Reference-based Super-Resolution via Image-based Retrieval-Augmented Generation Diffusion

Byeonghun Lee[1]* Hyunmin Cho[1]* Hong Gyu Choi[2]
Soo Min Kang[2] Iljun Ahn[2] Kyong Hwan Jin[1]†
[1]Korea University [2]Independent Researcher
{byeonghun_lee, hyun_cho, kyong_jin}@korea.ac.kr, {hgc950909, kang.soomin, ijahn0828}@gmail.com

Abstract

Most existing diffusion models have primarily utilized reference images for image-to-image translation rather than for super-resolution (SR). In SR-specific tasks, diffusion methods rely solely on low-resolution (LR) inputs, limiting their ability to leverage reference information. Prior reference-based diffusion SR methods have shown that incorporating appropriate references can significantly enhance reconstruction quality; however, identifying suitable references in real-world scenarios remains a critical challenge. Recently, Retrieval-Augmented Generation (RAG) has emerged as an effective framework that integrates retrieval-based and generation-based information from databases to enhance the accuracy and relevance of responses. Inspired by RAG, we propose an image-based RAG framework (iRAG) for realistic super-resolution, which employs a trainable hashing function to retrieve either real-world or generated references given an LR query. Retrieved patches are passed to a restoration module that generates high-fidelity super-resolved features, and a hallucination filtering mechanism is used to refine generated references from pre-trained diffusion models. Experimental results demonstrate that our approach not only resolves practical difficulties in reference selection but also delivers superior performance over existing diffusion and non-diffusion RefSR methods. Code is available at https://github.com/ByeonghunLee12/iRAG.

1. Introduction

Single image super-resolution (SISR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) input, presenting a long-standing challenge in computer vision. Recent deep learning methods [6, 10, 17, 32, 40, 62] have significantly improved the performance of SISR, but often result in over-smoothed details when optimized only for fidelity metrics (e.g., PSNR, SSIM [61]). To address this limitation, leveraging adversarial training emphasizes perceptually plausible high-frequency details, often leveraging adversarial training [4, 59, 67]. However, GAN-based models can be unstable during training and can introduce unnatural artifacts [34], spurring interest in alternative generative formulations such as diffusion probabilistic models [22, 46].

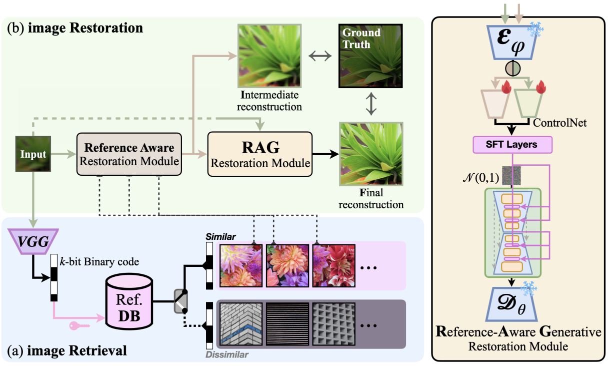


Figure 1. **Concept of image-based RAG (iRAG).** Our framework comprises two key components: image restoration and data retrieval. The image restoration process is facilitated by two modules—the Reference-Aware Restoration Module and the RAG Restoration Module. The data retrieval component manages internal communication through hash keys, enabling efficient and rapid retrieval. Additionally, the database is constructed using a subset of real datasets combined with synthetically generated data produced by a pre-trained generative model.

Diffusion probabilistic models have gained prominence by capturing complex, multimodal data distributions without the mode collapse problems often associated with GANs [20, 51]. By modeling a forward noise-adding and reverse-denoising process, these models excel in high-quality image generation [13, 23] and have been extended to inpainting, colorization, and super-resolution [47, 52]. Although operating in pixel space achieves state-of-the-art fidelity, it demands substantial computational resources in both training and inference [13]. To address these con-

*Equal Contribution.
†Corresponding Author.

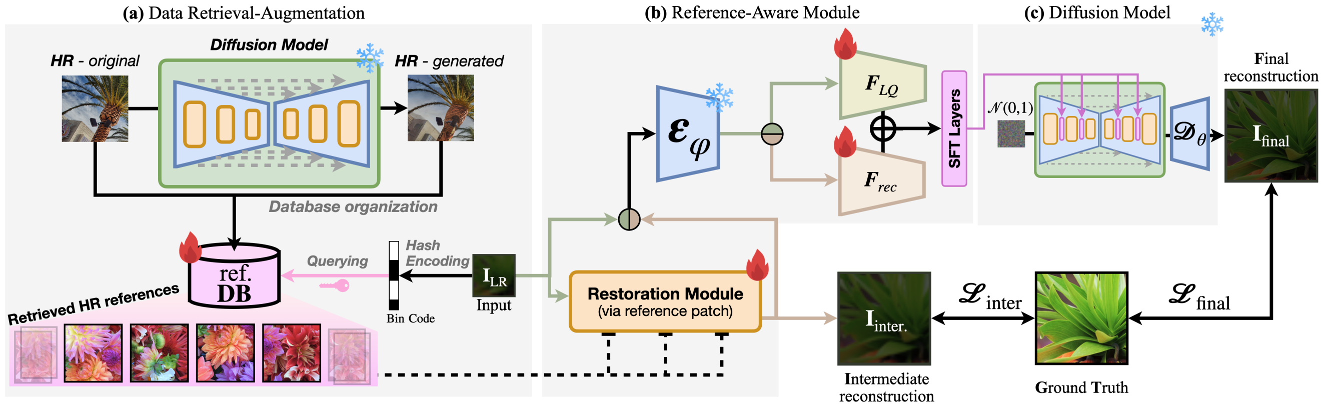


Figure 2. **Flowchart of image-based RAG.** **(a)** A database of high-resolution references is hashed for efficient querying, and retrieved references are used to generate additional training samples via a diffusion model. **(b)** The low-resolution input is fused with the retrieved reference patches through a dedicated restoration module, leveraging spatial feature transform (SFT) layers. **(c)** The final stage refines the intermediate reconstruction to produce high-fidelity outputs.

straints, latent diffusion models (LDMs) [30] project images into a lower-dimensional latent space, enabling efficient training and sampling while retaining high-quality reconstructions. This approach democratizes large-scale diffusion training and broadens its applicability to various conditional image-to-image transformations [38, 68].

In parallel, reference-based super-resolution (RefSR) utilizes HR auxiliary images with semantically or texturally similar regions to guide the reconstruction of the LR input. Early RefSR methods relied on optical flow [71] or patch similarity [48, 70] to align and transfer relevant textures from the reference image. Contemporary techniques employ attention mechanisms [65] or contrastive learning [24, 63] to handle complex alignments and improve texture fidelity. While existing RefSR methods *assign* HR auxiliary images that are semantically or structurally similar to LR input images, they face two key challenges: (1) matching difficulty, especially when the reference image differs substantially from the input LR image in illumination or pose [24, 41]; and (2) robust texture transfer, to ensure that only relevant high-frequency details are mapped to the LR input while minimizing artifacts from mismatched regions.

Recently, the Retrieval-Augmented Generation (RAG) framework has been widely used to enhance the quality of retrieved responses in NLP tasks [19, 33]. In the text-to-image generation domain, Re-Imagen [9] leverages cross-attention with retrieved images from an image database to improve semantic fidelity. We observe the analogy between retrieval-augmented systems and reference-based SR in terms of using additional retrieved patches for their own tasks.

In this work, we propose a novel reference-based super-resolution diffusion model follows RAG that unites the rich generative priors of latent diffusion [30] with the targeted detail enhancement strategy of RefSR [24, 41] which follows RAG [33] dedicated to the image domain such as [9]. Our approach tackles the reference selection and matching challenges by:

- We introduce a hashing code vector strategy to efficiently retrieve relevant references within a compact latent space, making it robust against illumination and style variations between LR and HR images.
- We integrate additional modules within our diffusion-based super-resolution framework to refine domain alignment between the LR input and the selected reference, thereby efficiently leveraging reference information.
- Extensive experiments show that our diffusion-based RefSR framework outperforms previous methods and provides a robust real-world RAG-based solution even when exact references are absent.

2. Related Works

Diffusion Probabilistic Models Diffusion probabilistic models have emerged as powerful generative frameworks for high-fidelity image synthesis. Earlier generative models such as variational autoencoders and flow-based models [14, 15, 28] focused on likelihood-based training and efficient sampling, yet they often fell short of the visual quality achieved by Generative Adversarial Networks (GANs) [20]. Autoregressive models [7, 11] offered strong density estimation, but were hampered by slow sampling.

Diffusion models [51] have since demonstrated excellent sample quality [13] and robust density estimation [27]. However, directly operating in pixel space is computationally expensive. Latent diffusion models (LDMs) [30] address this challenge by projecting images into a compressed latent space, reducing computational demands while preserving the quality of the synthesis. This approach supports a wide range of conditional and unconditional image generation tasks, including applications like super-resolution.

Realistic Image Super-Resolution Realistic image super-resolution (Real-ISR) targets high-fidelity, perceptually convincing, and artifact-free outputs in real-world settings. Conventional methods [16, 36] optimizing fidelity metrics (e.g., PSNR, SSIM) tend to produce over-smoothed details, whereas adversarial approaches [20, 31, 58] enhance texture sharpness via discriminative training, albeit with potential instability and artifacts [34]. Recently, diffusion models [22, 46] have been employed to capture the natural image distribution more effectively, resulting in finer textures and improved training stability. By integrating diffusion priors with tailored degradation models and objectives, these methods demonstrate superior restoration performance under challenging conditions.

Reference-Based Image Super-Resolution and RAG Reference-based image super-resolution (RefSR) improves reconstruction quality by transferring high-frequency details from an auxiliary high-resolution (HR) reference image to the low-resolution (LR) input. Traditional methods align the reference to the LR image using optical flow [71] or patch matching [48, 70], while advanced techniques employ attention mechanisms, contrastive learning, or teacher-student distillation for more robust texture alignment [24, 63, 65]. Recently, Retrieval-Augmented Generation (RAG) [19, 33] has emerged as a promising paradigm that dynamically sources and integrates external references, addressing the challenges of precise matching in real-world scenarios and further enriching the reconstruction process. RAG has been applied across multiple task domains: text-to-image generation [9], vision–language models [21], autonomous driving [60], etc.

Data Retrieval via Neural Hash Network Neural hashing has advanced data retrieval by learning compact binary codes that capture both semantic and structural information. Supervised methods [18, 43, 64, 66] leverage label information to ensure similar images yield similar codes, preserving fine details for various vision tasks. Meanwhile, unsupervised approaches [37, 44] derive representations directly from the data distribution, revealing intrinsic patterns.

Hallucination in Diffusion Models Diffusion models are well-known for generating high-fidelity images; however, they can also produce anomalous outputs—often termed *hallucinations*—such as images where hands exhibit extra fingers. These irregularities not only compromise image quality but may also undermine model robustness. Prior works [49, 50] have demonstrated that training on recursively generated data can erode critical, rare features, potentially leading to model collapse. Other works [2, 26] have examined the origins of these hallucinations and proposed various strategies for mitigation.

3. Methodology

In the context of image super-resolution (SR), an auxiliary high-resolution (HR) image, containing semantically or texturally similar information to the input low-resolution (LR) image, is often leveraged to guide the restoration of fine details and structural integrity. However, obtaining such a reference image for real-world datasets is challenging. Large datasets, such as ImageNet [12], typically consist of single-view images, making it difficult for direct use as reference patches. Furthermore, the process of curating a reference image from these extensive datasets is computationally onerous. To address this challenge, we propose a reference-based SR method that involves three key steps: (i) *augmenting* the existing dataset by enriching it with auxiliary HR images, (ii) *retrieving* a relevant HR image from a large database to match the target LR image, and (iii) *generating* a high-quality HR image by integrating LR and reference features into a diffusion model, as illustrated in Fig. 2.

3.1. Data Augmentation

The collection and annotation of real-world data can incur significant expenses, pose privacy concerns, and be subject to bias, potentially compromising the diversity of auxiliary data useful for LR image super-resolution [3, 45]. To address these challenges, we employ a diffusion model [46, 54] for data augmentation, thereby expanding the available dataset. Specifically, we sample latent noise σ from $\mathcal{N}(0, \mathbf{1})$ and use a pre-trained diffusion editing model, $\mathcal{G}$ [42], to generate a synthesized image $\mathbf{I}_{\text{gen}}$ conditioned on a real image $\mathbf{I}_{\text{HR}}$ as follows:

$$\mathbf{I}_{\text{gen}} = \mathcal{G}\big(\tilde{\mathbf{z}}_{\text{HR}}\big), \quad \tilde{\mathbf{z}}_{\text{HR}} \triangleq \mathbf{z}_{\text{HR}} + \alpha \cdot \sigma, \tag{1}$$

where $\tilde{\mathbf{z}}_{\text{HR}}$ is noisy latent obtained by adding noise σ to $\mathbf{z}_{\text{HR}}$, an encoded latent of $\mathbf{I}_{\text{HR}}$.

Recent analyses indicate that generative models are prone to producing out-of-distribution artifacts, commonly referred to as "hallucinations." These hallucinations can degrade downstream performance if used naively as training data [49, 50]. To mitigate these risks, we have incorporated a variance-based filtering method [2] to remove extreme hallucinations from the $\mathbf{I}_{\text{gen}}$. This method is based on the assumption that the high variance of a sampling trajectory can lead to the generation of hallucinated samples. Suppose $\{\hat{x}_0^{(i)}\}_{i=T_1}^{T_2}$ denotes the sequence of predicted images at timestep i during the reverse diffusion process, and $\bar{\hat{x}}_0$ denotes their mean. We leverage the hallucination metric proposed in [2]:

$$\text{Hal}(x) = \frac{1}{|T_2 - T_1|} \sum_{i=T_1}^{T_2} (\hat{x}_0^{(i)} - \bar{\hat{x}}_0)^2. \tag{2}$$

Training Loss		Pixel Loss		GAN Loss			Diffusion Loss		
Using References		✓	✓	✗	✗	✗	✗	✗	✓
Benchmarks	Metrics	TTSR [65]	DATSR [5]	Real-ESRGAN [59]	BSRGAN [67]	DASR [35]	LDM [46]	StableSR [56]	**Ours**
DIV2K Valid [1]	PSNR ↑	27.560	27.843	23.499	23.861	24.746	22.629	22.044	22.323
	SSIM ↑	0.783	0.788	0.686	0.683	0.707	0.661	0.626	0.632
	LPIPS ↓	0.264	0.257	0.239	0.253	0.172	0.228	0.248	0.228
	CLIP-IQA ↑	0.451	0.452	0.534	0.580	0.612	0.618	0.672	0.702
	MUSIQ ↑	58.702	59.036	63.387	64.817	65.340	66.910	67.066	68.248
RealSR [4]	PSNR↑	31.144	31.546	25.347	26.349	28.021	24.228	23.702	23.951
	SSIM↑	0.902	0.906	0.789	0.799	0.838	0.773	0.721	0.717
	LPIPS↓	0.176	0.129	0.194	0.199	0.140	0.223	0.262	0.263
	CLIP-IQA↑	0.458	0.454	0.526	0.574	0.554	0.657	0.664	0.678
	MUSIQ↑	64.292	63.493	67.536	69.660	67.964	69.902	70.196	70.524
DRealSR	PSNR↑	36.426	37.090	29.893	30.418	34.532	24.228	27.049	27.144
	SSIM↑	0.949	0.952	0.865	0.865	0.905	0.773	0.765	0.776
	LPIPS↓	0.110	0.105	0.199	0.221	0.116	0.223	0.305	0.300
	CLIP-IQA↑	0.425	0.422	0.516	0.573	0.455	0.657	0.687	0.682
	MUSIQ↑	49.237	47.575	60.056	63.027	52.984	69.902	64.338	64.861

Table 1. **Quantitative comparison on test benchmarks.** The best and second results are in red and orange . All the models on the table are trained with Flickr2K , DIV2K-train and OST.

This metric ensures that latent codes exhibiting unusually high variance are flagged as hallucinations and excluded from the augmentation set. As illustrated in Sec. 4.2, the method successfully filters out hallucinated samples (i.e., unrealistic data), enabling robust data augmentation.

3.2. Hashing for Data Retrieval

To search for the HR exemplar relevant to the input LR image from an external database, our pipeline incorporates a *retrieval* mechanism [44]. Specifically, for each input $\mathbf{I}_{\text{LR}}$, we retrieve $\mathbf{I}_{\text{ref}}$ from ref.DB, the database of reference images, that best matches the content or style of $\mathbf{I}_{\text{LR}}$ under a compact hashing scheme. To achieve this, we first project the image into an embedding space, where image $\mathbf{I}$ is mapped to vector $\mathbf{h} = H_\phi(\mathbf{I})$. We then retrieve the reference image $\mathbf{I}_{\text{ref}}$ from ref.DB whose hash code is closest to the encoded query image's hash code $\mathbf{h}_{\text{LR}} = H_\phi\big(\mathbf{I}_{\text{LR}}\big)$, as follows:

$$\mathbf{I}_{\text{ref}} = \arg\min_{\mathbf{I}_{\text{DB}} \in \text{ref.DB}} \left(\text{dist}(\mathbf{h}_{\text{LR}}, H_\phi(\mathbf{I}_{\text{DB}}))\right).$$

To facilitate contrastive learning, we construct positive pairs by generating correlated views for each image. Following standard convention, we apply arbitrary transformations (e.g., rotation, reflection, etc.) to each image, thereby producing two correlated views for the k-th image $x^{(k)}$ in the database:

$$v_i^{(k)} := \mathcal{T}(x^{(k)}), \quad i \in \{1, 2\},$$

where $\mathcal{T}$ denotes an arbitrary transform operator. Next, these views are passed through a pre-trained VGG encoder network $f(\cdot)$ to obtain the corresponding latent representations:

$$z_i^{(k)} = f(v_i^{(k)}), \quad h_i^{(k)} = g_\phi(z_i^{(k)}),$$

where $z_i^{(k)}$ represents the encoded latent vector of the i-th view of the k-th image and $h_i^{(k)}$ denotes the corresponding binary hash code generated by the learnable function $g_\phi(\cdot)$. The primary objective of our learning process is to minimize the distance between the binary hash codes $h_1^{(k)}$ and $h_2^{(k)}$ derived from the same image. We quantify the similarity between binary hash codes [8] using a cosine similarity function $\mathcal{C}$. To amplify the differences between positive and negative pairs, we exponentiate the cosine similarity, yielding:

$$\mathcal{S}(h_1^{(k)}, h_2^{(k)}) := e^{\mathcal{C}(h_1^{(k)}, h_2^{(k)})}/\tau,$$

where τ is a normalization term. Next, we compute the log-likelihood of correctly identifying the positive match among all candidate pairs:

$$\ell_1^{(k)} := -\log \frac{\mathcal{S}(h_1^{(k)}, h_2^{(k)})}{\mathcal{S}(h_1^{(k)}, h_2^{(k)}) + \sum_{i, n \neq k} \mathcal{S}(h_1^{(k)}, h_i^{(n)})}.$$

The summation in the denominator aggregates the similarities between the anchor hash code $h_1^{(k)}$ and all negative samples. To symmetrize the objective, we also compute an analogous term $\ell_2^{(k)}$ for the second view. The final contrastive loss is defined as the average over all instances:

$$\mathcal{L}_{cl} := \frac{1}{N} \sum_{k=1}^{N} \left(\ell_1^{(k)} + \ell_2^{(k)}\right).$$

By minimizing $\mathcal{L}_{cl}$, the model learns an embedding space where semantically related samples cluster together.

3.3. Reference-based Latent Diffusion Model

To reconstruct the LR image using a retrieved reference, we proceed in two stages. First, we generate an intermediate reconstruction, $\mathbf{I}_{\text{inter}}$, using a transformer-based reference SR

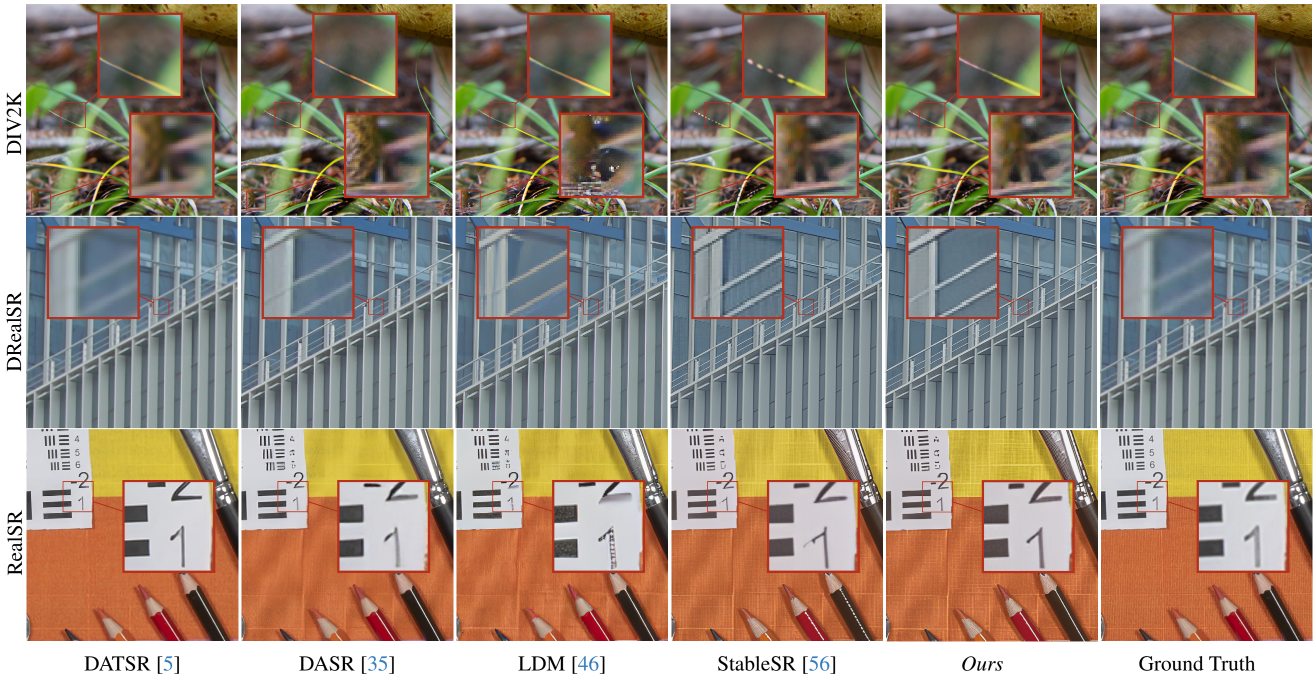


Figure 3. **Qualitative comparison on the test benchmarks with respect to our reference DB (Sec. 4.1).** We compare DATSR [5], DASR [35], LDM [46], StableSR [56], and our method. Zoomed-in regions highlight differences in detail preservation.

module [5, 65], which contains enriched features from the $\mathbf{I}_{\text{LR}}$. Then, we feed both the $\mathbf{I}_{\text{LR}}$ and $\mathbf{I}_{\text{inter}}$ into a diffusion-based refinement module to guide the final reconstruction.

3.3.1. Intermediate Reference-based SR

Given $\mathbf{I}_{\text{LR}}$ and its retrieved reference $\mathbf{I}_{\text{ref}}$, we employ a reference-based SR network, [5, 65] R_ϕ, to produce an intermediate super-resolved image:

$$\mathbf{I}_{\text{inter}} = R_\phi(\mathbf{I}_{\text{LR}}, \mathbf{I}_{\text{ref}}). \tag{3}$$

We minimize a combined loss that incorporates both L1 loss and a perceptual loss to ensure that the network captures the global structure and texture of the reference image. The resulting image, $\mathbf{I}_{\text{inter}}$, enriches the feature representation of $\mathbf{I}_{\text{LR}}$, ensuring that the diffusion model benefits from faithful structural and texture cues.

Next, we feed both $\mathbf{I}_{\text{LR}}$ and $\mathbf{I}_{\text{inter}}$ into a feature encoder ε_φ. Specifically, we adapt SFT layers [56] to condition the reverse diffusion process on the features extracted from $\mathbf{I}_{\text{LR}}$ and $\mathbf{I}_{\text{inter}}$. Formally, let:

$$\boldsymbol{F}_{\text{cond}} \triangleq \text{concat}\big(\varepsilon_\varphi(\mathbf{I}_{\text{LR}}), \varepsilon_\varphi(\mathbf{I}_{\text{inter}})\big), \tag{4}$$

be the condition features encoded from both images, which are injected into each diffusion block through the SFT layers. We integrate the pre-cleaned features into the diffusion process via SFT layers, which perform affine transformations on the diffusion model's intermediate features. Let $\boldsymbol{F}^n_{dif}$ be the feature map in the n-th residual block of the Stable Diffusion U-Net. The SFT layer computes two affine parameters, α^n and β^n, based on both the LR feature $\boldsymbol{F}^n_{LR}$ and the reference feature $\boldsymbol{F}^n_{inter}$. Formally, these parameters are computed as:

$$\alpha^n(\boldsymbol{F}^n_{LR}, \boldsymbol{F}^n_{inter}),\ \beta^n(\boldsymbol{F}^n_{LR}, \boldsymbol{F}^n_{inter}) = \mathcal{K}^n_\theta\big(\boldsymbol{F}^n_{LR}, \boldsymbol{F}^n_{inter}\big),$$

where $\mathcal{K}^n_\theta(\cdot)$ is a small network (e.g., a series of convolutional and activation layers) that learns to predict the affine parameters. The transformed diffusion feature map $\hat{\boldsymbol{F}}^n_{dif}$ is then given by:

$$\hat{\boldsymbol{F}}^n_{dif} = \alpha^n(\boldsymbol{F}^n_{LR}, \boldsymbol{F}^n_{inter}) \odot \boldsymbol{F}^n_{dif} + \beta^n(\boldsymbol{F}^n_{LR}, \boldsymbol{F}^n_{inter}),$$

where $\odot$ denotes element-wise multiplication. By injecting $\boldsymbol{F}^n_{inter}$ and $\boldsymbol{F}^n_{LR}$ into the diffusion process, the network can selectively emphasize or suppress certain features, effectively leveraging the intermediate features extracted from the reference image to guide super-resolution. During reverse diffusion, we iteratively denoise a latent $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ down to $\mathbf{x}_0$, guided by the condition features $\boldsymbol{F}_{\text{cond}}$. The final super-resolved output is:

$$\mathbf{I}_{\text{final}} = \mathcal{D}\big(\mathcal{U}(\mathbf{x}_T, \boldsymbol{F}_{\text{cond}})\big), \tag{5}$$

where $\mathcal{U}$ and $\mathcal{D}$ denote the pre-trained denoising U-Net and the latent decoder from [46], respectively.

3.3.2. Training Objectives

We supervise both the intermediate and final outputs using the following losses. Specifically, we define the loss for intermediate reconstruction as:

$$\mathcal{L}_{\text{inter}} = \|\mathbf{I}_{\text{inter}} - \mathbf{I}_{\text{HR}}\|_1 + \lambda \cdot \text{GAN}\left(\mathbf{I}_{\text{inter}}, \mathbf{I}_{\text{HR}}\right),$$

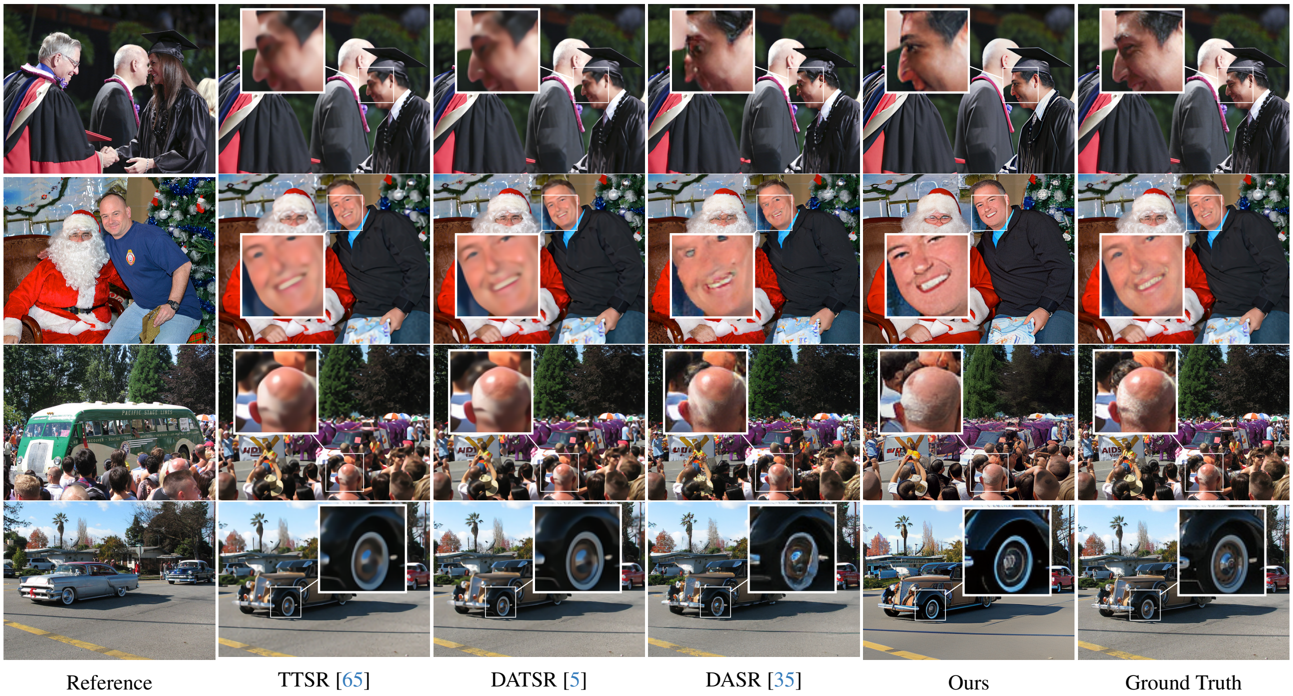


Figure 4. **Qualitative comparison on the CUFED5 dataset with respect to CUFED5 references.** We compare TTSR [65], DATSR [5], DASR [35] and our method against the ground truth.

Quant. Result: **Evaluation on *CUFED5* (i.e., Reference-Provided setting)**				
Metrics	**Models for Eval**			
	TTSR [65]	DATSR [5]	DASR [35]	*Ours*
PSNR↑	25.224	26.71	20.69	20.56
SSIM↑	0.782	0.838	0.649	0.686
LPIPS↓	0.222	0.152	0.183	0.219
CLIP-IQA↑	0.372	0.435	0.536	0.693
MUSIQ↑	66.329	69.149	68.355	73.573

Table 2. **Model performance comparison** on reference-provided settings to assess the intrinsic performance of the models. All models were trained on DF2K-OST.

and the final loss as:

$$\mathcal{L}_{\text{final}} = \|\mathbf{I}_{\text{final}} - \mathbf{I}_{\text{HR}}\|_2 \,. \tag{6}$$

The overall loss is a weighted sum of the intermediate and final losses:

$$\mathcal{L} = \mathcal{L}_{\text{inter}} + \alpha \cdot \mathcal{L}_{\text{final}}, \tag{7}$$

where α is a hyperparameter that balances the importance of intermediate reconstruction versus final refinement. By jointly optimizing these two stages, the network learns to leverage both reference-based SR cues and diffusion refinement to produce high-quality high-resolution outputs.

4. Experiments

4.1. Experiment and Evaluation Settings

We trained our model using the Adam optimizer [29]. All experiments were performed on NVIDIA RTX 3090 GPUs.

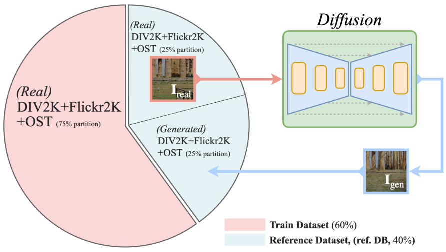


Figure 5. **Composition of the database**. We partition the real-domain dataset into a training set (75% of the data) and a reference set (25%). We then augment the reference set at a 1:1 ratio.

Reference Database (ref. DB) and Training Detail Fig. 5 illustrates the configuration of the dataset. We split the DF2K-OST dataset [1, 57] such that 75% of the images form the training set for image restoration. The remaining 25% is used to build the reference database, which is further augmented with synthetic images. To enrich this reference database, we add synthetic images in a one-to-one ratio with the real patches. Specifically, from each reference image, we extracted 512 × 512 patches (using reflective padding if needed) and trained an unsupervised hashing model on these patches using positive pairs generated via standard data augmentation [8]. Synthetic images were generated using the diffusers pipeline [54] and SDEdit [42], with guidance scales randomly chosen from $[7, 10]$ and noise scales from $[0.55, 0.65]$. A hallucination threshold (average variance of 0.03) was applied; if exceeded, the generation was repeated up to 10 times, retaining the sample with the lowest variance. All processes utilized SD2.1 [46].

Evaluation Settings We evaluated the performance of our model on three benchmark datasets: the DIV2K validation set [1], the RealSR dataset [4], and the DRealSR dataset. For each dataset, we computed quality metrics—including PSNR, LPIPS [69], SSIM [61], CLIP-IQA [55], and MUSIQ [25]—to assess performance. These metrics comprehensively assess both the pixel-level fidelity and the perceptual quality of the generated images.

Quantitative & Qualitative Results We validated our proposed approach using two distinct datasets: (i) a reference-provided dataset and (ii) a real-world image dataset as described in Fig. 5. As illustrated in Tab. 2, our method demonstrates superior perceptual quality in the matched dataset compared to other baselines. In particular, as shown in Fig. 4, fine details and realistic textures are preserved more effectively. Furthermore, as demonstrated in Fig. 3 and Tab. 1, our method consistently outperforms baseline models in real-world scenarios, thereby confirming its robustness and practical applicability.

Ablation Study In Tab. 3, we compare our full method (*Ours*) with three ablated variants on the RealSR dataset. In *Ours* (-H), we remove hashing-based reference matching during training and instead use random pairing. This leads to a noticeable drop in performance (e.g., PSNR from 23.957 to 23.614, SSIM from 0.715 to 0.698), underscoring the importance of accurate reference selection. In *Ours* (-RM), the reference restoration module is removed, causing further degradation in visual quality (e.g., PSNR decreases to 23.374, SSIM to 0.680), indicating that this module is essential for effectively leveraging reference patches. Finally, in *Ours* (-R), hashing-based reference retrieval is replaced with random patch selection during inference, which also impairs reconstruction quality (PSNR drops to 23.709, SSIM to 0.708). As shown in Tab. 4, varying the loss-balancing weight α indicates that $\alpha = 1.0$ strikes the best trade-off between distortion metrics and perceptual quality.

Quant. Result: **Evaluation on each modules**

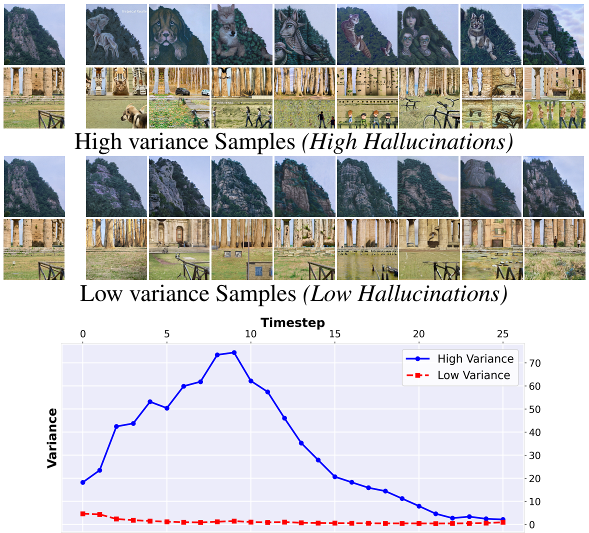
Ablation	PSNR↑	SSIM↑	LPIPS↓	CLIP-IQA↑	MUSIQ↑
Ours	23.957	0.715	0.264	0.677	70.521
Ours (-H)	23.614	0.698	0.302	0.639	66.146
Ours (-RM)	23.374	0.680	0.289	0.645	68.122
Ours (-R)	23.709	0.708	0.276	0.661	69.182

Table 3. **Ablation study** on different variants of the proposed method on the RealSR dataset.

Quant. Result: **Ablation study on weight α**

Ablation	PSNR↑	SSIM↑	LPIPS↓	CLIP-IQA↑	MUSIQ↑
Ours ($\alpha = 0.7$)	22.431	0.637	0.229	0.665	67.102
Ours ($\alpha = 1.0$)	22.323	0.632	0.228	0.702	68.248
Ours ($\alpha = 1.3$)	22.007	0.617	0.239	0.716	67.913

Table 4. **Impact of the loss-balancing weight α** (Eq. (7)) on distortion and perceptual quality.


High variance Samples *(High Hallucinations)*

Low variance Samples *(Low Hallucinations)*

Figure 6. **Effectiveness of variance (Hallucination)** in diffusion trajectories during sampling [39]. Our empirical results indicate that the effect shown in [2] is also observed in the latent domain. Furthermore, the high variance sampling process even shows the oscillation in trajectory.

4.2. Robust Data Augmentation via Generation

We apply rejection sampling, as described in Eq. (2), to generate robust image samples. As illustrated in Fig. 6, higher variance in the sampling trajectory tends to yield unrealistic images (e.g., faces appearing on mountains), whereas lower variance produces images that preserve the core realistic attributes while incorporating desired edits. Furthermore, samples generated with high variance tend to exhibit persistently high variance values and lack a decaying trend during the initial timesteps. In contrast, samples generated with low variance exhibit an almost monotonic decrease in the trajectory, indicating smooth sampling in the vector field that likely guides the sample toward the proper mode.

4.3. Augmented Data Improves Hashing

We evaluate our approach by measuring the hit rate of the hash function, a key retrieval indicator. For each real image, augmented samples are generated as in Fig. 5; each CUFED query is paired with four references ranked by similarity. As shown in Tab. 5, generated-only training underperforms real-only training, but combining the two (*Real*+*Gen*) yields the highest hit rates and retrieves the correct patches in proper order (Fig. 7). These results show that training with diverse patches strengthens invariant representations and overall retrieval performance. Crucially, even when we restrict the real set to just 10% of its original size, augmenting it with a nine-fold amount of generated patches (*Real*$_{lim}$+*Gen*) almost restores—and in some ranks surpasses—the full-data baseline, highlighting generation's value in data-scarce scenarios.

Quant. Result: **Hit rate (%) of correct CUFED ref. patch retrieval**						
Train Configuration		Rank of the ref. patch Similarity				
Train. DB	#samples	R1	R2	R3	R4	R5
Real	13K	45	59	64	70	75
Gen	13K	40	52	59	63	66
Real+Gen	26K	52	60	66	71	73
Real_{lim}	1K	48	55	60	63	67
Real_{lim} + Gen	10K	**52**	**62**	**63**	**67**	**71**

Table 5. **Effectiveness of generated samples.** Evaluation is conducted on the CUFED dataset to compare the hit rate for models trained on Real vs. Gen patches, as well as a combination of both.

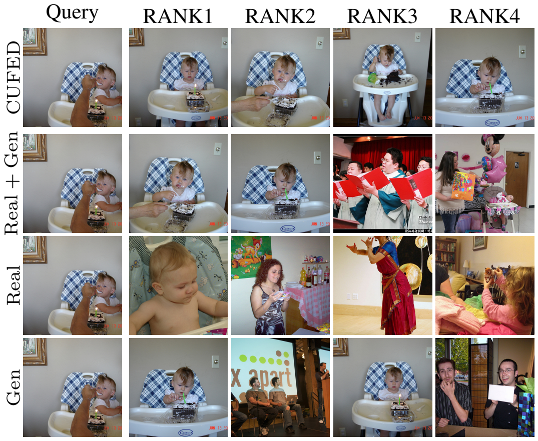


Figure 7. **Effectiveness of generated samples.** Hash trained with both (Gen+Real) dataset retrieves the most similar patches.

4.4. Hashing Time and Resource Usage

Tab. 6 shows that our hashing-based retrieval approach reduces both storage memory and matching cost compared to VGG-based feature matching. We assume 26,500 reference images (3×512×512), where each reference feature or hash code is precomputed, and 80,000 query images are matched in batches of 64. By compressing the code length from 4096 to just 16, we shrink the memory footprint from more than 800 MB to less than 52 KB while also cutting the retrieval time from 4338.2 ms to 67.8 ms.

Quant. Result: **Computational Cost**			
Algorithm	Evaluation Metric		
	Code Length	Storage Memory	Cost
VGG Matching	4096	828.13 M	4338.2 ms
Hashing	16	51.76 KB	67.8 ms

Table 6. **Effectiveness of Hash** on computational cost. Comparison conducted with VGG-based matching and our hashing method. We assume 26,500 reference images with precomputed codes, and 80,000 query images matched.

4.5. Integrating Hashing into Existing Ref-SR

As described in Tab. 8 and Fig. 8, integrating hashing into conventional Ref-SR models [65] efficiently retrieves structurally correlated regions from references, enabling more precise alignment and fusion for higher-fidelity reconstructions. In diffusion-based Ref-SR, CoSeR [53] first generates a reference from the LR input; replacing or augmenting it with our hash-selected references yields consistent gains in PSNR/SSIM and no-reference IQA (CLIP-IQA, MUSIQ), as shown in Tab. 7.

Quant. Result: **Reference adaptation in CoSeR**					
Ref. Type	PSNR↑	SSIM↑	LPIPS↓	CLIP-IQA↑	MUSIQ↑
Gen	20.23	0.517	0.405	0.562	56.419
Hash	20.30	0.519	0.408	0.566	57.686

Table 7. **Impact of hash-retrieved references** on diffusion-based RefSR [53]; metrics compare generated (Gen) and hash-selected (Hash) reference strategies.

Quant. Result: **PSNR/SSIM**			
model: **TTSR**		**Ref. at Eval**	
Dataset	**Ref. at Train**	Random	Hash
DIV2K	Random	28.26 / 0.771	28.25 / 0.770
	Hash	29.08 / 0.793	29.16 / 0.794
CUFED	Random	24.74 / 0.736	24.73 / 0.736
	Hash	25.34 / 0.748	25.65 / 0.762

Table 8. **Effectiveness of Hash** for conventional RefSR [65]. Evaluation was conducted using different reference-selection strategies on DIV2K and CUFED. Each cell shows the average PSNR/SSIM under two factors: reference selection (*Random* vs *Hash*) at training and evaluation.

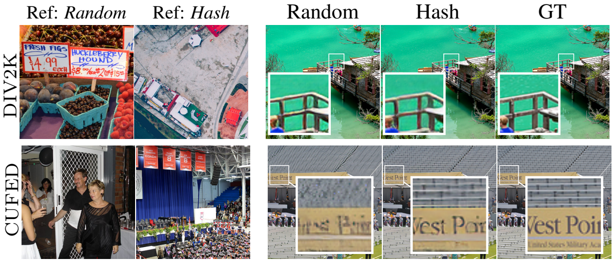


Figure 8. **Effectiveness of Hash**. The two groups on the left compare reference patches selected randomly or those retrieved using a hashing. Hashing mechanism enables the RefSR model to perform effectively even when an exact reference patch is unavailable.

5. Conclusion

We proposed a novel image-based Retrieval Augmented Generation framework that combines latent diffusion models with an efficient hashing code vector strategy achieving robust reference matching and realistic reference-based SR. Operating in a compact latent space by short binary hash codes, our method addressed the challenges of reference selection and improves domain consistency between low-resolution inputs and high-resolution references. Experiments on real-world datasets demonstrate that our approach outperforms existing diffusion-based super-resolution methods and reference-based methods in terms of fidelity, perceptual quality, and computational efficiency.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 4, 6, 7

[2] Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in Neural Information Processing Systems*, 37:134614–134644, 2025. 3, 7

[3] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[4] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3086–3095, 2019. 1, 4, 7

[5] 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 *European conference on computer vision*, pages 325–342. Springer, 2022. 4, 5, 6

[6] Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18257–18267, 2023. 1

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 4, 6

[9] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2, 3

[10] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 1

[11] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[14] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2

[16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[17] Huiyuan Fu, Fei Peng, Xianwei Li, Yejun Li, Xin Wang, and Huadong Ma. Continuous optical zooming: A benchmark for arbitrary-scale image super-resolution in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3035–3044, 2024. 1

[18] Hua Gao, ChenChen Hu, Guang Han, Jiafa Mao, Wei Huang, and Kaiyuan Wan. Hashneck is a boosting tool for deep learning to hashing. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 83–91, 2024. 3

[19] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023. 2, 3

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2, 3

[21] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021. 3

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[23] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 1

[24] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2021. 2, 3

[25] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 7

[26] Seunghoi Kim, Chen Jin, Tom Diethe, Matteo Figini, Henry FJ Tregidgo, Asher Mullokandov, Philip Teare, and Daniel C Alexander. Tackling structural hallucination in image translation with local diffusion. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 3

[27] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[28] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[30] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2

[31] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3

[32] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1929–1938, 2022. 1

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 2, 3

[34] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 1, 3

[35] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022. 4, 5, 6

[36] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 3

[37] Kevin Lin, Jiwen Lu, Chu-Song Chen, and Jie Zhou. Learning compact binary descriptors with unsupervised deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1183–1192, 2016. 3

[38] Hanyuan Liu, Jinbo Xing, Minshan Xie, Chengze Li, and Tien-Tsin Wong. Improved diffusion-based image colorization via piggybacked models. *arXiv preprint arXiv:2304.11105*, 2023. 2

[39] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 7

[40] Xiaoyi Liu and Hao Tang. Difffno: Diffusion fourier neural operator. *arXiv preprint arXiv:2411.09911*, 2024. 1

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2021. 2

[42] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 6

[43] Xiushan Nie, Xingbo Liu, Jie Guo, Letian Wang, and Yilong Yin. Supervised discrete multiple-length hashing for image retrieval. *IEEE Transactions on Big Data*, 9(1):312–327, 2022. 3

[44] Zexuan Qiu, Qinliang Su, Zijing Ou, Jianxing Yu, and Changyou Chen. Unsupervised hashing with contrastive information bottleneck. *arXiv preprint arXiv:2105.06138*, 2021. 3, 4

[45] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 4, 5, 6

[47] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 1

[48] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8425–8434, 2020. 2, 3

[49] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023. 3

[50] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631 (8022):755–759, 2024. 3

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1, 2

[52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[53] Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. *arXiv preprint arXiv:2311.16512*, 2023. 8

[54] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 3, 6

[55] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 7

[56] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024. 4, 5

[57] Xintao Wang, Ke Yu, Chao Dong, and Chen Change 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*, pages 606–615, 2018. 6

[58] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[59] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1, 4

[60] Yujin Wang, Quanfeng Liu, Jiaqi Fan, Jinlong Hong, Hongqing Chu, Mengjian Tian, Bingzhao Gao, and Hong Chen. RAC3: Retrieval-augmented corner case comprehension for autonomous driving with vision-language models. *arXiv preprint arXiv:2412.11050*, 2024. 3

[61] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 7

[62] Min Wei and Xuesong Zhang. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18247–18256, 2023. 1

[63] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *European Conference on Computer Vision*, pages 230–245. Springer, 2020. 2, 3

[64] Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1445–1451, 2020. 3

[65] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 2, 3, 4, 5, 6, 8

[66] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3083–3092, 2020. 3

[67] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 4

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[70] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7982–7991, 2019. 2, 3

[71] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 88–104, 2018. 2, 3