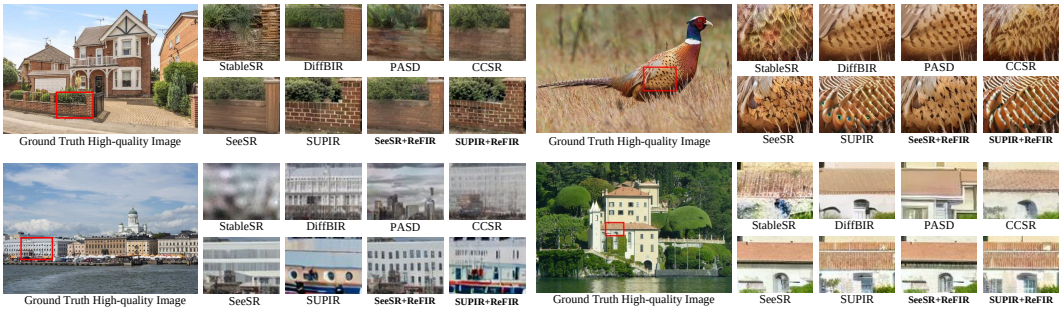


Figure 5: Quantitative comparison on RefSR dataset. The results using our ReFIR are **bolded**. Please zoom in for better visualization.

5.2 Comparison to State-of-the-Arts

Restoration with ideal reference. We first compare on the RefSR dataset with real-world degradation. The compared methods includes state-of-the-art RefSR methods [47, 48, 49], GAN-based methods [9, 8], and recent Diffusion-based methods [16, 19, 18, 17, 15]. Tab. 1 gives the results. It can be seen that our method brings significant gains in *all* metrics on both fidelity (PSNR, SSIM) and perceptual quality (LPIPS, NIQE, FID) for the LRMs. Taking SUPIR as an example, our method brings a FID improvement of even 19.57 on the CUFED5 dataset. Moreover, similar performance gains can also be observed in SeeSR. For instance, equipping our ReFIR to SeeSR can lead to 0.38dB PSNR improvement, demonstrating the generalization of our ReFIR. It is noteworthy that the above superiority is obtained without any training or fine-tuning. Moreover, we also give visual comparisons in Fig. 5, and it can be seen that our method can generate details that are faithful to the original scene with the help of external knowledge from retrieved reference images.

Restoration in the wild. The above experiments on RefSR datasets focus on utilizing the already provided reference images from the dataset, which applies when the user has relevant HQ images. In this section, we turn to more challenging scenarios in which the reference image has to be obtained by retrieval. Since the ground truth of RealPhoto datasets is unavailable, we use non-reference image quality assessment metrics, *i.e.* NIQE, MUSIQ, and CLIPIQA for evaluation. As shown in Tab. 2, our approach continues to produce significant gains over its non-ReFIR counterparts. For instance, our SeeSR+ReFIR surpasses the original SeeSR by 0.2866 NIQE and 1.59 MUSIQ. Since the retrieved image can not serve as an ideal reference, the above favorable results demonstrate the robustness of our ReFIR in the face of real-world retrieved images. We also give quantitative results in Fig. 6. Even under severe real-world degradation, our method maintains good perceptual quality.

https://doi.org/10.52202/079017-1478

Table 2: Quantitative comparison on real-world degradation with RealPhoto datasets.

Metrics	StableSR [16]	DiffBIR [15]	PASD [17]	CCSR [42]	SeeSR [18]	SUPIR [19]	SeeSR+ReFIR (Ours)	SUPIR+ReFIR (Ours)
NIQE↓	3.7695	2.8458	5.1603	5.5082	4.7432	3.5076	4.4566(+0.2866)	3.4593(+0.0483)
MUSIQ↑	51.95	65.20	49.01	32.26	55.54	59.84	57.13(+1.59)	60.49(+0.65)
CLIPIQA↑	0.6852	0.7845	0.5863	0.4568	0.6575	0.5692	0.6732(+0.0157)	0.5722(+0.003)

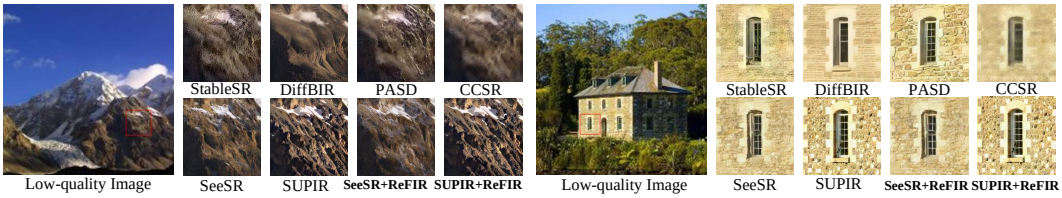


Figure 6: Quantitative comparison on RealPhoto dataset. More results are provided in Appendix H.

Table 3: Comparison of model complexity before and after incorporating our ReFIR. We use an input image with the resolution of 2048×2048 to evaluate the GPU memory and the inference time on one single 80G NVIDIA A100 GPU.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	#param↓	GPU Memory↓	Inference Time↓
SeeSR [18]	19.94	0.5195	0.3660	142.92	2.04B	24.4G	76.5s
SUPIR [19]	18.97	0.4665	0.4807	168.26	3.87B	37.3G	146.4s
SeeSR+ReFIR	20.32	0.5289	0.3338	134.62	2.04B	40.9G	170.7s
SUPIR+RefIR	19.00	0.4729	0.4341	148.69	3.87B	51.4G	322.8s

Complexity analysis. Tab. 3 gives the comparison of the computational complexity, including the number of parameters, GPU cost, and the inference latency. We also give the restoration performance for a more comprehensive comparison. As for the parameters, our ReFIR can facilitate both fidelity and realistic image restoration using the same #param as the original base LRMs. For the GPU memory, since our ReFIR uses two images as input, *i.e.*, one LQ image, and one reference image, the GPU cost will become larger than the original one. For instance, it rises 1.38 times the increase of SUPIR+ReFIR than the original SUPIR model. Moreover, the inference time also increases due to more inputs as well as the additional interaction between two chains. In the future, we will delve deep into the effective utilization of retrieved images while maintaining efficiency.

5.3 Ablation Studies

Effectiveness of the reference retriever. In order to obtain content-relevant retrieved images, we present a simple but inference-efficient retriever $\mathcal{R}$ that uses the high-level semantic vectors from the pre-trained deep models for similarity matching in the high-quality image dataset $\mathcal{D}$. Despite the simple design, we here demonstrate its effectiveness in Fig. 7. Since semantically consistent images usually contain similar textures, *e.g.*, the texture in the first elephant image can help in the restoration of the LQ elephant image, and thus the proposed retriever can yield satisfactory retrieval results. Although texture-based retrieval may be a better choice for image restoration, it usually necessitates additional training of new retrieval models. For simplicity, we adopt semantic-based retrieval and leave the exploration of more advanced reference retrievers for future work.

Ablation on cross image injection. In the proposed cross image injection, we use separate attention (SA), spatial adaptive gating (SG), and distribution alignment (DA) for effective external knowledge injection. Here, we ablate to validate the effectiveness of different components. We use SUPIR+ReFIR as a representative on the CUFED5 dataset and use the scalar weighted sum when SG is removed. The results are shown in Tab. 4. One can see that using fixed scalar weights instead of spatial adaptive gating results in a 0.18 NIQE drop. This is because not all pixels of the reference image are useful, and thus fine-grain gated mask is needed. Moreover, removing the distribution alignment also impairs performance, *e.g.*, 4.36 FID drop, since the distribution of raw fusion results O_{fuse} does not match $\mathcal{C}_T$, and directly inject O_{fuse} to the denoising chain of $\mathcal{C}_T$ can cause sub-optimal results.

https://doi.org/10.52202/079017-1478

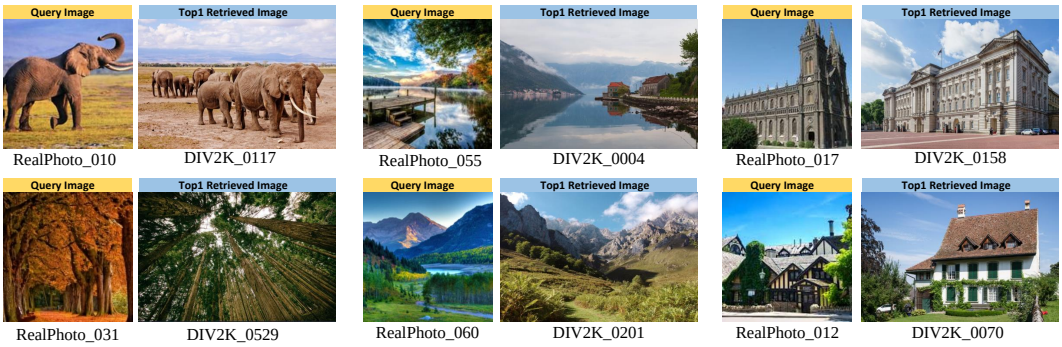


Figure 7: The retrieval results with RealPhoto60 dataset [19] as the query images and DIV2K [58] as the HQ image database.

Table 4: Effectiveness of different components in cross image injection.

SA	SG	DA	PSNR↑	SSIM↑	NIQE↓	FID↓
			18.97	0.4665	4.5624	168.26
✓		✓	19.09	0.4799	4.3893	150.81
✓	✓		19.12	0.4724	4.2275	153.05
✓	✓	✓	19.00	0.4729	4.2085	148.69

Table 5: Ablation experiments on different positions of cross-image injection.

Injection Position	PSNR↑	SSIM↑	NIQE↓	FID↓
No Injection	18.97	0.4665	4.5624	168.26
Encoder&Decoder	19.09	0.4773	4.5241	153.43
Encoder Only	19.08	0.4689	4.6977	168.26
Decoder Only	19.00	0.4729	4.2085	148.69

Domain preference problem. The motivation behind the proposed separate attention is to address the domain preference problem, *i.e.*, the attention in $\mathcal{C}_T$ will prefer to use latent from the same chain even though the latent from $\mathcal{C}_S$ is more helpful for reconstruction. To verify the existence of the domain preference, we use the ground truth I_{HR} as the input of $\mathcal{C}_S$ and compute the normalized attention scores between Q_T and K_T, Q_T and K_S. It can be seen in Fig. 8 that even using the spatially strictly aligned I_{HQ} as the reference, Q_T still has significantly high attention for the latent from the same chain, indicating that the domain preference problem interferes with the $\mathcal{C}_T$'s utilization of external knowledge in $\mathcal{C}_S$. By contrast, the proposed separate attention can effectively mitigate this problem by forcing the Q_T to separately attend K_T and K_S.

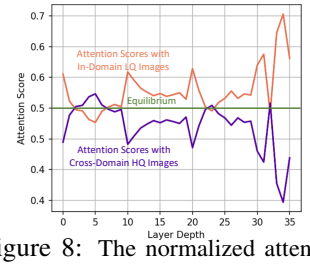


Figure 8: The normalized attention scores are obtained by averaging all samples and all time steps.

Other choices on injection position. In Sec. 3, we find the diffusion decoder is responsible for restoring textures. Based on this observation we propose to apply cross-image injection on the UNet decoder. Here, we ablate to analyze the impact of different cross-image injection positions. The results are shown in Tab. 5. It can be seen that performing cross-image injection only on the encoder will cause 19.57 FID drops. This is because the encoder focuses on the structure reconstruction, thus transferring the structure of $\mathcal{C}_S$ will destroy the layout of the $\mathcal{C}_T$. Moreover, performing injection only in the decoder achieves the best results since it can transfer the high-quality textures from the $\mathcal{C}_S$. Due to the page limit, more ablation experiments can be seen in Appendix C.

5.4 Discussions

What is the impact of the control scale? The scale s in Eq. (2) can control the extent to which the LRMs use external knowledge from the retrieved reference image for restoration. Here, we conduct an ablation study to explore the effect of s. The results are shown in Fig. 9. It can be seen that when s takes smaller values, the model mainly uses the internal knowledge embedded in its own parameters, which can make the model hallucinate when the degradation is severe. For example, the model produces incorrect textures when $s = 0$. As s increases, the model starts to use external knowledge from the retrieved reference image, from which the model's hallucination problem can be alleviated. We also provide quantitative ablation experiments on s in Appendix C.

How much do the reference images affect performance? In the proposed framework, the retrieved images $\mathbf{I_R}$ is crucial in alleviating hallucinations. Here, we try to answer the role of $\mathbf{I_R}$ during restoration process, by manually controlling different types of retrieved images. As shown in Tab. 6,

https://doi.org/10.52202/079017-1478

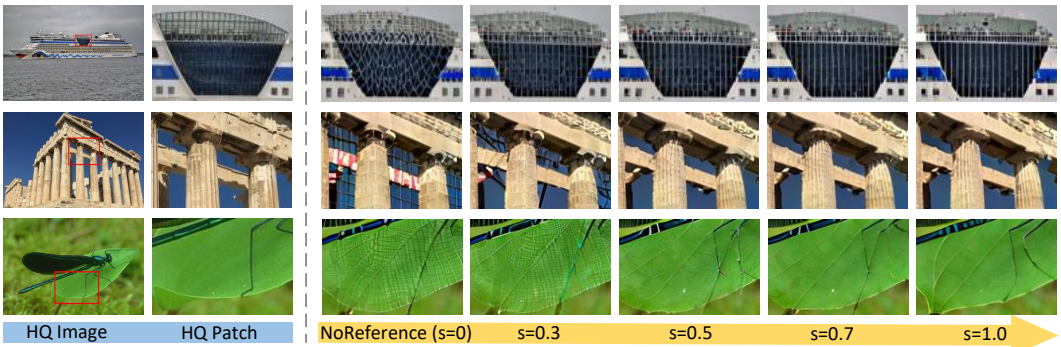
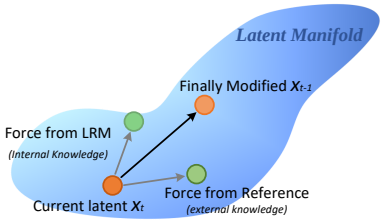


Figure 9: Ablation visualization on the control scale s. As s increases, the LRM utilizes the external knowledge from retrieved reference images to mitigate hallucination. Zoom in for better effects.

Table 6: The performance impact of reference images. NoRef means no reference image is used. HQRef denotes the corresponding I_{HQ} is used as the reference. SelfRef represents using $\times 4$ bicubic upsampling of I_{LQ} for reference. Random means randomly selecting a high-quality image as the reference.

Settings	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓
NoRef	18.97	0.4665	0.4807	4.5624	168.26
HQRef	19,41	0.5033	0.3928	4.0764	137.52
SelfRef	19.16	0.4795	0.4761	4.5501	163.94
Random	19.53	0.5138	0.5354	5.3796	223.47
Baseline	19.00	0.4729	0.4341	4.2085	148.69

Figure 10: An explanation of how the proposed retrieval augmented framework affects the restoration process of existing LRMs.



we find that using the exact ground truth I_{HQ} as the $\mathbf{I_R}$ can further improve the performance, which can be seen as an ideal up-bound. Interestingly, using I_{LQ} itself as its own retrieved image instead brings a slight improvement compared with no retrieval, which we attribute to the regularization effect from the distribution alignment strategy. Finally, randomly selecting a high-quality reference image even resulted in a huge performance degradation, suggesting that the content correlation is more important than the image quality for a favorable retrieved reference image.

How does the proposed ReFIR work? Extensive experiments have shown the state-of-the-art performance of our ReFIR. However, it seems not straightforward to understand how the retrieved reference images influence the image restoration process of the original LRMs. Here, we give an intuitive explanation. As shown in Fig. 10, for the latent at the t-th time step on the latent manifold, there are two forces in different directions pulling it to produce the latent at the next $t-1$-th time step. One force is from the internal knowledge of frozen weights in LRMs, and the other is the external knowledge from the retrieved reference image through the proposed cross image injection mechanism. These two forces ultimately determine the latent of the next time step. Therefore, a restored image from our ReFIR can utilize both the internal knowledge in the original LRMs as well as the external knowledge in the retrieved image, thus alleviating the hallucination of the LRMs.

6 Conclusion

This paper presents ReFIR, a training-free and generic framework that can alleviate the hallucination of LRMs to facilitate high-fidelity and photo-realistic restoration results through retrieval augmentation. We introduce the nearest neighbor lookup as a simple retriever to obtain relevant high-quality images and further propose the cross-image injection which employs separate attention to transfer knowledge while avoiding the domain preference problem, the spatial adaptive gating to address the spatial misalignment, and the distribution alignment to mitigate the domain gap during injection. Through expanding the knowledge boundary using the additional external knowledge from retrieved images, our ReFIR exhibits significant improvements on both fidelity and perceptual quality, as demonstrated through extensive qualitative and quantitative evaluations. Moreover, with its training-free and generic nature, our ReFIR can be easily applied to multiple LRMs.

https://doi.org/10.52202/079017-1478

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309,62171248), Shenzhen Science and Technology Program (JCYJ20220818101014030, JCYJ20220818101012025), and the PCNL KEY project (PCL2023AS6-1).

References

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014. 1, 3

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 1, 3

[3] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024. 1, 3

[4] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, pages 3929–3938, 2017. 1, 3

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 1

[6] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 1

[7] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 1

[8] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 1, 6, 7

[9] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 1, 7

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[12] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pages 9935–9946, 2023. 1

[13] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *NeurIPS*, 35:23593–23606, 2022. 1

[14] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 1

[15] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 1, 3, 7, 8

[16] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 1, 3, 6, 7, 8

[17] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 1, 3, 6, 7, 8

[18] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. *arXiv preprint arXiv:2311.16518*, 2023. 1, 3, 6, 7, 8, 15

[19] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling Up to Excellence: Practicing model scaling for photo-realistic image restoration in the wild. *arXiv preprint arXiv:2401.13627*, 2024. 1, 2, 3, 4, 6, 7, 8, 9, 15, 21

 https://doi.org/10.52202/079017-1478

[20] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Distillation-free one-step diffusion for real-world image super-resolution. *arXiv preprint arXiv:2410.04224*, 2024. 1

[21] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FactScore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023. 1, 3

[22] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022. 1, 3

[23] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. 2, 3

[24] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023. 2, 3

[25] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Retrieval-based language models and applications. In *ACL*, pages 41–46, 2023. 2, 3

[26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 3

[29] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023. 3, 4, 16

[30] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. *NeurIPS*, 36, 2024. 3, 4, 16

[31] Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. PhotoSwap: Personalized subject swapping in images. *NeurIPS*, 36, 2024. 3, 4, 16

[32] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *NeurIPS*, 35:15309–15324, 2022. 3

[33] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your Pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI*, volume 38, pages 4117–4125, 2024. 3

[34] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-Your-Click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*, 2024. 3

[35] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-Your-Emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 3

[36] Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. MagicStick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*, 2023. 3

[37] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. COVE: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[39] Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. ResDiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8975–8983, 2024. 3

[40] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 36, 2024. 3

[41] Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. *NeurIPS*, 36, 2024. 3

[42] Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability of diffusion models for content consistent super-resolution. *arXiv preprint arXiv:2401.00877*, 2023. 3, 8

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3

[44] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

[45] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 3

[46] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-SR: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, pages 6368–6377, 2021. 3

[47] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via C2-matching. In *CVPR*, pages 2103–2112, 2021. 3, 6, 7, 15

[48] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *ECCV*, pages 325–342. Springer, 2022. 3, 7, 15

[49] Lin Zhang, Xin Li, Dongliang He, Fu Li, Errui Ding, and Zhaoxiang Zhang. Lmr: A large-scale multi-reference dataset for reference-based super-resolution. In *ICCV*, pages 13118–13127, 2023. 3, 7, 15

[50] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 4, 16

[51] Namuk Park and Songkuk Kim. How do vision transformers work? In *ICLR*, 2021. 4

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763. PMLR, 2021. 5

[55] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 6

[56] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, pages 7982–7991, 2019. 6

[57] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. Event-specific image importance. In *CVPR*, pages 4810–4819, 2016. 6

[58] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, pages 114–125, 2017. 6, 9

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[60] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[61] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 6

[62] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 6

 https://doi.org/10.52202/079017-1478

[63] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI*, volume 37, pages 2555–2563, 2023. 6

[64] Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE TPAMI*, 43(4):1445–1451, 2020. 17

[65] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. VoP: Text-video co-operative prompt tuning for cross-modal retrieval. In *CVPR*, pages 6565–6574, 2023. 17

Appendix

A Adaptive Multi-reference Injection

Technical details. In the main paper, we mainly focus on the case of one single retrieved image. However, in practice, there may be multiple available reference images at hand, and using multiple reference images for resemblance could intuitively gain better performance. To this end, we extend the original cross-image injection to allow to incorporation of multiple reference images for reconstruction. Our key idea is to modify the scale factor s in Eq. (2) from a scalar into a vector: $\mathbf{s} = \{s_1, s_2, \cdots, s_k\}$, where $\sum s_n = 1$. Each s_n can be obtained by computing the cosine similarity between I_{LQ} and the corresponding n-th retrieved image in $\mathbf{I_R}$ followed by Softmax normalization. Then we can modify the original single-reference cross-image injection of Eq. (2) to the following multi-reference version:

$$O_{fuse} = (\mathbf{1} - \sum_{n=1}^{k} s_n \mathcal{M}_n) \otimes O_{inter} + (\sum_{n=1}^{k} s_n \mathcal{M}_n) \otimes O_{intra}, \tag{4}$$

where $\mathcal{M}_n$ denotes the gated mask of the n-th reference image.

Experiments with multiple reference images. For experiments with multiple reference images, we use SUPIR+ReFIR as a representative. Since the CUFED5 dataset contains multiple reference images, we directly use the provided images as the retrieved reference for reproducibility. Tab. 8 gives the results. It can be seen that using multiple reference images produces better results than one single reference image, *e.g.* the 2.08 improvement in FID. However, it is worth noting that the marginal gain from adding reference images is diminishing, accompanied by a notable increase in computational cost. Therefore, in practice, we use one single reference image to balance the model performance and inference efficiency.

B More Discussions

Difference from the other methods. Our ReFIR uses retrieved images as the reference for high-fidelity restoration. Despite both RefSR methods and ours appears the reference image, we would like to clarify the difference between our ReFIR and previous RefSR methods. **Firstly**, current RefSR models [47, 48, 49] are typically small-scale (#param <50M) and use simple Bicubic degradation, while our ReFIR focuses on the recent diffusion-based large-scale restoration model (#param >1B) for more challenging real-world SR. **Secondly**, most RefSR methods can only use one reference image and even fail to work in the absence of reference images, by contrast, our ReFIR can flexibly use $0 \sim k$ images. **Thirdly**, different from RefSR models that require training, our method can be applied in various LRMs in a training-free manner.

Performance under extreme conditions. Since our ReFIR relies on the retrieved images, it is interesting to explore extreme situations when highly relevant and high-quality reference images are scarce or even unavailable. To this end, we introduce the fallback strategies to handle this situation. Specifically, since our method does not modify the parameters of LRMs, we can directly use the original inference pipeline of the LRM without using reference images. We denote this as origin_lrm. In addition, we also use the BLIP model to caption the LR image to obtain the text prompt, which will then be fed into the StableDiffusion2.0 model to generate semantic-similar high-quality images as the reference. We denote this as gen_ref. We use SeeSR [18] as a representative, on the real-world degradation dataset RealPhoto60 [19]. We first give the results in which all LR images adopt the fallback strategies in Tab. 7. It can be seen that using the SD2.0 generated images as the fallback image can bring slightly improvement compared with noReference. After that, we further develop task-oriented adaptive strategies to enhance the performance of ReFIR in real-world scenarios. In detial, we respectively use the retrieved images and the gen_ref to generate the results. And then we select the one with a larger task score as the final result. We denote it as ada_gen_ref. From Tab. 7, it can be seen that the task-oriented strategy achieves a significant

Table 7: Results of all LR images using the fallback strategies.

setup	NIQE↓	MUSIQ↑	CLIPIQA↑
origin_lrm	4.7432	55.54	0.6575
gen_ref	4.6923	55.98	0.6602
ada_gen_ref	4.3464	57.68	0.6942
ReFIR	4.4986	57.01	0.6759

https://doi.org/10.52202/079017-1478

Table 8: Experiments on extending to use multiple retrieved images for restoration. The inference time is evaluated on A100 GPU.

settings	GPU Memory↓	Inference Time↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓
NoRef	37.3G	146.4s	18.97	0.4665	0.4807	4.5624	168.26
OneRef	51.4G	322.8s	18.86	0.4623	0.4492	4.2317	156.10
TwoRef	65.5G	499.2s	18.78	0.4676	0.4296	4.2315	154.02

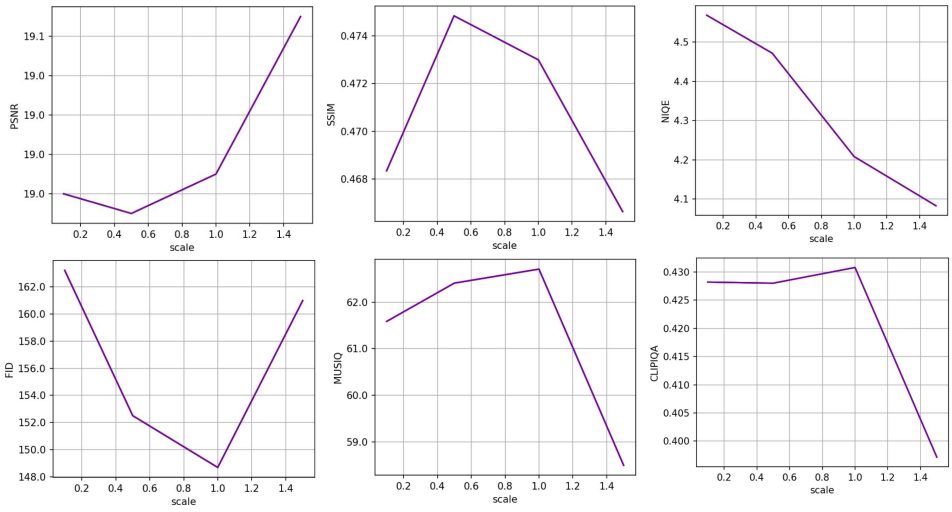


Figure 11: Quantitative ablation results on the control scales using SUPIR+ReFIR on CUFED5.

performance improvement against previous ReFIR baselines, *e.g.*, 0.0183 CLIPIQA improvements, due to the fact that it works in the output end. However, this setup is accompanied by a larger inference time, and further acceleration on this fallback strategies can be an promising future work.

Computational overhead from retrieval and attention modification. Since we employ additional Ref images as input and modify the attention layers, we adiscuss the impact of these trchnuques on the inferenve efficiency. First, in order to reduce the computational overhead of the retrieval process, we pre-calculated the feature vectors of all images in the retrieval database before inference. Furthermore, the cosine similarity between the LR image vectors and all retrieval vectors is computed in parallel. These strategy results in an almost negligible (less than 3% inference time) cost of computational overhead. Second, the modification of self-attention layers only happens in the last 20 timestep in the decoder layers, *i.e.*, only 12% attention layers are modified while the left is kept intact. These analysis is also supported by practice, in which we find these two process only take up <5% inference time, with most computational cost coming from the original LRM. Future LRM acceleration (*e.g.* pruning, quantization, one-step diffusion) will benefit our ReFIR, and we will explore more efficient implementation in the future.

Why use the self-attention as the external knowledge? In the proposed cross-image injection, we use the features of the self-attention layer of the $\mathcal{C}_S$'s decoder as external knowledge to guide $\mathcal{C}_T$ to produce textures faithful to the original scene. Here, we give the reason behind this. **Firstly**, previous image-to-image efforts [29, 31, 30, 50], *e.g.*, image editing, has demonstrated through extensive experiments that the self-attention layer of the diffusion model contains important spatial correlations in images, which inspired us to follow this clue to utilize this prior. **Secondly**, leveraging the attention mechanism allows $\mathcal{C}_T$ to query features in $\mathcal{C}_S$ without any training, whereas using features from other parts of $\mathcal{C}_S$ may require introducing additional training.

What about the quality of cross-image attention? In the proposed cross-image injection, the inter-chain attention is used to perform attention between Q_T and K_S. Considering the domain gap between $\mathcal{C}_T$ and $\mathcal{C}_S$ due to the input quality difference, one may ask whether the results of the inter-chain attention are meaningful. Here, We visualize the attention map to validate the effectiveness of the inter-chain attention (see Fig. 13). It can be seen that for a given query pixel query in $\mathcal{C}_T$,

the inter-chain attention can effectively overcome the spatial misalignment, and find relevant pixel features in $\mathcal{C}_S$ for reference.

C More Ablation Results

Quantitative ablation on the control scale. We also provide quantitative ablation results on the control scale s in Fig. 11. It can be seen that when s is too small, the LRM will mainly use the knowledge contained within its parameters to restore high-quality images, which can lead to performance degradation due to the hallucination problem. On the other hand, when s is too large, the LRM will overuse the content in the retrieved reference image, thus producing patterns that are not present in the original LQ image. In practice, we adopt a moderate $s = 0.5$ to trade off the hallucination and the overuse of the reference image.

Other choices for cross image injection. The proposed cross image injection mitigates the domain preference problem by using separate attention to promote latent in $\mathcal{C}_T$ to attend $\mathcal{C}_S$. Here, we conduct ablation to study the impact of different design choices of cross image injection. As shown in Tab. 9, directly replacing the original self-attention results from O_{intra} in $\mathcal{C}_T$ with corresponding latent in $\mathcal{C}_S$ causes severe performance degradation, due to the significant loss of original knowledge in $\mathcal{C}_T$. In addition, using Q_T to query the concatenation results of K_T and K_S also causes a performance drop, which further confirms that the domain preference problem, *i.e.*, Q_T prefers to use latent from the same chain $\mathcal{C}_T$, even though $\mathcal{C}_S$ is more helpful for reconstruction.

Table 9: Ablation experiments on different cross image injection designs.

settings	PSNR↑	SSIM↑	NIQE↓	FID↓
replace	18.84	0.4385	4.26	182.82
concat	18.89	0.4691	4.19	156.03
baseline	19.00	0.4729	4.21	148.69

D Statistical Significance on Performance

In Tab. 1 of the main paper, we give the performance gains of incorporating the proposed ReFIR into the existing LRMs. Considering the randomness of the generative models, we give the performance fluctuations of ReFIR under multiple trials with exactly the same experimental setting and random seed. The results are given in Tab. 10. It can be seen that the randomness of the diffusion-based generative model is very small when using a fixed seed, reducing the disturbance from noise errors for evaluation. In addition, we further use hypothesis testing to verify the significance of performance gains, and the test results reject the original hypothesis H0 at 95% confidence level on all metrics and datasets, indicating that the performance gains from the proposed method are statistically significant.

E Extension to Specific Restoration Scenarios

An important application of our method is in scenarios with high fidelity demands, such as scene text images with a specific stylistic structure, or face images with identity preservation, and here we preliminarily explore the application of the proposed ReFIR to real-world face image restoration. The results are given in Fig. 15. It can be seen that by using a high quality image of a specific person's identity as a reference, the resulting restoration results can better preserve the person's attributes. However, it should be noted that this experiment is just a preliminary attempt, and we will leave the further improvement of our ReFIR for specific downstream restoration tasks for future work.

F Limitation and Future Works

Although the proposed ReFIR can effectively mitigate the hallucination of LRMs by introducing external knowledge from retrieved reference images, the proposed framework can be further improved in the following aspects. First, since the computational complexity of the current LRMs is costly, the computational complexity will be further increased when using the proposed method, which may hinder the use of resource-constrained mobile devices. With the advent of accelerated diffusion-based image restoration methods in the future, we believe that the proposed method can further improve its efficiency. In addition, this paper proposes a simple retriever based on semantic vector matching for presentation. With the development of image retrieval techniques [64, 65], designing specialized

https://doi.org/10.52202/079017-1478

Table 10: Performance fluctuations under different experiment trials. We use ten trails to obtain a stable fluctuation range.

settings	CUDED5					WR-SR				
	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓
SeeSR+ReFIR	20.32	0.5289	0.3338	3.7831	134.62	21.86	0.5664	0.3460	3.9089	61.22
Numerical fluctuations	±0.0013	±0.0001	±0.0002	±0.0001	±0.03	±0.001	± 0.0001	±0.0002	±0.0001	± 0.02
SUPIR+ReFIR	19.00	0.4729	0.4341	4.2085	148.69	21.02	0.5497	0.3785	3.7478	71.82
Numerical fluctuations	±0.005	±0.0003	±0.001	±0.012	±0.5	±0.002	±0.0003	±0.0006	±0.007	±0.3

retrievers, *e.g.*, using textures as key matching cues, will further improve the performance. Finally, for some slightly degraded images, which can already be handled well by only using the internal knowledge of the LRMs, designing hyper-networks to adaptively decide whether to use retrieval augmentation or not is also promising. We leave the above considerations for future work.

G Broader Impact

The development of our ReFIR offers significant positive societal impacts, including advancements in medical imaging, historical preservation, and media restoration by enhancing the fidelity and realism of image restoration. However, it also poses potential negative societal impacts, such as the misuse of improved restoration capabilities for generating disinformation, deepfakes, and surveillance, raising ethical concerns about privacy, security, and fairness. To mitigate these risks, implementing safeguards like gated releases of models, monitoring mechanisms, and transparency in model training and deployment is crucial. Continuous ethical evaluation and adherence to strict guidelines are essential to prevent potential harms.

H Additional Visual Results

In this section, we provide more visual results, which are organized as follows:

- In Fig. 12, we give more samples of the PCA visualization on the top three principal components of the self-attention layer latent.
- In Fig. 13, we give a visualization of the attention map from the cross-image injection, to help better understand the feasibility of cross-image attention.
- Fig. 14 gives more quantitative comparison results against the state-of-the-art method on real-world degradation without ground truth.
- Fig. 15 gives the visualization results of the extension experiments of applying the proposed ReFIR to blind face image restoration.

https://doi.org/10.52202/079017-1478

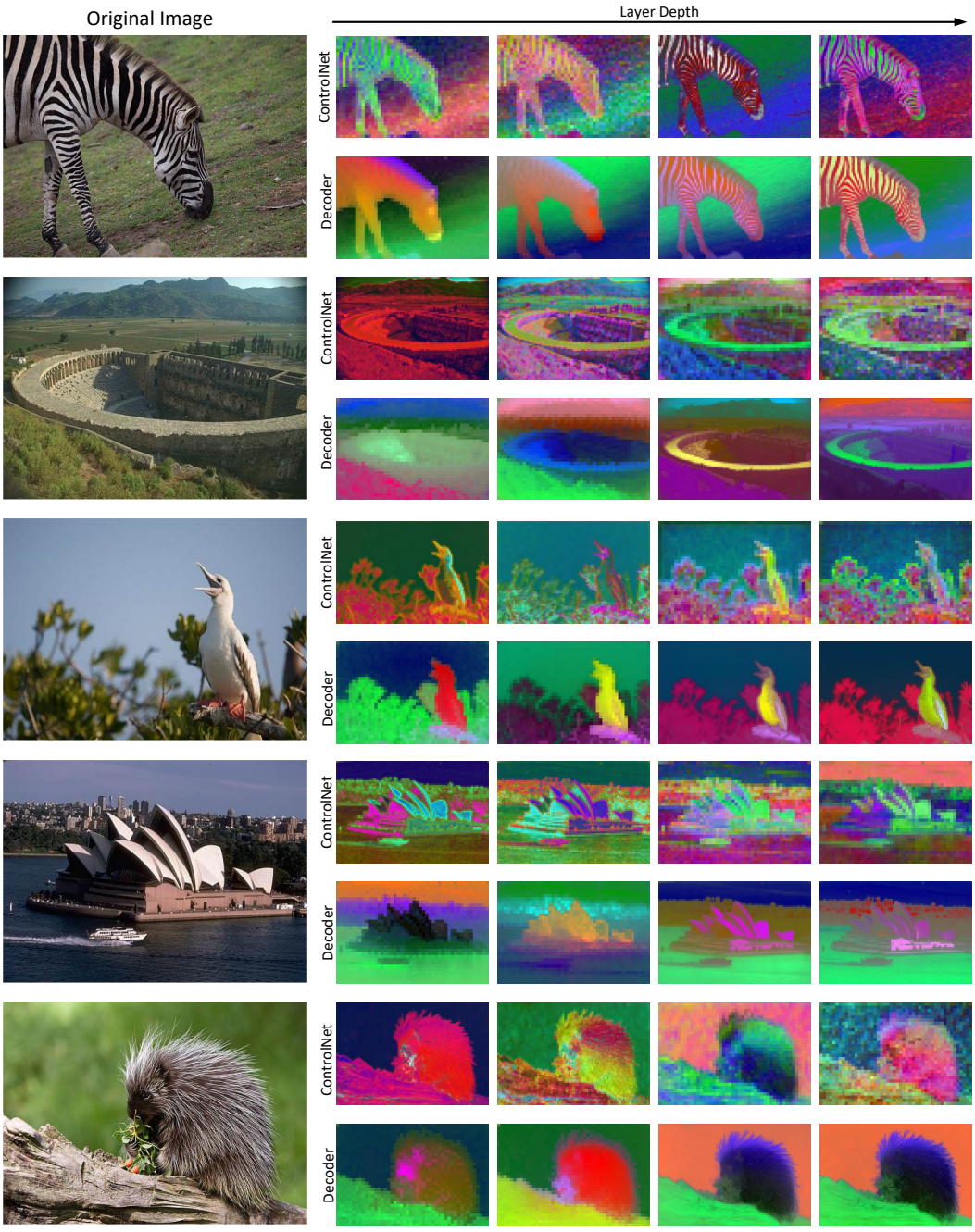


Figure 12: Additional visualization on the top three principal components of the self-attention layer latent of PCA. The latent is extracted from the first self-attention layer within blocks of the control net and unet decoder.

https://doi.org/10.52202/079017-1478

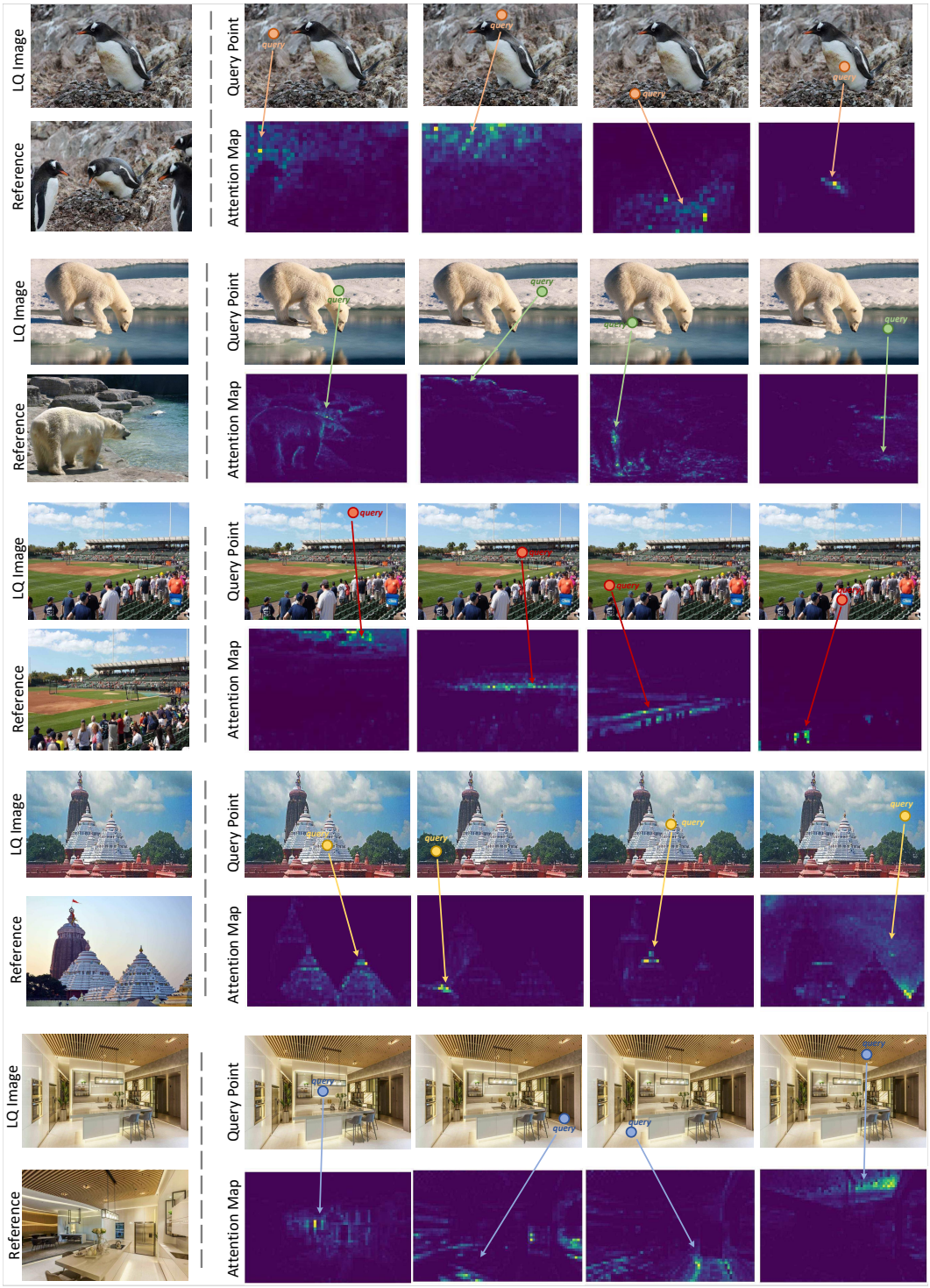


Figure 13: Additional visualization on the attention maps from the cross image injection. It can be seen that the query pixel in the $\mathcal{C}_T$ can well attend similarly region from $\mathcal{C}_S$.

https://doi.org/10.52202/079017-1478

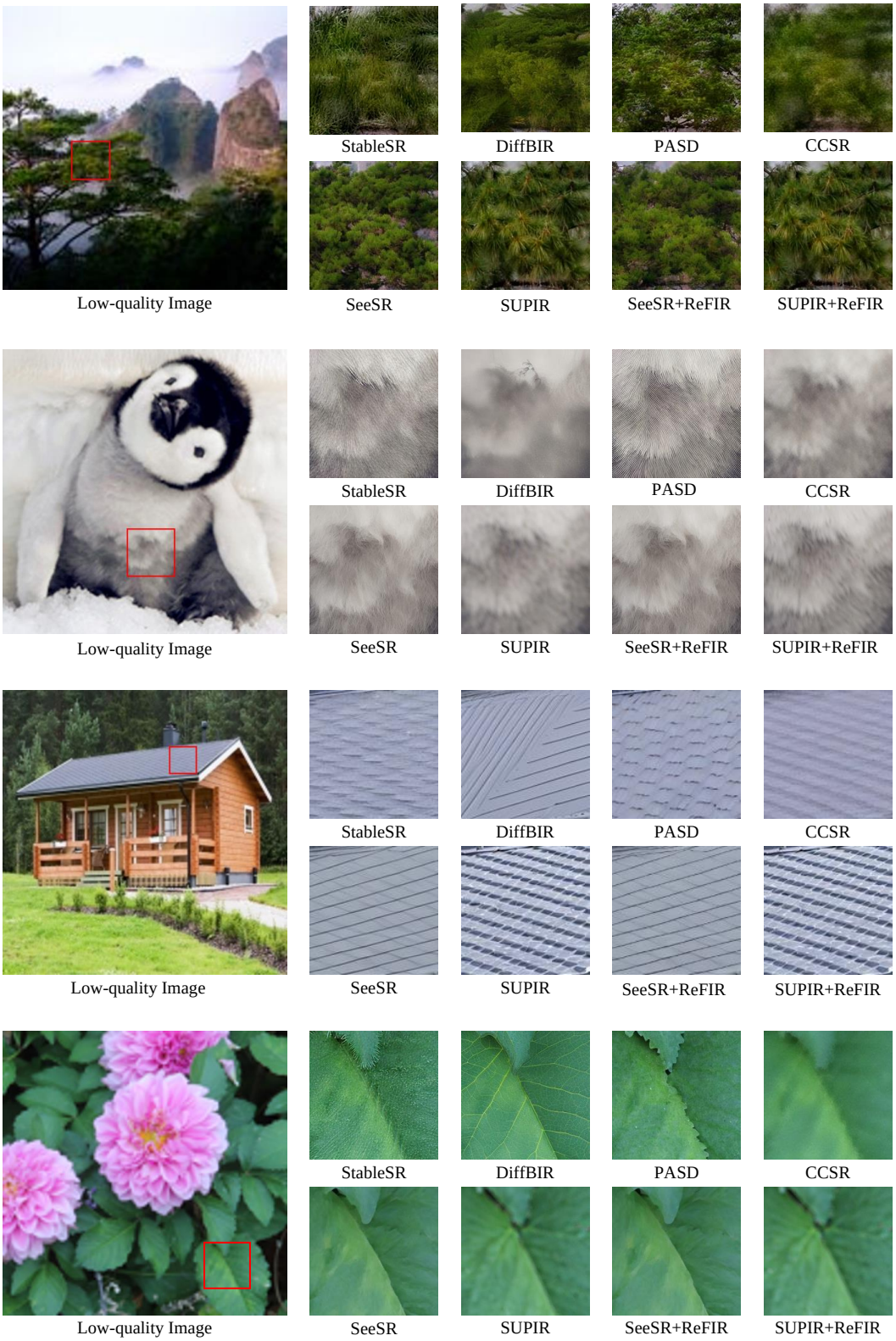


Figure 14: Additional qualitative comparison with state-of-the-art methods on RealPhoto60 [19]. Please zoom in for better effects.

https://doi.org/10.52202/079017-1478

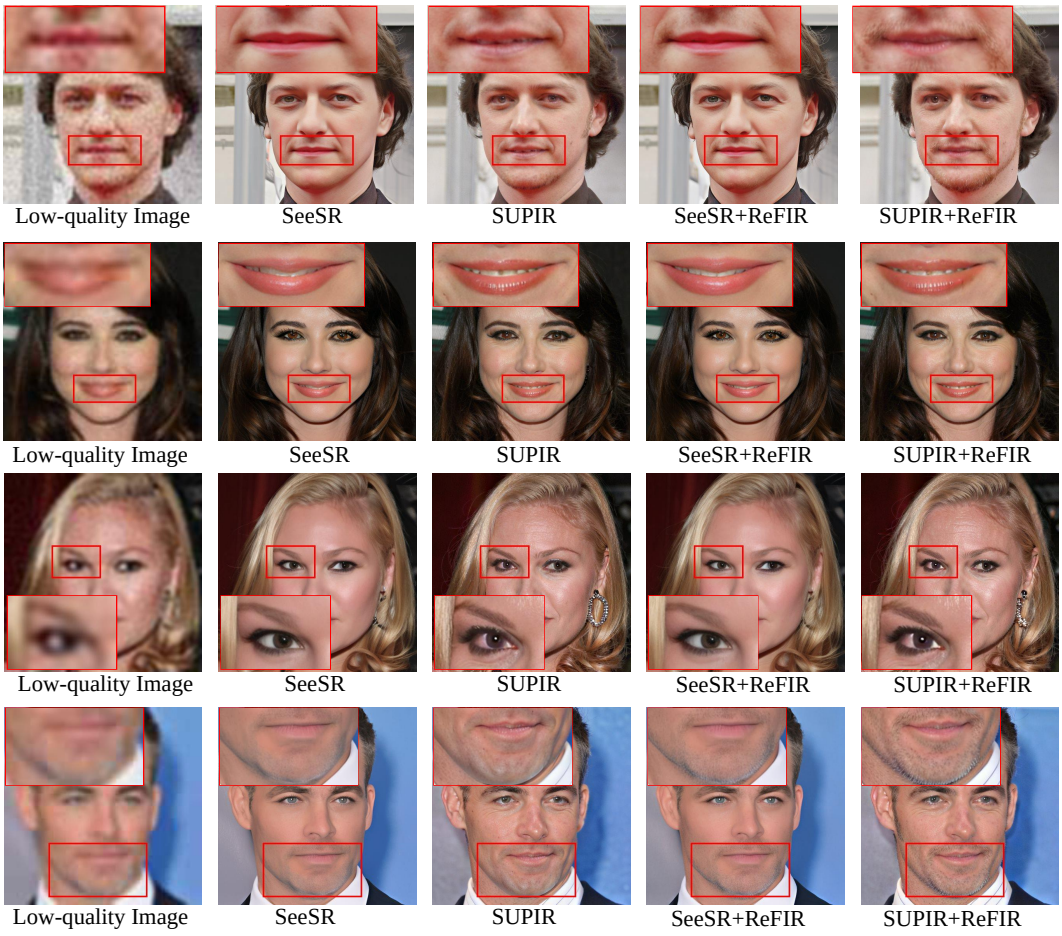


Figure 15: Visualization results of applying the proposed ReFIR to the downstream specific domain of blind face image restoration. Please zoom in for better effects.

https://doi.org/10.52202/079017-1478

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clearly outlined the main contributions of this paper in the abstract and introduction sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed in detail the limitation of this work in Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-1478

Justification: All assumptions in the paper are either derived from the conclusions of previous research or validated through extensive experiments.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided a detailed description of our algorithm's process and implementation specifics in the paper, and we will release our code after review.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release our code after review, but we have already provided a detailed explanation of how to implement our algorithm and the specific implementation details in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided detailed experimental details in both the paper and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have provided the performance fluctuation under different random seeds in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-1478

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the computational resources we used for experiments in the **??**.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have checked in every respect that this work conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have provided detailed discussion on possible both positive and negative societal impact in Appendix G.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, the creators and original owners of assets used in this paper are properly credited, and the license and terms of use are explicitly mentioned and respected, as evidenced by thorough citations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/079017-1478

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1478