Diffusion Models, Image Super-Resolution And Everything: A Survey

Brian B. Moser[1,2], Arundhati S. Shanbhag[1,2], Federico Raue[1], Stanislav Frolov[1,2], Sebastian Palacio[1], Andreas Dengel[1,2]
[1] German Research Center for Artificial Intelligence (DFKI), Germany
[2] Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Germany
first.second@dfki.de

Abstract—Diffusion Models (DMs) represent a significant advancement in image Super-Resolution (SR), aligning technical image quality more closely with human preferences and expanding SR applications. DMs address critical limitations of previous methods, enhancing overall realism and details in SR images. However, DMs suffer from color-shifting issues, and their high computational costs call for efficient sampling alternatives, underscoring the challenge of balancing computational efficiency and image quality. This survey gives an overview of DMs applied to image SR and offers a detailed analysis that underscores the unique characteristics and methodologies within this domain, distinct from broader existing reviews in the field. It presents a unified view of DM fundamentals and explores research directions, including alternative input domains, conditioning strategies, guidance, corruption spaces, and zero-shot methods. This survey provides insights into the evolution of image SR with DMs, addressing current trends, challenges, and future directions in this rapidly evolving field.

Index Terms—Diffusion Models, Image Super-Resolution, Artificial Intelligence, Survey.

1 Introduction

In the ever-evolving field of computer vision, Super-Resolution (SR) – the enhancement of Low-Resolution (LR) images into High-Resolution (HR) ones – stands as a longstanding yet still perplexing challenge, owing to its inherently ill-posed nature: Several HR images can be valid for any given LR image due to many aspects like brightness and coloring [1]. Spanning applications from natural [2], [3] to sophisticated satellite [4] and medical imaging [5], SR's progress has been fueled by rapid strides in deep learning.

Diffusion Models (DMs) have recently risen to prominence as a leading family of generative models, challenging the longstanding dominance of Generative Adversarial Networks (GANs) [6], [7], [8], [9]. While earlier generative models showcased impressive image-generation capabilities, they carried inherent limitations. For instance, Autoregressive models can be prohibitively costly for generating HR images [10], [11], [12]. On the other hand, NFs and VAEs often produce samples of sub-optimal quality. Furthermore, GANs demand carefully designed regularization and optimization strategies to manage optimization instability and prevent mode collapse [13]. Nevertheless, these models collectively contributed significantly to the field of SR.

The advent of DMs heralded a new era of generative models and profoundly impacted the realm of generative AI, reinvigorating the field of image SR once again. However, as the volume of research on DMs continues to expand, it is becoming more challenging, particularly for those new to the field, to stay updated on the latest developments. The sheer breadth of this research area can obscure overarching trends and potentially hinder the progression of further research. We tackle these issues by delivering a comprehensive overview of the current research on DMs in image SR.

This survey builds upon our previous work *Hitchhiker's Guide to Super-Resolution* [14], which gives a broad overview of the field of image SR in general. We aim to provide a valuable entry point for those new to the domain of DMs in image SR while offering a broader perspective to those already experienced.

The structure of this paper is organized as follows:
Section 2 - Super-Resolution Basics: This section provides fundamental definitions and introduces standard datasets, methods, and metrics for assessing image quality commonly utilized in image SR publications.
Section 3 - Diffusion Models Basics: Introduces the principles and various formulations of DMs, including Denoising Diffusion Probabilistic Models (DDPMs), Score-based Generative Models (SGMs), and Stochastic Differential Equations (SDEs). This section also explores how DMs relate to other generative models.
Section 4 - Improvements for Diffusion Models: Common practices for enhancing DMs, focusing on efficient sampling techniques and improved likelihood estimation.
Section 5 - Diffusion Models for Image SR: Presents concrete realizations of DMs in SR, explores alternative domains (latent space and wavelet domain), discusses architectural designs and multiple tasks with Null-Space Models, and examines alternative corruption spaces.
Section 6 - Domain-Specific Applications: DM-based SR applications, namely medical imaging, blind face restoration, atmospheric turbulence in face SR, and remote sensing.
Section 7 - Discussion and Future Work: Common problems of DMs for image SR and noteworthy research avenues for DMs specific to image SR.
Section 8 - Conclusion: Summarizes the survey.

2 SUPER-RESOLUTION BASICS

Image Super-Resolution (SR) represents a collection of methodologies designed to construct High-Resolution (HR) images from one or more Low-Resolution (LR) images. The domain of SR can be broadly categorized into two areas [14]: Single Image Super-Resolution (SISR) and Multi-Image Super-Resolution (MISR). In the SISR paradigm, a single LR image is processed to yield a corresponding HR image. In contrast, MISR methods utilize multiple LR images to synthesize one or many HR outputs. This section primarily focuses on the SISR facet of SR for the context of DMs, exploring relevant datasets, established SR models, and techniques employed to assess image quality.

2.1 Single Image Super-resolution (SISR)

Single Image Super-Resolution (SISR) defines the recovery of a High-Resolution (HR) image from a Low-Resolution (LR) image. Given a LR image $\mathbf{x} \in \mathbb{R}^{\bar{w} \times \bar{h} \times c}$, the objective is to produce a corresponding HR image $\mathbf{y} \in \mathbb{R}^{w \times h \times c}$, where $\bar{w} < w$ and $\bar{h} < h$. This process is often represented through a degradation mapping, defined as

$$\mathbf{x} = \mathcal{D}(\mathbf{y}; \Theta) \tag{1}$$

where $\mathcal{D}$ is the degradation map $\mathcal{D} : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}^{\bar{w} \times \bar{h} \times c}$ and Θ signifies the degradation parameters, encompassing aspects such as blur type, noise, scaling factor, among others. The degradation process is typically uncertain, posing the challenge for SISR to discern the inverse mapping of $\mathcal{D}$. The inherent complexity of this task arises from the ill-posed nature of predicting Θ, leading to an optimization task aimed at minimizing the discrepancy between the predicted SR image $\hat{\mathbf{y}}$ and the original HR image $\mathbf{y}$:

$$\hat{\Theta} = \text{argmin}_{\Theta} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda \phi(\Theta), \tag{2}$$

where $\mathcal{L}$ represents the loss between the predicted SR image and the actual HR image. Here, λ is a balancing parameter, while $\phi(x)$ is introduced as a regularization term.

2.2 Datasets

Several datasets, offering a variety of images, resolutions, and content types, are available for SISR and instrumental in effectively training and evaluating SR models. Typically, these datasets consist of LR and HR image pairs. However, certain datasets contain only HR images, with the pairs created by downsampling them using techniques such as bicubic downsampling with anti-aliasing - a default setting for *imresize* in MATLAB [15].

For natural images, commonly employed datasets include the Berkeley Segmentation Dataset (BSDS300) [3], Flickr2K [16], and the Diverse 2K resolution (DIV2K) dataset [17]. Additionally, datasets such as ImageNet [18] and the Visual Object Classes (VOC2012) [19], which are favored in the Computer Vision (CV) community, are also utilized for SISR tasks. In face SR, datasets featuring high-quality, diverse images of human faces are required. To meet this demand, datasets like CelebA-HQ [20] and Flickr-Faces-HQ (FFHQ) [21] are commonly employed.

2.3 SR Models

The primary objective is to design a SR model $\mathcal{M} : \mathbb{R}^{\bar{w} \times \bar{h} \times c} \rightarrow \mathbb{R}^{w \times h \times c}$, such that it inverses Equation 1:

$$\hat{\mathbf{y}} = \mathcal{M}(\mathbf{x}; \theta), \tag{3}$$

where $\hat{\mathbf{y}}$ is the predicted HR approximation of the LR image $\mathbf{x}$ and θ the parameters of $\mathcal{M}$. Consequently, this translates into an optimization objective that minimizes Equation 2, the difference between the estimation $\hat{\mathbf{y}}$ and the ground-truth HR image $\mathbf{y}$ under a given loss function $\mathcal{L}$. The following sections will focus on standard methods to develop such a SR model, especially conventional deep learning methods, before we examine in detail how diffusion models fulfill the role of a SR model. This includes a brief overview of conventional methods and a short examination of state-of-the-art deep learning methods.

2.3.1 Traditional Methods

Traditional methods for image SR define a range of methodologies, such as statistical [22], edge-based [23], [24] patch-based [25], [26], prediction-based [27], [28] and sparse representation techniques [29]. These methodologies fundamentally rely on image statistics and the information inherent in existing pixels to generate HR images. Despite their utility, a noteworthy drawback of these traditional methods is the potential introduction of noise, blur, and visual artifacts.

2.3.2 Deep Learning Methods

With advancements in Deep Learning (DL) and computational power, image SR has significantly evolved, with DL-based approaches providing notable improvements over traditional methods. DL-based SR methods mainly employ Convolutional Neural Networks (CNNs) for end-to-end mapping from LR to HR images. Initial models, such as SRCNN [30], FSRCNN [31], and ESPCNN [32], utilized simple CNNs of diverse depth and feature maps sizes.

Subsequently, concepts from the broader CV domain were incorporated into SR models, e.g., ResNet led to SRResNet [33], where residual information was propagated to successive network layers. Two approaches were introduced to learn residuals - local and global, each differing in their objectives and internal network connections. DenseNet [34] was also adapted to this task through the use of dense blocks where each layer receives the feature maps generated in all the preceding layers in addition to the inputs like in SRDenseNet [35] and ESRGAN [36]. Recursive CNNs that recursively use the same module to learn representations were also inspired by other CV methods for DL-based SR methods in DRCN [37], DRRN [38] and CARN [39]. More recently, attention mechanisms have been incorporated to enhance feature extraction by focusing on regions of interest in images, predominantly via the channel and spatial attention mechanisms [14].

DL-based models can be further categorized into three directions based on the objective used to train the neural network: Pixel-based, GAN-based, and Flow-based techniques, which are introduced next.

2.3.3 Pixel-based Methods

Pixel-based techniques utilize loss functions designed to minimize the difference directly in pixel space. Commonly used loss functions are the L1 and L2 losses. Models trained with pixel-based loss functions tend to generate images with a tendency toward blurriness and excessive smoothing, which can result in less high-frequency details and even lead to unrealistic visual appearances [14]. One way to overcome this limitation is to use generative models, such as GANs, flow-based methods, or diffusion models.

2.3.4 Generative Adversarial Networks (GANs)

GANs use two CNNs: A generator G and a discriminator D, which are trained simultaneously. The generator aims to produce HR samples that are as close to the original as to fool the discriminator, which tries to distinguish between generated and real samples. This framework, e.g., in SRGAN [40] or ESRGAN [36], is optimized using a combination of adversarial loss and content loss to produce less-smoothed images. The resultant images of state-of-the-art GANs are sharper and more detailed. Due to their capability to generate high-quality and diverse images, they have received much attention lately. However, they are susceptible to mode collapse, have a sizeable computational footprint, sometimes fail to converge, and require additional methods for stabilization [13].

2.3.5 Flow-based Methods

Flow-based methods employ optical flow algorithms to generate SR images [41]. They were introduced in an attempt to counter the ill-posed nature of image SR by learning the conditional distribution of plausible HR images given a LR input. They introduce a conditional normalized flow architecture that aligns LR and HR images by calculating the displacement field between them and then uses this information to recover SR images. They employ a fully invertible encoder capable of mapping any input HR image to the latent flow space and ensuring exact reconstruction. This framework enables the SR model to learn rich distributions using exact log-likelihood-based training [41]. This facilitates flow-based methods to circumvent training instability but incurs a substantial computational cost.

2.4 Image Quality Assessment (IQA)

Image quality is a multifaceted concept that addresses various properties, such as sharpness, contrast, and absence of noise. Hence, a fair evaluation of SR models based on produced image quality forms a non-trivial task. This section presents the essential methods, especially for diffusion models, to assess image quality in the context of image SR, which fall under the umbrella term Image Quality Assessment (IQA) [1]. At its core, IQA refers to any metric that resembles the perceptual evaluations from human observers, specifically, the level of realism perceived in an image after the application of SR techniques. During this section, we will use the following notation: $N_{\mathbf{x}} = w \cdot h \cdot c$, which defines the number of pixels of an image $\mathbf{x} \in \mathbb{R}^{w \times h \times c}$ and $\Omega_{\mathbf{x}} = \{(i,j,k) \in \mathbb{N}_1^3 | i \leq h, j \leq w, k \leq c\}$ that defines the set of all valid positions in $\mathbf{x}$.

1. More SR-related IQA methods can be found in Moser *et al.* [14].

2.4.1 Peak Signal-to-Noise Ratio (PSNR)

The Peak Signal-to-Noise Ratio (PSNR) is one of the most widely used techniques to evaluate SISR reconstruction quality. It represents the ratio between the maximum pixel value L and the Mean Squared Error (MSE) between the SR image $\hat{\mathbf{y}}$ and the actual image $\mathbf{y}$.

$$\mathrm{PSNR}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = 10 \cdot \log_{10}\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{y} - \hat{\mathbf{y}}\right]^2}\right) \tag{4}$$

Despite being one of the most popular IQA methods, it does not accurately match human perception [42]. It focuses on pixel differences, which can often be inconsistent with the subjectively perceived quality: the slightest shift in pixels can result in worse PSNR values while not affecting human perceptual quality. Due to its pixel-level calculation, models trained with correlated pixel-based loss tend to achieve high PSNR values [14], whereas generative models tend to produce lower PSNR values [42].

2.4.2 Structural Similarity Index (SSIM)

The SSIM, like the PSNR, is a popular evaluation method that focuses on the differences in structural features between images. It independently captures the structural similarity by comparing luminance, contrast, and structures. SSIM estimates for an image $\mathbf{y}$ the luminance $\mu_{\mathbf{y}}$ as the mean of the intensity, while it is estimating contrast $\sigma_{\mathbf{y}}$ as its standard deviation:

$$\mu_{\mathbf{y}} = \frac{1}{N_{\mathbf{y}}}\sum_{p \in \Omega_{\mathbf{y}}} \mathbf{y}_p, \tag{5}$$

$$\sigma_{\mathbf{y}} = \frac{1}{N_{\mathbf{y}} - 1}\sum_{p \in \Omega_{\mathbf{y}}} \left[\mathbf{y}_p - \mu_{\mathbf{y}}\right]^2 \tag{6}$$

To capture the similarity between the computed entities, the authors introduced a comparison function S:

$$S\left(x, y, c\right) = \frac{2 \cdot x \cdot y + c}{x^2 + y^2 + c}, \tag{7}$$

where x and y are the scalar variables being compared, and $c = (k \cdot L)^2$, $0 < k \ll 1$ is a constant for numerical stability. For a HR image $\mathbf{y}$ and its approximation $\hat{\mathbf{y}}$, the luminance ($\mathcal{C}_l$) and contrast ($\mathcal{C}_c$) comparisons are computed using $\mathcal{C}_l\left(\mathbf{y}, \hat{\mathbf{y}}\right) = S\left(\mu_{\mathbf{y}}, \mu_{\hat{\mathbf{y}}}, c_1\right)$ and $\mathcal{C}_c\left(\mathbf{y}, \hat{\mathbf{y}}\right) = S\left(\sigma_{\mathbf{y}}, \sigma_{\hat{\mathbf{y}}}, c_2\right)$, where $c_1, c_2 > 0$. The empirical covariance

$$\sigma_{\mathbf{y},\hat{\mathbf{y}}} = \frac{1}{N_{\mathbf{y}} - 1}\sum_{p \in \Omega_{\mathbf{y}}} \left(\mathbf{y}_p - \mu_{\mathbf{y}}\right) \cdot \left(\hat{\mathbf{y}}_p - \mu_{\hat{\mathbf{y}}}\right), \tag{8}$$

defines the structure comparison ($\mathcal{C}_s$), which is the correlation coefficient between $\mathbf{y}$ and $\hat{\mathbf{y}}$:

$$\mathcal{C}_s\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \frac{\sigma_{\mathbf{y},\hat{\mathbf{y}}} + c_3}{\sigma_{\mathbf{y}} \cdot \sigma_{\hat{\mathbf{y}}} + c_3}, \tag{9}$$

where $c_3 > 0$. Finally, the SSIM is defined as:

$$\mathrm{SSIM}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \left[\mathcal{C}_l\left(\mathbf{y}, \hat{\mathbf{y}}\right)\right]^{\alpha} \cdot \left[\mathcal{C}_c\left(\mathbf{y}, \hat{\mathbf{y}}\right)\right]^{\beta} \cdot \left[\mathcal{C}_s\left(\mathbf{y}, \hat{\mathbf{y}}\right)\right]^{\gamma} \tag{10}$$

where $\alpha > 0, \beta > 0$, and $\gamma > 0$ are parameters that can be adjusted to tune the relative importance of the components.

2.4.3 Mean Opinion Score (MOS)

The MOS is a subjective measure that leverages human perceptual quality for the evaluation of the generated SR images. Human viewers are shown SR images and asked to rate them with quality scores that are then mapped to numerical values and later averaged. Typically, these range from 1 (bad) to 5 (good) but may vary. While this method is a direct evaluation of human perception, it is more time-consuming and cumbersome to conduct compared to objective metrics. Moreover, due to the highly subjective nature of this metric, it is susceptible to bias.

2.4.4 Consistency

Consistency measures the degree of stability of non-deterministic SR methods, such as generative models like GANs or DMs. Like flow-based methods, generative approaches are intentionally designed to generate a spectrum of plausible outputs for the same input. However, low consistency is not desirable. Minor variations lessen the influence of a relatively consistent method in the input. Nevertheless, consistency can vary depending on the requirements. One commonly employed metric to quantify consistency is the Mean Squared Error.

2.4.5 Learned Perceptual Image Patch Similarity (LPIPS)

Contrary to the pixel-based evaluation of PSNR and SSIM, the Learned Perceptual Image Patch Similarity (LPIPS) utilizes a pre-trained CNN φ, e.g., VGG [43] or AlexNet [44], and generates L feature maps from the SR and HR image, and subsequently calculates the similarity between them. Given h_l and w_l as the height and width of the l-th feature map respectively, and a scaling vector $\alpha_l \in \mathbb{R}^{C_l}$, the LPIPS metric is formulated as follows:

$$\text{LPIPS}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \sum_{l=1}^{L} \sum_{p} \frac{\left\| \alpha_l \odot \left(\varphi^l\left(\hat{\mathbf{y}}\right) - \varphi^l\left(\mathbf{y}\right)\right)_p \right\|_2^2}{h_l \cdot w_l} \tag{11}$$

LPIPS operates by projecting images into a perceptual feature space through φ and evaluating the difference between corresponding patches in SR and HR images, scaled by α_l. This methodology allows for a more human-centric evaluation, given that it is better aligned with human perception than traditional metrics such as PSNR and SSIM [14].

2.4.6 Fool Rate

The Fool Rate is a 2-Alternative Forced-Choice (2AFC) paradigm that measures how humans discriminate true images from model-generated ones using two tasks [42]. In the first task, the subjects are asked to pick the HR image they believe is of higher quality. They are shown an LR input between two HR images of the ground truth and the generated SR image. In task 2, they are asked to guess which image they believe to be from a camera between the HR ground truth and generated SR image. The fool rate is the fraction of trials in which the subjects chose the generated SR over the ground truth.

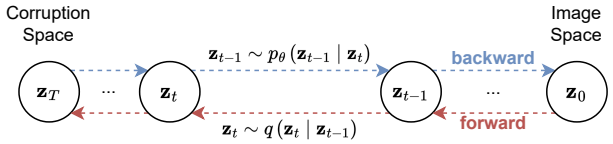

Fig. 1: Principle of Diffusion Models. The forward diffusion adds noise iteratively (red). The backward diffusion, the iterative refinement process, reverts the process (blue).

3 Diffusion Models Basics

Diffusion Models (DMs) have profoundly impacted the realm of generative AI, and many approaches that fall under the umbrella term DM have emerged. At their core, these models share a fundamental principle called diffusion: generate data through the gradual introduction of noise into an initial, often simplistic, data sample, followed by the reversal of this process until it converges to the desired complex data distribution. What sets DMs apart from earlier generative models is their execution over iterative time steps, both forward and backward in time and denoted by t, as depicted in Figure 1. The forward and backward diffusion processes are distinguished by:

Forward Diffusion q - degrade input data using noise iteratively, forward in time (i.e., t increases).

Backward Diffusion p - denoise the degraded data, thereby reversing the noise iteratively, backward in time (i.e., t decreases).

3.1 Basics

Consider a dataset of LR-HR image-pairs, $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$. For each time step t, the random variable $\mathbf{z}_t$ describes the current state. The time step t increases incrementally during forward diffusion, whereas it propagates towards 0 during backward diffusion. In literature, there is no clear distinction between $\mathbf{z}_t$ in the forward diffusion and $\mathbf{z}_t$ in the backward diffusion. During forward diffusion, we assume $\mathbf{z}_t \sim q\left(\mathbf{z}_t \mid \mathbf{z}_{t-1}\right)$. Conversely, in the backward diffusion, we assume $\mathbf{z}_{t-1} \sim p\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t\right)$. We will denote T with $0 < t \leq T$ as the maximal time step for finite cases. The initial data distribution ($t = 0$) is represented by $\mathbf{z}_0 \sim q\left(\mathbf{x}\right)$, which is then slowly injected with noise (additive). Vice versa, DMs remove noise therein by running a parameterized model $p_\theta\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t\right)$ in the reverse time direction that approximates the ideal (but unattainable) denoised distribution $p\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t\right)$.

The explicit implementation of the forward diffusion q and backward diffusion p, approximated by p_θ, is influenced by the specific DM in use. Three primary formulations are prevalent in this domain: Denoising Diffusion Probabilistic Models (DDPMs), Score-Based Generative Models (SGMs), and Stochastic Differential Equations (SDEs) [45]. Each of these types will be discussed next.

3.2 Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs) [6] use two Markov chains to enact the forward and backward diffusion across a finite amount of time steps.

Forward Diffusion: It transforms the data distribution into a prior distribution, typically designed manually (e.g., Gaussian), given by:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t \mid \sqrt{1-\alpha_t}\,\mathbf{z}_{t-1}, \alpha_t \mathbf{I}), \tag{12}$$

where the hyper-parameters $0 < \alpha_{1:T} < 1$ represent the variance of noise incorporated at each time step. While the Gaussian kernel is commonly adopted, alternative kernel types can also be employed. This formulation can be condensed to a single-step calculation, as shown by:

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t \mid \sqrt{\gamma_t}\,\mathbf{z}_0, (1-\gamma_t)\mathbf{I}), \tag{13}$$

where $\gamma_t = \prod_{i=1}^{t}(1-\alpha_i)$ [46]. Consequently, $\mathbf{z}_t$ can be directly sampled regardless of what ought to happen on previous time steps by

$$\mathbf{z}_t = \sqrt{\gamma_t} \cdot \mathbf{z}_0 + \sqrt{1-\gamma_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{14}$$

Backward Diffusion: The goal is to directly learn the inverse of the forward diffusion and generate a distribution that resembles the prior $\mathbf{z}_0$, usually the HR image in SR. In practice, we use a CNN to learn a parameterized form of p. Since the forward process approximates $q(\mathbf{z}_T) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$, the formulation of the learnable transition kernel becomes:

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1} \mid \mu_\theta(\mathbf{z}_t, \gamma_t), \Sigma_\theta(\mathbf{z}_t, \gamma_t)), \tag{15}$$

where μ_θ and Σ_θ are learnable. Similarly, the conditional formulation $p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{x})$ conditioned on $\mathbf{x}$ (e.g., a LR image) is using $\mu_\theta(\mathbf{z}_t, \mathbf{x}, \gamma_t)$ and $\Sigma_\theta(\mathbf{z}_t, \mathbf{x}, \gamma_t)$ instead.

Optimization: To guide the backward diffusion in learning the forward process, we minimize the Kullback-Leibler (KL) divergence of the joint distribution of the forward and reverse sequences

$$p_\theta(\mathbf{z}_0, ..., \mathbf{z}_T) = p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t), \text{ and} \tag{16}$$

$$q(\mathbf{z}_0, ..., \mathbf{z}_T) = q(\mathbf{z}_0) \prod_{t=1}^{T} q(\mathbf{z}_t \mid \mathbf{z}_{t-1}), \tag{17}$$

which leads to minimizing

$$\begin{aligned} &\mathrm{KL}(q(\mathbf{z}_0, ..., \mathbf{z}_T) \| p_\theta(\mathbf{z}_0, ..., \mathbf{z}_T)) \\ &= -\mathbb{E}_{q(\mathbf{z}_0, ..., \mathbf{z}_T)}[\log p_\theta(\mathbf{z}_0, ..., \mathbf{z}_T)] + c \\ &\overset{(i)}{=} \mathbb{E}_{q(\mathbf{z}_0, ..., \mathbf{z}_T)}\left[-\log p(\mathbf{z}_T) - \sum_{t=1}^{T} \log \frac{p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t)}{q(\mathbf{z}_t \mid \mathbf{z}_{t-1})}\right] + c \\ &\overset{(ii)}{\geq} \mathbb{E}[-\log p_\theta(\mathbf{z}_0)] + c, \end{aligned} \tag{18}$$

where (i) is possible because both terms are products of distributions and (ii) is the product of Jensen's inequality. The constant c is unaffected and, therefore, irrelevant in optimizing θ. Note that Equation 18 without c is the Variational Lower Bound (VLB) of the log-likelihood of the data $\mathbf{z}_0$, which is commonly maximized by DDPMs.

Ho *et al.* [6] recommend reweighting the VLB for improved sample quality:

$$\mathop{\mathbb{E}}_{\substack{t \sim \mathcal{U}(1,T) \\ \mathbf{z}_0 \sim q(\mathbf{z}_0) \\ \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}} \left[\lambda(t) \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|^2\right], \tag{19}$$

where $\lambda(t) > 0$ is a weighting function, $\mathbf{z}_t$ is derived by Equation 14 and ϵ_θ is a NN that predicts the noise vector ϵ given $\mathbf{z}_t$ and t. For a specific choice of $\lambda(t)$, Equation 19 reduces to Equation 18. We present this formulation to show later that it intriguingly bears a resemblance to the objective of denoising score matching for training Score-based Generative Models (SGMs), which will be introduced next, and draws the connection between both approaches.

3.3 Score-based Generative Models (SGMs)

Score-based Generative Models (SGMs), much like DDPMs, utilize diffusion processes but employ an alternative mathematical foundation. Instead of using probability density function $p(\mathbf{z})$ directly, Song *et al.* [9] propose to work with its (Stein) score function, which is defined as the gradient of the log probability density $\nabla_{\mathbf{z}} \log p(\mathbf{z})$. Mathematically, the score function preserves all information about the density function, but computationally it is easier to work with. Furthermore, the decoupling of model training from the sampling procedure grants greater flexibility in defining sampling methods and training objectives.

Forward Diffusion: Let $0 < \sigma_1 < ... < \sigma_T$ be a finite sequence of noise levels. Like DDPMs, the forward diffusion, typically assigned to a Gaussian noise distribution, is

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t \mid \mathbf{z}_0, \sigma_t^2 \mathbf{I}). \tag{20}$$

This equation results in a sequence of noisy data densities $q(\mathbf{z}_1), ..., q(\mathbf{z}_T)$ with $q(\mathbf{z}_t) = \int q(\mathbf{z}_t) q(\mathbf{z}_0) \mathrm{d}\mathbf{z}_0$. Consequently, the intermediate step $\mathbf{z}_t = \mathbf{z}_0 + \sigma_t \cdot \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ can be sampled agnostic from previous time steps in a single step.

Backward Diffusion: To revert the noise during the backward diffusion, we need to approximate $\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)$ and choose a method for estimating the intermediate states $\mathbf{z}_t$ from that approximation. For the gradient approximation at each time step t, we use a trained predictor, denoted as s_θ and called Noise-Conditional Score Network (NCSN), such that $s_\theta(\mathbf{z}_t, t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)$ [9]. The training of the NCSN will be covered in the next section; for now, we focus on the sampling process using NCSN. Sampling with NCSN involves generating the intermediate states $\mathbf{z}_t$ through an iterative approach, using $s_\theta(\mathbf{z}_t, t)$. Note that this iterative process is different from the iterations done during the diffusion as it addresses solely the generation of $\mathbf{z}_t$.

There are various ways to perform this iterative generation, but we will concentrate on a specific method known as Annealed Langevin Dynamics (ALD), introduced by Song et al. [8]. Let N be the number of estimation iterations for $\mathbf{z}_t$ at time step t and $\alpha_t > 0$ the corresponding step size, which determines how much the estimation moves from one estimate $\mathbf{z}_{t-1}^{(i)}$ towards $\mathbf{z}_{t-1}^{(i+1)}$. The initial state is $\mathbf{z}_T^{(N)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For each $0 < t \leq T$, we initialize $\mathbf{z}_{t-1}^{(0)} = \mathbf{z}_t^{(N)} \approx \mathbf{z}_t$, which is the latest estimation of the previous intermediate state. In order to get $\mathbf{z}_{t-1}^{(N)} \approx \mathbf{z}_{t-1}$ iteratively, ALD uses the following update rules for $i = 0, ..., N-1$:

$$\epsilon^{(i)} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{21}$$

$$\mathbf{z}_{t-1}^{(i+1)} \leftarrow \mathbf{z}_{t-1}^{(i)} + \frac{1}{2}\alpha_{t-1} s_\theta(\mathbf{z}_{t-1}^{(i)}, t-1) + \sqrt{s_{t-1}}\epsilon^{(i)} \tag{22}$$

This update rule guarantees that $\mathbf{z}_0^{(N)}$ converges to $q(\mathbf{z}_0)$ for $\alpha_t \to 0$ and $N \to \infty$ [47].

Similar to DDPMs, we can turn SGMs into conditional SGMs by integrating the condition $\mathbf{x}$, e.g., a LR image, into $s_\theta(\mathbf{z}_t, \mathbf{x}, t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{x})$.

Optimization: Without specifically formulating the backward diffusion, we can train a NCSN such that $s_\theta(\mathbf{z}_t, t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)$. Estimating the score can be done by using the denoising score matching method [48]:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T)\\ \mathbf{z}_0\sim q(\mathbf{z}_0)\\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t)\sigma_t^2 \|\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t) - s_\theta(\mathbf{z}_t, t)\|^2\right] \qquad (23)\\
\overset{(i)}{=}\ &\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T)\\ \mathbf{z}_0\sim q(\mathbf{z}_0)\\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t)\sigma_t^2 \|\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{z}_0) - s_\theta(\mathbf{z}_t, t)\|^2\right] + c\\
\overset{(ii)}{=}\ &\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T)\\ \mathbf{z}_0\sim q(\mathbf{z}_0)\\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t) \| - \frac{\mathbf{z}_t - \mathbf{z}_0}{\sigma_t} - \sigma_t s_\theta(\mathbf{z}_t, t)\|^2\right] + c\\
\overset{(iii)}{=}\ &\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T)\\ \mathbf{z}_0\sim q(\mathbf{z}_0)\\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}} \left[\lambda(t) \|\epsilon + \sigma_t s_\theta(\mathbf{z}_t, t)\|^2\right] + c
\end{aligned}$$

where $\lambda(t) > 0$ is a weighting function, σ_t the noise level added at time step t, (i) derived by Vincent *et al.* [48], (ii) from Equation 20, (iii) from $\mathbf{z}_t = \mathbf{z}_0 + \sigma_t\epsilon$ and with c again a constant unaffected in the optimization of θ. Note that there are other ways to estimate the score, e.g., based on score matching [49] or sliced score matching [50].

As mentioned in the chapter before, Equation 19 resembles Equation 23, which becomes clear once we set $\epsilon_\theta(\mathbf{z}_t, t) = -\sigma_t s_\theta(\mathbf{z}_t, t)$ in Equation 19. Thus, there is a mathematical connection between DDPMs and SGMs.

3.4 Stochastic Differential Equations (SDEs)

So far, we have discussed DMs that deal with finite time steps. A generalization to infinite time steps is made by formulating these as solutions to Stochastic Differential Equations (SDEs), also known as Score SDEs [51].

Forward Diffusion: We can represent the forward diffusion by the following SDE:

$$\mathrm{d}\mathbf{z} = \mathrm{f}(\mathbf{z}, t)\mathrm{d}t + \mathrm{g}(t)\mathrm{d}\mathbf{w}, \qquad (24)$$

where f and g are the drift and diffusion functions, respectively, and $\mathbf{w}$ is the standard Wiener process (also known as Brownian motion). The SDE for DDPMs is given by:

$$\mathrm{d}\mathbf{z} = -\frac{1}{2}\alpha(t)\mathbf{z}\mathrm{d}t + \sqrt{\alpha(t)}\mathrm{d}\mathbf{w}, \qquad (25)$$

with $\alpha(\frac{t}{T}) = T\alpha_t$ for $T \to \infty$. For SGMs, the SDE is

$$\mathrm{d}\mathbf{z} = \sqrt{\frac{\mathrm{d}\left[\sigma(t)^2\right]}{\mathrm{d}t}}\mathrm{d}\mathbf{w}, \qquad (26)$$

with $\sigma(\frac{t}{T}) = \sigma_t$ for $T \to \infty$. From now on, we denote with $q_t(\mathbf{z})$ the distribution of $\mathbf{z}_t$ in the diffusion process.

Backward Diffusion: The reverse-time SDE is formulated by Anderson *et al.* [52] as:

$$\mathrm{d}\mathbf{z} = \left[\mathrm{f}(\mathbf{z}, t) - \mathrm{g}(t)^2\nabla_{\mathbf{z}} \log q_t(\mathbf{z})\right]\mathrm{d}t + \mathrm{g}(t)\mathrm{d}\tilde{\mathbf{w}}, \qquad (27)$$

where $\tilde{\mathbf{w}}$ is the standard Wiener process when time flows backwards and $\mathrm{d}t$ an infinitesimal negative time step. Solutions to Equation 27 can be viewed as diffusion processes that gradually convert noise to data. The existence of a corresponding probability flow Ordinary Differential Equation (ODE), whose trajectories possess the same marginals as the reverse-time SDE, was proven by Song *et al.* [51] and is

$$\mathrm{d}\mathbf{z} = \left[\mathrm{f}(\mathbf{z}, t) - \frac{1}{2}\mathrm{g}(t)^2\nabla_{\mathbf{z}} \log q_t(\mathbf{z})\right]\mathrm{d}t. \qquad (28)$$

Thus, the reverse-time SDE and the probability flow ODE enable sampling from the same data distribution.

Optimization: Similar to the approach in SGMs, we define a score model such that $s_\theta(\mathbf{z}_t, t) \approx \nabla_{\mathbf{z}} \log q_t(\mathbf{z})$. Additionally, we extend Equation 23 to continuous time as follows:

$$\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(0,T)\\ \mathbf{z}_0\sim q(\mathbf{z}_0)\\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t)\|s_\theta(\mathbf{z}_t, t) - \nabla_{\mathbf{z}_t} \log q_t(\mathbf{z}_t \mid \mathbf{z}_0)\|^2\right], \qquad (29)$$

where $\lambda(t) > 0$ is a weighting function.

3.5 Relation to other Image SR Generative Models

The exploration of generative models in image SR is fascinating because it offers the potential to improve the quality of upsampled images beyond interpolation significantly. They differ primarily in how they approach the task of generating HR images from LR inputs. These differences stem from the underlying architecture and training objectives. While they offer significant advantages, they come with their individual set of challenges, such as training stability and computational costs.

GAN: One prominent category of generative models is Generative Adversarial Networks (GANs) [53], which have demonstrated state-of-the-art performance in various vision-related tasks, including text-to-image synthesis [13] and image super-resolution (SR) [40]. GANs are known for their adversarial training, where a generator competes against a discriminator. While DMs do not use a discriminator in the same way, they also employ an adversarial-like training strategy, where noise is iteratively added and removed to encourage the generation of realistic data. However, approaches with GANs often suffer from non-convergence, training instability, and high computational costs. They require careful hyperparameter tuning due to the interplay between the generator and the discriminator.

VAE: Variational Autoencoders (VAEs) [54] are designed as autoencoders with a variational latent space, which is especially interesting in addressing the ill-posedness of image SR. The core objective of a VAE centers around establishing the variational lower bound of the log data likelihood, akin to the fundamental principle underlying DMs. In a comparative context, one can consider DMs as a variation of VAEs but with a fixed VAE encoder responsible for perturbing the input data, while the VAE decoder resembles the backward diffusion process in DMs. Still, unlike VAEs, which compress the input into smaller dimensions in the latent space, DMs often maintain the same spatial size.

ARM: Autoregressive Models (ARMs) treat images as sequences of pixels and generate each pixel based on the

values of previously generated pixels in a sequential manner [12]. The probability of the entire image is given as the product of conditional probability distributions for each individual pixel. This makes ARMs computationally expensive for HR image generation. Conversely, DMs generate data by gradually diffusing noise into an initial data sample and then reverse this process. Noise is diffused across the entire image simultaneously rather than sequentially.

NF: Normalizing Flows (NF) [55] are a distinct category of generative models renowned for their capacity to represent data as intricate and complex distributions. Like DMs and VAEs, these models are optimized based on the log-likelihood of the data they generate. However, what sets NFs apart is their unique ability to learn an invertible parameterized transformation. Importantly, this transformation possesses a tractable Jacobian determinant, making it feasible to compute. The concept of DiffFlow [56] enters the picture as an innovative algorithm that marries the principles of DMs with those of NFs. This combination offers the promise of enhanced generative modeling capabilities. Yet, while promising, NFs are often considered challenging to train and can be computationally demanding [57].

4 Improvements for Diffusion Models

This section introduces enhancements made to the foundational principles we have previously outlined. Two particular advancements, namely, efficient sampling and enhanced likelihood estimation, stand out due to their direct relevance and adaptability to the domain of image SR.

4.1 Efficient Sampling

Efficient sampling plays a pivotal role in addressing one of the central challenges hindering the widespread adoption of DMs - the considerable computational resources required for sampling. To put it in perspective, generating samples using DMs can demand processing power equivalent to hundreds of GPU days. For instance, sampling 50,000 images of size 32x32 with a DDPM can take about 20 hours, while a GAN can achieve this in less than a minute on a Nvidia 2080 Ti GPU; the challenge escalates for larger images, as sampling 50,000 images of size 256x256 could require close to 1,000 hours on the same hardware [58]. Advanced sampling methodologies have emerged as a solution, expediting the sampling process while upholding the quality of the generated images. A key aspect of those models to achieve this is the independence of training and inference schedules. In other words, while a model may undergo training over 1000 steps, the subsequent inference phase in practice may require only a fraction. Efficient sampling strategies can be categorized into Training-based and Training-free sampling.

Training-based sampling methods speed up data generation using a trained sampler that approximates the backward diffusion process instead of a traditional numerical solver. This can be done either entirely or partially. For example, Watson et al. [59] developed a dynamic programming algorithm that identifies optimal inference paths using a fixed number of refinement steps, significantly reducing the computation required. Diffusion Sampler Search [60] offers another approach, optimizing fast samplers for pre-trained diffusion models by adjusting the Kernel Inception Distance.

Knowledge distillation is also used to accelerate sampling. It involves transferring knowledge from a complex, slower sampler (the teacher model) to simpler, faster models (student models) [61], [62]. As demonstrated by Salimans et al. [63], this method progressively reduces the number of sampling steps, trading off a slight decrease in sample quality for increased speed. Similarly, Xiao et al. [64] addressed the slow sampling issue associated with the Gaussian assumption in denoising steps, which is usually only effective for small step sizes. They proposed Denoising Diffusion GANs that use conditional GANs for the denoising steps, allowing for larger step sizes and faster sampling.

Lastly, truncated diffusion is a technique that improves speed by prematurely ending the forward diffusion process [65], [66]. This early termination results in outputs that are not purely Gaussian random noise, presenting computational challenges. These challenges are overcome using proxy distributions from pre-trained VAEs or GANs, which match the diffused data distribution and facilitate efficient backward diffusion.

Training-free sampling methods aim to speed up sampling by minimizing the number of discretization steps while solving the Stochastic Differential Equation (SDE) or Probability Flow Ordinary Differential Equation (ODE) [67], [68].

Denoising Diffusion Implicit Models (DDIMs) [67] introduced by Song *et al.* generalizes the Markovian forward diffusion of DDPMs into non-Markovian ones. This generalization allows the DDIMs to learn a Markov chain to reverse the non-Markovian forward diffusion, resulting in higher sampling speeds with minimal loss in sample quality.

In another noteworthy development, Jolicoeur-Martineau *et al.* [68] have devised an efficient SDE solver with adaptive step sizes for the accelerated generation of score-based models. This method has been found to generate samples more rapidly than the Euler-Maruyama method without any compromise in sample quality.

4.2 Improved Likelihood

Log-likelihood improvement is directly coupled with enhancing the performance of various applications and methods, including but not limited to compression [69], semi-supervised learning [70], and image SR. Given that DMs do not directly optimize the log-likelihood, e.g., SGMs utilize a weighted combination of score-matching losses, an objective that forms an upper bound on the negative log-likelihood needs to be optimized.

Song *et al.* [71] proposed a method called *likelihood weighting* to address this need. This method minimizes the weighted combination of score matching losses for score-based DMs. A carefully selected weighting function provides an upper bound to the negative log-likelihood within the weighted score-matching objective. Upon minimization, this results in an elevation of the log-likelihood.

Kingma *et al.* [72] explored methods that simultaneously train the noise schedule and diffusion parameters to maximize the variational lower bound within Variational Diffusion Models.

Additionally, the Improved Denoising Diffusion Probabilistic Models (iDDPMs) proposed by Nichol and Dhariwal *et al.* [73] implement a cosine noise schedule. This gradually introduces noise into the input, contrasting with the linear schedules that tend to degrade the information quicker. Using the cosine noise schedule leads to better log-likelihoods and facilitates faster sampling.

5 Diffusion Models for Image SR

So far, our discussion has revolved around the theoretical framework of DMs. This section explores the practical applications and recent publications in image SR.

5.1 Concrete Realization of DDPMs

Among the pioneering DM efforts in SR is SR3 [42], a method that concretely realizes DDPMs for image SR. Like typical in DDPMs, it adds Gaussian noise to the LR image until $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and generates a target HR image $\mathbf{z}_0$ iteratively in T refinement steps. SR3 employs the denoising model to predict the noise ϵ_t. The denoising model, $\varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t)$, takes the LR image $\mathbf{x}$, the noise variance γ_t, and the noisy target image $\mathbf{z}_t$ as inputs.

With the prediction of ϵ_t provided by φ_θ, we can reformulate Equation 14 to approximate $\mathbf{z}_0$ as follows:

$$\begin{aligned} \mathbf{z}_t &= \sqrt{\gamma_t} \cdot \hat{\mathbf{z}}_0 + \sqrt{1-\gamma_t} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t) \\ \iff \hat{\mathbf{z}}_0 &= \frac{1}{\sqrt{\gamma_t}} \cdot \left(\mathbf{z}_t - \sqrt{1-\gamma_t} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t)\right) \end{aligned} \tag{30}$$

The substitution of $\hat{\mathbf{z}}_0$ into the posterior distribution to parameterize the mean of $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})$ leads to:

$$\mu_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t) = \frac{1}{\sqrt{\alpha_t}} \left[\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t)\right] \tag{31}$$

In SR3, the authors simplified the variance Σ_θ to $(1-\alpha_t)$ for ease of computation. Consequently, each refinement step with $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ can be represented as:

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left[\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t)\right] + \sqrt{1-\alpha_t} \cdot \epsilon_t \tag{32}$$

Concurrent work focused on a similar implementation of SR3 but shows different variations implementing the denoising model $\varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t)$, which we will discuss later.

A notable mention is SRDiff [74], published around the same time and follows a close realization of SR3. The main distinction between SRDiff and SR3 is that SR3 predicts the HR image directly, whereas SRDiff predicts the residual information between the LR and HR image, i.e., the difference. Thus, it focuses on an alternative target domain, which is part of the next section.

5.2 Alternative Domains

So far, we have discussed methods that operate on the pixel space directly. This section will look at different methods that tackle the problem by mapping it to latent and frequency space. Apart from particular challenges arising from the alternative input domain, these methods incur an additional step that maps the pixel domain into their own, as illustrated in Figure 2. Moreover, we will introduce the residual target domain.

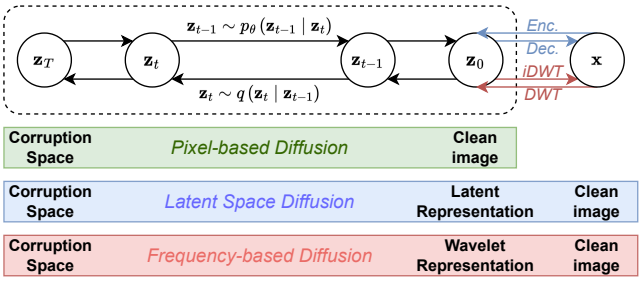

Fig. 2: Overview of alternative domains. The green bar shows the vanilla DM that operates in pixel space. The blue bar shows the exploit of the latent space domain via Autoencoders. The red bar shows the application of DMs within the wavelet domain.

5.2.1 *Latent Space Domain*

Models like SR3 [42] and SRDiff [74] have achieved high-quality SR results by operating in the pixel domain. However, these models are computationally intensive due to their iterative nature and the high-dimensional calculations in RGB space. To reduce computational demands, Latent Diffusion Models (LDMs), introduced by Rombach et al. [75], perform diffusion processes in a low-dimensional latent space. This approach significantly lowers resource requirements without compromising performance and allows the reuse of latent representations across multiple models.

LDMs typically follow a two-stage process. First, a pre-trained autoencoder, like the VQ-GAN [11], compresses input data into a low-dimensional latent representation. This encoder is trained using perceptual and adversarial losses. In the second stage, a time-conditional U-Net model is trained in the latent space, incorporating image-specific biases. For image SR, low-resolution images are concatenated as a condition to this model.

Similarly, Latent Score-based Generative Models (LSGMs), as introduced by Vadhat *et al.* [76], apply SGMs within the latent space. Unlike LDMs, they jointly train their SGM with a VAE. They achieve faster sampling speeds by pre-training their VAE. Another latent model, but for SDEs, is presented by Luo *et al.* [77], [78] with **Refusion** (image **Re**storation with dif**fusion** models). It employs a mean-reverting Image Restoration (IR)-SDE to transform HR images into their degraded counterparts. They leverage an autoencoder to compress the LR image into its latent representation, with skip connections for accessing multi-scale details. Also, the authors introduce nonlinear activation-free blocks (NAFBlocks) [79] into the IR-SDE [78]. These NAFBlocks replace all non-linear activations with an element-wise operation that splits feature channels into two parts and multiplies them to produce one output.

Chen *et al.* [80] employ a similar approach by proposing a two-stage strategy called the Hierarchical Integration Diffusion Model (HI-Diff). In the first stage, a Latent Encoder (LE) compresses the ground truth image to a highly compact latent space representation. This is treated as a prior feature to a Hierarchical Integration Module (HIM) transformer jointly trained with the LE using L1 loss. A DM is trained to generate the prior feature in the next stage.

TABLE 1: PSNR and SSIM comparison on CelebA-HQ face SR 16×16 → 128×128. Consistency measures MSE ($\times 10^{-5}$) between LR inputs and the downsampled SR outputs.

Methods	PSNR ↑	SSIM ↑	Consistency ↓
PULSE [81]	16.88	0.44	161.1
FSRGAN [82]	23.01	0.62	33.8
SR3 (regression) [42]	23.96	0.69	2.71
SR3 (diffusion) [42]	23.04	0.65	2.68
DiWa [83]	23.34	0.67	-
IDM [73]	24.01	0.71	2.14

5.2.2 Wavelet Target Domain

Wavelets provide a novel outlook on SR [14], [84]. The conversion from the spatial to the wavelet domain is lossless and offers significant advantages. The spatial size of an image can be downsized by a factor of four, thereby allowing faster diffusion during the training and inference stages. Moreover, wavelet transformations naturally segregate high-frequency details into distinct channels, which facilitates a more concentrated and intentional focus on high-frequency information, offering a higher degree of control during the SR process [85]. Besides, it can be conveniently incorporated into existing DMs as a plug-in feature.

The diffusion process can interact directly with all wavelet bands as proposed in DiWa [83] or specifically target certain bands while the remaining bands are predicted via standard CNNs. For instance, WaveDM [86] modifies the low-frequency band, whereas WSGM [87] or ResDiff [88] conditions the high-frequency bands relative to the low-resolution image. Altogether, the wavelet domain presents a promising avenue for future research. It provides potential for significant performance acceleration while maintaining, if not enhancing, the quality of SR results.

5.2.3 Residual Target Domain

As mentioned, SRDiff [74] is the pioneering work that advocates for shifting the image generation process into the residual space. This strategy aims to generate the difference between the upsampled LR and the HR image. This enables the DM to focus on residual details, speeds up convergence, and stabilizes the training [14]. Another notable contribution is ResDiff [88], which also seeks to predict the residuals. Whang *et al.* [89] also employs residual predictions as a fundamental component of their predict-and-refine approach for image deblurring. In a different vein, Yue *et al.* [90] presents a technique that constructs a Markov chain of transformations between HR and LR images by manipulating the residual between them. This novel approach substantially enhances transition efficiency, providing an innovative perspective on image manipulation strategies.

5.3 Conditioning Diffusion Models

DMs heavily depend on conditioning information to guide the sampling process toward a reasonable HR prediction. This section reviews various methods for seamlessly integrating conditioning information into backward diffusion.

TABLE 2: Results for 4× SR of general images on DIV2K val. Note that EDSR, FxSR-PD, CAR, and RRDB are regression-based methods that generally produce better PSNR and SSIM scores than generative approaches [42].

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Bicubic	26.70	0.77	0.409
EDSR [91]	28.98	0.83	0.270
FxSR-PD [92]	29.24	0.84	0.239
RRDB [36]	29.44	0.84	0.253
CAR [1]	32.82	0.88	-
RankSRGAN [93]	26.55	0.75	0.128
ESRGAN [36]	26.22	0.75	0.124
SRFlow [94]	27.09	0.76	<u>0.120</u>
SRDiff [74]	27.41	0.79	0.136
IDM [73]	27.59	0.78	-
DiWa [83]	28.09	0.78	0.104

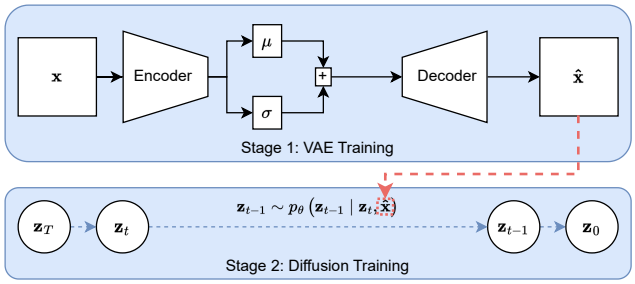

Fig. 3: Overview of DiffuseVAE. The two-stage approach employs a VAE (first stage), which generates variational prediction as a condition for the DM (second stage).

5.3.1 Low Resolution Reference

As SR3 [95] exemplifies, high-quality SR predictions can be achieved through a straightforward channel concatenation. The LR image is concatenated with the denoised result from time step $t-1$ and serves as the conditioning input for noise prediction at time step t. In contrast, Iterative Latent Variable Refinement (ILVR) by Choi *et al.* [96] conditions the generative process of an unconditional LDM. This approach offers the advantage of shorter training times, as it leverages a pre-trained DM. To integrate conditioning information, the low-frequency components of the denoised output are replaced with their corresponding counterparts from the LR image. Thus, the latent variable is aligned with a provided reference image at each generation process stage, ensuring precise control and adaptation during generation.

5.3.2 Super-Resolved Reference

An alternative to conditioning the denoising on the LR image involves learned priors from pre-trained SR models to predict a reference image. E.g., CDPMSR [97] conditions the denoising process with a predicted SR reference image obtained using existing and standard SR models. ResDiff [88], on the other hand, leverages a pre-trained CNN to predict a low-frequency, content-rich image that includes partial high-frequency components. This image guides the noise towards the residual space, offering an alternative means of conditioning the generative process.

Pandey *et al.* [98] introduced an exciting idea of varying predicted conditions with DiffuseVAE as illustrated in Figure 3. This approach integrates the stochastic predictions generated by a VAE as conditioning information for the DM, capitalizing on the advantages offered by both models. They use a two-stage approach called the *generator-refiner* framework. In the first stage, a VAE is trained on the training data. In the subsequent stage, the DM is conditioned using varying, often blurred, reconstructions generated by the VAE. The essential advantage of this method lies in the diversity in the generated samples, which is defined within the lower-dimensional latent space of the VAE. This characteristic creates a more favorable balance between sampling speed and sample quality. It is advantageous in scenarios where multiple predictions are required, similar to the use cases for Normalizing Flows.

5.3.3 Feature Reference

Another avenue for conditioning involves relevant features extracted from pre-trained networks. SRDiff [74] leverages a pre-trained encoder to encode LR image features at each step of the backward diffusion. These features serve as guidance, aiding in the generation of higher-resolution outputs. Implicit DMs (IDMs) [73] take a different approach by conditioning their denoising network with a neural representation, which enables the learning of a continuous representation at various scales. They encode the image as a function within continuous space and seamlessly integrate it into the DM. These extracted features are adapted to multiple scales and are used across multiple layers within the DMs. To comprehensively understand the performance differences between these approaches, comparisons can be found in Table 1 and Table 2.

5.3.4 Exploiting Text-To-Image

By incorporating conditioning information that goes beyond the LR image (e.g., its SR prediction, direct concatenation of the LR image, or its feature representation), one can add Text-To-Image (T2I) information. The incorporation of T2I information proves advantageous as it allows the usage of pre-trained T2I models. These models can be fine-tuned by adding specific layers or encoders tailored to the SR task, facilitating the integration of textual descriptions into the image generation process. This approach enables a richer source of guidance, potentially leading to improved image synthesis and interpretation in SR tasks.

Wang *et al.* have put this concept into practice with StableSR [99]. Central to StableSR is a time-aware encoder trained in tandem with a frozen Stable DM, essentially a LDM. This setup seamlessly integrates trainable spatial feature transform layers, enabling conditioning based on the input image. To further augment the flexibility of StableSR and achieve a delicate balance between realism and fidelity, they introduce an optional controllable feature wrapping module. This module accommodates user preferences, allowing for fine-tuned adjustments based on individual requirements. The inspiration for this feature comes from the methodology introduced in CodeFormer [100], which enhances the versatility of StableSR in catering to diverse user needs and preferences.

TABLE 3: Results for 4× SR of general images on resized DIV2K val ($128 \times 128 \rightarrow 512 \times 512$).

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
BSRGAN [103]	23.41	0.61	0.426
Real-ESRGAN [104]	23.15	0.62	0.403
LDL [105]	22.74	0.62	0.416
FeMaSR [106]	21.86	0.54	0.410
SwinIR-GAN [107]	22.65	0.61	0.406
LDM [75]	21.48	0.56	0.450
SD Upscaler [75]	21.21	0.55	0.430
StableSR [99]	20.88	0.53	0.438
PASD [101]	21.85	0.52	0.403

Likewise, Yang *et al.* introduce a method known as Pixel-Aware Stable Diffusion (PASD) [101]. PASD takes conditioning a step further by incorporating text embeddings of the LR input using a CLIP text encoder [102] and its feature representation. This approach augments the model's ability to generate images by incorporating textual information, thus allowing for more precise and context-aware image synthesis. Comparisons between PASD and other approaches can be found in Table 3, demonstrating the impact of this text-based conditioning on image SR results.

5.4 Guidance in Training

We have examined the learning of conditional distributions, which form the backbone of image SR [42], [108]. As such, the condition $\mathbf{x}$, such as an LR image, is integrated into the backward diffusion, i.e., $p_\theta\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{x}\right)$ for DDPMs or in $s_\theta(\mathbf{z}_t, \mathbf{x}, t)$ for SGMs/SDEs. However, this simple formulation can result in a model that overlooks the conditioning information.

A principle known as guidance can mitigate this issue by controlling the weighting of the conditioning information at the expense of sample diversity. This principle is categorized into Classifier and Classifier-Free guidance. To our knowledge, while effectively used for improving DMs, they have not been applied to the task of SR. We include a brief overview of these methods as they comprise an exciting avenue for future work on image SR.

5.4.1 Classifier Guidance

Classifier guidance employs a classifier to guide the diffusion process by merging the score estimate of the DM with the gradients of the classifier during sampling [109]. This process is similar to low temperature or truncated sampling in BigGANs [110] and facilitates a trade-off between mode coverage and sample fidelity. The classifier is trained concurrently with the DM to predict the conditional information $\mathbf{x}$ from $\mathbf{z}_t$. For weighting of the conditioning information, the score function becomes:

$$\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t \mid \mathbf{x}) = \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t) + \lambda \nabla_{\mathbf{z}_t} \log q(\mathbf{x} \mid \mathbf{z}_t), \quad (33)$$

where $\lambda \in \mathbb{R}^+$ is a hyper-parameter for controlling the weighting. The downside of this approach is its dependence on a learned classifier that can handle arbitrarily noisy inputs, a capability most existing pre-trained image classification models lack.

5.4.2 Classifier-Free Guidance

Classifier-Free guidance aims to achieve similar results without a classifier [111]. It exploits the vanilla formulation such that Equation 33 can be rearranged into

$$\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{x}) = (1-\lambda)\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t) + \lambda \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t \mid \mathbf{x}). \tag{34}$$

As a result, we have a standard unconditional DM and a conditional DM that has the score estimate $\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t \mid \mathbf{x})$. The unconditional DM remains when $\lambda = 0$, and for $\lambda = 1$, it aligns with the vanilla formulation of the conditional DM. The interesting scenario arises when $\lambda > 1$, where the DM prioritizes conditional information and moves away from the unconditional score function, thus reducing the likelihood of generating samples disregarding conditioning information. However, the major downside of this approach is its computational cost for training two separate DMs. This can be mitigated by training a single conditional model and substituting the conditioning information with a null value in the unconditional score function [112].

5.5 Zero-Shot

Zero-shot image SR aims to develop methods that do not depend on prior image examples or training [14], [119]. Typically, these methods harness the inherent redundancy within a single image for improvement. They often leverage pre-trained DMs for generation, incorporating LR images as conditions during the sampling process, in contrast to other conditioning methods discussed earlier. Additionally, they differ from guidance-based methods, where conditioning information is used to weight the training of a DM from scratch. A recent study by Li *et al.* [113] categorizes diffusion-based methods into projection-based, decomposition-based, and posterior estimation, which are introduced in this section. The discussed methods are compared in Table 4.

5.5.1 Projection-Based

Projection-based methods aim to extract inherent structures or textures from LR images to complement the generated images at each step and to ensure data consistency. An illustrative example of a projection-based method in the realm of inpainting tasks is RePaint [120]. In RePaint, the diffusion process is selectively applied to the specific area requiring inpainting, leaving the remaining image portions unaltered. Taking inspiration from this concept, YODA [121] applies a similar technique, but for image SR. YODA incorporates importance masks derived from DINO [122] to define the areas for diffusion during each time step, but it is not a zero-shot approach.

One zero-shot method is ILVR [96], which projects the low-frequency information from the LR image to the HR image, ensuring data consistency and establishing an improved DM condition. A more sophisticated method is Come-Closer-Diffuse-Faster (CCDF) [123], which modifies the unified projection method to SR as follows:

$$\hat{\mathbf{z}}_{t-1} = \mathrm{f}(\mathbf{z}_t, t) + \mathrm{g}(\mathbf{z}_t, t) \cdot \varepsilon_t \tag{35}$$

$$\mathbf{z}_{t-1} = (\mathbf{I} - \mathbf{P}) \cdot \hat{\mathbf{z}}_{t-1} + \hat{\mathbf{x}}, \quad \hat{\mathbf{x}} \sim q(\mathbf{z}_t|\mathbf{z}_0 = \mathbf{x}), \tag{36}$$

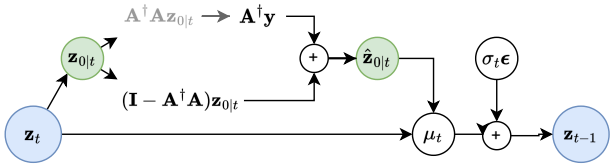

Fig. 4: Overview of DDNM [117]. It utilizes the range-null space decomposition for constructing a general solution for multiple tasks, such as image SR, colorization, inpainting, and deblurring.

where f, g depend on the type of DMs, $\mathbf{P}$ is the degradation process of the LR image, and $\hat{\mathbf{x}}$ is the LR image with the added and time-dependent noise.

5.5.2 Decomposition-Based

Decomposition-based methods view image SR tasks as a linear reverse problem similar to Equation 1:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + b, \tag{37}$$

where $\mathbf{A}$ is the degradation operator and b contaminating noise. Among the earliest decomposition-based methods, we find SNIPS [114] and its subsequent work DDRM [115]. These methods employ diffusion in the spectral domain, enhancing SR outcomes. To achieve this, they apply singular value decomposition to the degradation operator $\mathbf{A}$, thereby facilitating a spectral-domain transformation that contributes to their improved SR results.

The Denoising Diffusion Null-space Model (DDNM) represents another decomposition-based zero-shot approach applicable to a broad range of linear IR problems [117] beyond image SR to tasks like colorization, inpainting, and deblurring [117]. It leverages the range-null space decomposition methodology [124], [125] to tackle diverse IR challenges effectively. DDNM approaches the problem by reconfiguring Equation 1 as a linear reverse problem, although it is essential to note that this approach differs from SNIPS and DDRM in that it operates in a noiseless context:

$$\mathbf{x} = \mathbf{A}\mathbf{y}, \tag{38}$$

with $\mathbf{y} \in \mathbb{R}^{D\times 1}$ as the linearized HR image and $\mathbf{x} \in \mathbb{R}^{d\times 1}$ the linearized degraded image. Furthermore, it has to conform to the following two constraints:

$$\textit{Consistency}: \quad \mathbf{A}\hat{\mathbf{y}} \equiv \mathbf{x}, \qquad \textit{Realness}: \quad \hat{\mathbf{y}} \sim p(\mathbf{y}), \tag{39}$$

with $p(\mathbf{y})$ as the distribution of ground-truth images and $\hat{\mathbf{y}}$ the predicted image. The range-null space decomposition allows constructing a general solution for $\hat{\mathbf{y}}$ in the form of:

$$\hat{\mathbf{y}} = \mathbf{A}^\dagger\mathbf{x} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\bar{\mathbf{y}}, \tag{40}$$

with $\mathbf{A}^\dagger \in \mathbb{R}^{D\times d}$ the pseudo-inverse that satisfies $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} \equiv \mathbf{A}$. Our goal is to find a proper $\bar{\mathbf{y}}$ that generates the null-space $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\bar{\mathbf{y}}$ and agrees with the range-space $\mathbf{A}^\dagger\mathbf{x}$ that also fulfills realness in Equation 39.

DDNM derives clean intermediate states, denoted as $\mathbf{z}_{0|t}$, for the range-null space decomposition from $\mathbf{z}_0$ at timestep t. This is achieved through the equation:

$$\mathbf{z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{z}_t - \epsilon_\theta(\mathbf{z}_t, t)\sqrt{1-\bar{\alpha}_t}\right) \tag{41}$$

TABLE 4: Comparison of zero-shot methods. Data in bold represents the best performance. Second-best is underlined. Values derived from Li *et al.* [113].

Methods	ImageNet 1K			CelebA 1K			Time	Flops
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	[s/image]	[G]
Bicubic	25.36	0.643	0.27	24.26	0.628	0.34	-	-
ILVR [96]	<u>27.40</u>	**0.871**	0.21	<u>31.59</u>	0.878	0.22	41.3	1113.75
SNIPS [114]	24.31	0.684	0.21	27.34	0.675	0.27	31.4	-
DDRM [115]	27.38	<u>0.869</u>	0.22	**31.64**	**0.946**	0.19	<u>10.1</u>	1113.75
DPS [116]	25.88	0.814	<u>0.15</u>	29.65	0.878	0.18	141.2	1113.75
DDNM [117]	**27.46**	**0.871**	<u>0.15</u>	**31.64**	<u>0.945</u>	**0.16**	15.5	1113.75
GDP [118]	26.51	0.832	**0.14**	28.65	0.876	<u>0.17</u>	**3.1**	1113.76

with $\epsilon_t = \epsilon_\theta(\mathbf{z}_t, t)$. To produce a $\mathbf{z}_0$ that fulfills the equation $\mathbf{A}\mathbf{z}_0 \equiv \mathbf{x}$, the model leaves the null-space unaltered while setting the range-space as $\mathbf{A}^\dagger\mathbf{y}$. This generates a rectified estimation, $\hat{\mathbf{z}}_{0|t}$, defined by:

$$\hat{\mathbf{z}}_{0|t} = \mathbf{A}^\dagger\mathbf{x} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{z}_{0|t}. \tag{42}$$

Finally, $\mathbf{z}_{t-1}$ is derived by sampling from $p(\mathbf{z}_{t-1}|\mathbf{z}_t, \hat{\mathbf{z}}_{0|t})$:

$$\mathbf{z}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{z}}_{0|t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{z}_t + \sigma_t\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{43}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$, illustrated in Figure 4.

The term $\mathbf{z}_{t-1}$ represents a noised version of $\hat{\mathbf{z}}_{0|t}$. This noise effectively mitigates the dissonance between the range-space contents, represented by $\mathbf{A}^\dagger\mathbf{x}$, and the null-space contents, denoted by $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{z}_{0|t}$. The authors of DDNM show additionally that $\hat{\mathbf{z}}_{0|t}$ conforms to consistency.

The last step involves defining $\mathbf{A}$ and $\mathbf{A}^\dagger$, the construction of which is contingent on the restoration task at hand. For instance, in SR tasks involving scaling by a factor of n, $\mathbf{A}$ can be defined as a $1 \times n^2$ matrix, representative of an average-pooling operator. The average-pooling operator, denoted as $\begin{bmatrix}\frac{1}{n^2} & \dots & \frac{1}{n^2}\end{bmatrix}$, functions to average each patch into a singular value. Similarly, we can construct its pseudo-inverse as $\mathbf{A}^\dagger \in \mathbb{R}^{n^2 \times 1} = \begin{bmatrix}1 & \dots & 1\end{bmatrix}^\top$. The original work provides further examples of tasks (such as colorization, inpainting, and restoration), illustrating how these methods are applied. In addition, it describes how compound operations consisting of numerous sub-operations function in these contexts. In their research, the authors also introduced DDNM$^+$ to support the restoration of noisy images. They utilized a technique analogous to the "back and forward" strategy implemented in RePaint [120]. This approach was leveraged to enhance the quality further.

Given this approach's novelty, only a handful of subsequent studies extend and build upon it, such as the work presented in CDPMSR [97]. This research direction promises exciting possibilities for future exploration, although it calls for further investigation. For example, it should be noted that the DDNM approach introduces additional computational expenses compared to the task-specific training carried out using DDPMs. Moreover, the degradation operator $\mathbf{A}$ is set manually, which can be challenging for certain tasks. Another potential drawback is the assumption that $\mathbf{A}$ functions as a linear degradation operator, which may not always hold true and thus could limit the model's effectiveness in certain scenarios.

5.5.3 Posterior Estimation

Most projection-based methods typically address the noiseless inverse problem. However, this assumption can weaken data consistency because the projection process can deviate the sample path from the data manifold [113]. To address this and enhance data consistency, some recent works [116], [126], [127] take a different approach by aiming to estimate the posterior distribution using the Bayes theorem:

$$p(\mathbf{z}_t \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z}_t) \cdot p(\mathbf{z}_t)}{p(\mathbf{x})}, \tag{44}$$

This Bayesian approach provides a more robust and probabilistic framework for solving inverse problems, ultimately improving results in various image processing tasks. It results in the corresponding score function:

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{x}) = \nabla_{\mathbf{z}_t} \log p_t(\mathbf{x} \mid \mathbf{z}_t) + s_\theta(\mathbf{x}, t), \tag{45}$$

where $s_\theta(\mathbf{x}, t)$ is extracted from a pre-trained model while $p_t(\mathbf{x}|\mathbf{z}_t)$ is intractable. Thus, the goal is precisely estimating $p_t(\mathbf{x}|\mathbf{z}_t)$. MCG [126] and DPS [116] approximate the posterior $p_t(\mathbf{x}|\mathbf{z}_t)$ with $p_t(\mathbf{x}|\hat{\mathbf{z}}_0(\mathbf{z}_t))$, where $\hat{\mathbf{z}}_0(\mathbf{z}_t)$ is the expectation given $\mathbf{z}_t$ as $\hat{\mathbf{z}}_0(\mathbf{z}_t) = \mathbb{E}[\mathbf{z}_0|\mathbf{z}_t]$ according to Tweedie's formula [116]. While MCG also relies on projection, which can be harmful to data consistency, DPS discards the projection step and estimates the posterior as:

$$\begin{aligned}\nabla_{\mathbf{z}_t} \log p_t(\mathbf{x} \mid \mathbf{z}_t) &\approx \nabla_{\mathbf{z}_t} \log p(\mathbf{x} \mid \hat{\mathbf{z}}_0(\mathbf{z}_t)) \\ &\approx -\frac{1}{\sigma^2}\nabla_{\mathbf{z}_t}\|\mathbf{x} - H(\hat{\mathbf{z}}_0(\mathbf{z}_t))\|_2^2,\end{aligned} \tag{46}$$

where H is a forward measurement operator. A further expansion of this formula to the unified form for the linear, non-linear, differentiable inverse problem with Moore Penrose pseudoinverse can be found in IIGDM [127].

A different approach to estimate $p_t(\mathbf{x}|\mathbf{z}_t)$ is demonstrated by GDP [118]. The authors noted that a higher conditional probability of $p_t(\mathbf{x}|\mathbf{z}_t)$ correlates with a smaller distance between the application of the degradation model $\mathcal{D}(\mathbf{z}_t)$ and $\mathbf{x}$. Thus, they propose a heuristic approximation:

$$p_t(\mathbf{x}|\mathbf{z}_t) \approx \frac{1}{Z}\exp(-[s\mathcal{L}(\mathcal{D}(\mathbf{z}_t), \mathbf{x})]) + \lambda\mathcal{Q}(\mathbf{z}_t), \tag{47}$$

where $\mathcal{L}$ and $\mathcal{Q}$ denote a distance and quality metric, respectively. The term Z is for normalization, and s is a scaling factor controlling the guidance weight. However, due to varying noise levels between $\mathbf{z}_t$ and $\mathbf{x}$, precisely defining the distance metric $\mathcal{L}$ can be challenging. To overcome this challenge, GDP substitutes $\mathbf{z}_t$ with its clean estimation $\hat{\mathbf{z}}_0$ in the distance calculation, providing a pragmatic solution to the noise discrepancy issue.

HR SR3

Fig. 5: Example of color shifting produced by vanilla SR3 in a $64 \times 64 \rightarrow 256 \times 256$ setting when trained with reduced batch size (8 instead of 256).

5.6 Corruption Space

Karras *et al.* [128] identified three pillars of DMs: the noise schedule, the network parameterization, and the sampling algorithm. Recently, many authors argued to consider also different types of corruption used during forward diffusion like Soft Score Matching [129]. It directly incorporates the filtering process within the SGM, training the model to predict a clean image. Upon corruption, this predicted image aligns with the diffused observation.

Cold Diffusion [130] presents another ingenious way of modifying the corruption space for DDPMs, which shows that the generative capability is not strongly dependent on the choice of image degradation. It reveals new experimental types of diffusion besides Gaussian noise can be effectively used, like animorphosis (i.e., human faces iteratively degrading to animal faces).

The Image-to-Image Schrödinger Bridge (I^2SB) goes in a similar direction but does not impose any assumptions on the underlying prior distributions [131]. In its diffusion process, the clean image represents the initial state, while the degraded image is the final state in both forward and backward diffusion processes. This is notable for its ability to provide a transparent and traceable path from a degraded image to its clean version. Consequently, it provides enhanced interpretability since the process between degraded and clean images is directly addressed, which is not commonly present in many DMs. Another benefit is its higher efficiency in backward diffusion since it requires fewer steps (often between 2 and 10) to achieve comparable performance. Its conditionality, however, limits its use specifically to paired data during training, which is unsuitable for unsupervised SR.

While Cold Diffusion and I^2SB show promising results for image restoration, an extensive and more detailed quantitative analysis of different corruption types for image SR remains an exciting and open research avenue.

5.7 Color Shifting

As a result of high computational costs, DMs can occasionally suffer from color shifting when limited hardware necessitates smaller batch sizes or shorter learning periods [132]. An example with SR3 is shown in Figure 5.

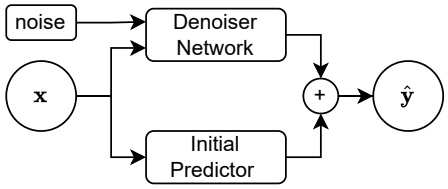

Fig. 6: Illustration of an initial predictor like in Whang *et al.* [89]. It divides the computation into a deterministic path (initial predictor) and a stochastic part (denoiser network), which predicts the missing residual information.

As presented by StableSR, a straightforward modification can address this issue by performing color normalization by adjusting the mean and variance with those of the LR image on the generated image [99]. Mathematically, it gives the following equation:

$$\hat{\mathbf{z}}_0 = \frac{\mathbf{z}_0^c - \mu_{\mathbf{z}_0}^c}{\sigma_{\mathbf{z}_0}^c} \cdot \sigma_{\mathbf{x}}^c + \mu_{\mathbf{x}}^c, \tag{48}$$

where $c \in \{r, g, b\}$ denotes the color channel, and $\mu_{\mathbf{z}_0}^c$ and $\sigma_{\mathbf{z}_0}^c$ (or $\mu_{\mathbf{x}}^c$ and $\sigma_{\mathbf{x}}^c$) are the mean and standard variance from the c-th channel of the predicted image $\mathbf{z}_0$ (or the input image $\mathbf{x}$), respectively.

You Only Diffuse Areas (YODA) [121], which targets diffusion on important image areas more frequently through time-dependent masks generated with DINO [122], also mitigates the color shift effect for image SR. This suggests that properly defined architecture and diffusion design are crucial to omit this effect. Further analysis of why this effect emerges must be obtained in future work.

5.8 Architecture Designs for Denoising

The design of the denoising model in DMs offers a range of options. The majority of DMs adopt the use of U-Net, as noted in most literature [110]. SR3 [42], for instance, employs residual blocks from BigGAN [110] and re-scales skip connections by a factor of $\frac{1}{\sqrt{2}}$. SRDiff takes a similar approach [74], although it opts for vanilla residual blocks without the re-scaling of skip connections and uses a LR encoder to incorporate the information of the LR image during the backward diffusion.

Whang *et al.* [89] exploit an initial predictor to combine the strengths of deterministic image SR models and DMs, as shown in Figure 6. It has the advantage that the DM only needs to learn the residuals that the deterministic image SR model (initial predictor) fails to predict, hence simplifying the learning target. Additionally, the removal of self-attention, positional encodings and group normalization from the SR3 U-Net enables their model to support arbitrary resolutions.

An initial predictor is also employed in the wavelet-based approach DiWa [83]. Moreover, wavelet SR models, such as DWSR [85] – a simple sequence of convolution layers of depth 10 – are utilized for denoising prediction in the wavelet domain. In WaveDM [86], a deterministic U-Net predictor is used for the high-frequency band, while diffusion is applied in the low-frequency band.

Latent Diffusion Models proposed by Rombach *et al.* [75] use a VQ-GAN [11] autoencoder in the latent space. For DiffIR [133], multiple variations of state-of-the-art Vision Transformers are employed [107], [134], [135]. Another common practice is pre-training deterministic components, as seen in models like SRDiff [74] or DiffIR [133]. Overall, the potential ways to design a denoising network are infinite, generally drawing inspiration from advancements made in general computer vision. The optimal denoising networks will vary based on the task, and the development of new models is anticipated.

6 Domain-Specific Applications

SR3 [42] produces photo-realistic and perceptually state-of-the-art images on faces and natural images but may not be suitable for other tasks like remote sensing. Some models are more suited to certain tasks as they tackle issues specific to the domain. This section highlights the applications of DMs to domain-specific SR tasks: Medical imaging, special cases of face SR (Blind Face Restoration and Atmospheric Turbulences), and remote sensing.

6.1 Medical Imaging

Magnetic Resonance Imaging (MRI) scans are widely used to aid patient diagnosis but can often be of low quality and corrupted with noise. Chung *et al.* [136] propose a combined denoising and SR network referred to as R2D2+ (Regularized Reverse Diffusion Denoiser + SR). They perform denoising of the MRI scans, followed by an SR module. Inspired by CCDF (i.e., a zero-shot method) from Chung *et al.* [123], they start their backward diffusion from an initial noisy image instead of pure Gaussian noise. The reverse SDE is solved using a non-parametric, eigenvalue-based method. In addition, they restrict the stochasticity of the DMs through low-frequency regularization. Particularly, they maintain low-frequency information while correcting the high-frequency ones to produce sharp and super-resolved MRI scans.

Mao *et al.* [137] addresses the lack of diffusion-based multi-contrast MRI SR methods. They propose a Disentangled Conditional Diffusion model (DisC-Diff) to leverage a multi-conditional fusion strategy based on representation disentanglement, enabling high-quality HR image sampling. Specifically, they employ a disentangled U-Net with multiple encoders to extract latent representations and use a novel joint disentanglement and Charbonnier loss function to learn representations across MRI contrasts. They also implement curriculum learning and improve their MRI model for varying anatomical complexity by gradually increasing the difficulty of training images.

6.2 Blind Face Restoration

Most previously discussed SR methods are founded on a fixed degradation process during training, such as bicubic downsampling. However, when applied practically, these assumptions frequently diverge from the actual degradation process and yield subpar results. Additionally, datasets with pairs of clean and real-world distorted images are usually unavailable. This issue is particularly researched in face SR, termed Blind Face Restoration (BFR), where datasets typically contain supervised samples $(\mathbf{x}, \mathbf{y})$ with an unknown degradation relationship.

A solution to BFR was proposed by Zongsheng *et al.* with DifFace [138] that leverages the rich generative priors of pre-trained DMs with parameters θ, which were trained to approximate $p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1})$. In contrast to existing methods that learn direct mappings from $\mathbf{x}$ to $\mathbf{y}$ under several constraints [139], [140], DifFace circumvents this by generating a diffused version $\mathbf{z}_N$ of the desired HR image $\mathbf{y}$ with $N < T$. They predict the starting point, the posterior $q(\mathbf{z}_N|\mathbf{x})$ via a transition distribution $p(\mathbf{z}_N|\mathbf{x})$. The transition distribution is formulated like the regular diffusion process, a Gaussian distribution, but uses an initial predictor $\varphi(\mathbf{x})$ to generate the mean, named diffused estimator. As their model borrows the reverse Markov chain from a pre-trained DM, DifFace requires no full retraining for new and unknown degradation mappings, unlike SR3.

A concurrent and better performing approach is DiffBFR [141] that adopts a two-step approach to BFR: A Identity Restoration Module (IRM), which employs two conditional DDPMs, and a Texture Enhancement Module (TEM), which employs an unconditional DDPM. In the first step within the IRM, a conditional DDPM enriches facial details at a low-resolution space same as $\mathbf{x}$. The downsampled version of $\mathbf{y}$ gives the target objective. Next, it resizes the output to the desired spatial size of $\mathbf{y}$ and applies another conditional DDPM to approximate the HR image $\mathbf{y}$. To ensure minimal deviation from the actual image, DiffBFR employs a novel truncated sampling method, which begins denoising at intermediate steps. The TEM further enhances realism through image texture and sharpened facial details. It imposes a diffuse-base facial prior with an unconditional DM trained on HR images and a backward diffusion starting from pure noise. DiffBFR produces high-quality, high-fidelity images close to the ground truth. The authors identify that DiffBFR has a much larger parameter scale than SR3 and requires optimization to accelerate sampling.

Another approach is DR2E presented by Wang *et al.* [142]. It also employs two stages: degradation removal and enhancement modules. For degradation removal, they use a pre-trained face SR DDPM to remove degradations from an LR image with severe and unknown degradations. In particular, they diffuse the degraded image $\mathbf{x}$ in T time steps to obtain $\mathbf{x}_T = \mathbf{z}_T$. Then, they use $\mathbf{x}_t$ to guide the backward diffusion such that the low-frequency part of $\mathbf{z}_t$ is replaced with that of $\mathbf{x}_t$, which are close in distribution. Theoretically, it produces visually clean intermediate results that are degradation-invariant. In the second stage, the enhancement module $p_\theta(y \mid \mathbf{z}_0)$, an arbitrary backbone CNN trained to map LR images to HR using a simple L2 loss, predicts the final output. DR2E can be slower than existing diffusion-based SR models for images with slight degradations and can even remove details from the input image.

6.3 Atmospheric Turbulence in Face SR

Atmospheric Turbulence (AT) results from atmospheric conditions fluctuations, leading to images' perceptual degradation through geometric distortions, spatially variant blur, and noise. These alterations negatively impact downstream vision tasks, such as tracking or detection.

Wang *et al.* [143] introduced a variational inference framework known as AT-VarDiff, which aims to correct AT in generic scenes. The distinctive feature of this approach is its reliance on a conditioning signal derived from latent task-specific prior information extracted from the input image to guide the DM.

Nair *et al.* [144] put forth another technique to restore facial images impaired by AT using SR. The method transfers class prior information from an SR model trained on clean facial data to a model designed to counteract turbulence degradation via knowledge distillation. The final model operates within the realistic faces manifold, which allows it to generate realistic face outputs even under substantial distortions. During inference, the process begins with noise- and turbulence-degraded images to ensure that the restored images closely resemble the distorted ones.

6.4 Remote Sensing

Remote Sensing Super-Resolution (RSSR) addresses the HR reconstruction from one or more LR images to aid object detection and semantic segmentation tasks for satellite imagery. RSSR is limited by the absence of small targets with complex granularity in the HR images. To produce finer details and texture, Liu *et al.* [145] present DMs with a Detail Complement mechanism (DMDC). They train their model similar to SR3 [42] and perform a detailed supplement task. To generate high-frequency information, they randomly mask several parts of the images to mimic dense objects. The SR images recover the occluded patches as the model learns small-grained information. Additionally, they introduce a novel pixel constraint loss to limit the diversity of DMDC and improve overall accuracy.

Ali *et al.* [146] design a new architecture for RS images that integrates Vision Transformers (ViT) with DMs as a Two-stage approach for Enhancement and Super-Resolution (TESR). In the first stage (SR stage), the SwinIR [107] model is used for RSSR. In the second stage (enhancement stage), the noisy images are enhanced by employing DMs to reconstruct the finer details.

Xu *et al.* [147] propose a blind SR framework based on Dual conditioning DDPMs for SR (DDSR). A kernel predictor conditioned on LR image encodings estimates the degradation kernel in the first stage. This is followed by an SR module consisting of a conditional DDPM in a U-Net with the predicted kernel and the LR encodings as guidance. An RRDB encoder extracts the encodings from LR images.

7 Discussion and Future Work

Though relatively new, DMs are quickly becoming a promising research area, especially in image SR. There are several avenues of ongoing research in this field, aiming to enhance the efficiency of DMs, accelerate computation speeds, and minimize memory footprint, all while generating high-quality, high-fidelity images. This section introduces common problems of DMs for image SR and examines noteworthy research avenues for DMs specific to image SR.

7.1 Color Shifting

Often, the most practical advancements come from a solid theoretical understanding. As discussed in Section 5.7, due to the substantial computational demands, DMs may occasionally exhibit color shifts when constrained by hardware limitations that demand smaller batch sizes or shorter training periods [132]. While well-defined diffusion methods [121] or color normalization [99] might mitigate this problem, a theoretical understanding of why this problem emerges is necessary.

7.2 Computational Costs

In a study conducted by Ganguli *et al.*, it was observed that the computing power needed for large-scale AI experiments has surged by over 300,000 times in the last decade [148]. Regrettably, this increase in resource intensity has been accompanied by a sharp decline in the share of these results originating from academic circles.

DMs are not immune to this issue; their computational demands add to the expanding gap between industry and academia. Therefore, there is a pressing need to reduce computational costs and memory footprints for practical applicability and research.

One strategy to alleviate computational demands is to examine smaller spatial-sized domains, as discussed in Section 5.2. Examples of such approaches include LDMs [11], [75] and wavelet-based models [83], [87]. However, the capability of LDMs to reconstruct data with high precision and fine-grained accuracy, as required in image SR, remains to be questioned. Therefore, further advancements in these methods for image SR are critically needed. On the other hand, wavelet-based models do not present a bottleneck regarding information preservation. This advantage suggests that they should be the subject of more intensive exploration.

7.3 Efficient Sampling

A defining characteristic and benefit of DMs is the possibility of decoupling training and inference schedules [149]. This decoupling allows for substantial enhancements in terms of curtailing the time required for inference in practical applications, providing a significant efficiency edge in real-world scenarios. While reducing the number of steps taken during inference is relatively simple, a systematic method for determining inference schedules for image SR has yet to be developed [150]. As outlined in Section 4.1, this research direction represents a promising avenue for future exploration.

An alternative for more efficient sampling is given by methods that use different corruption spaces, as discussed in Section 5.6. Unlike sampling from pure Gaussian noise, notable works such as Luo *et al.* [78], I^2SB [131], Come-Closer-Diffuse-Faster [123], or Cold Diffusion [130] define a process from the LR to the HR image directly.

Further, additional techniques for decreasing computation time, such as knowledge distillation, alternative noise schedulers, or truncated diffusion, demand further investigation concerning image SR [61], [62], [151], [152].

7.4 Comparability

The comparison of different DMs in image SR is challenging due to the use of varied datasets across studies. These datasets differ in aspects like resolution, content diversity, color distribution, and noise levels, affecting a model's performance. DM might excel with one dataset but not another, making it difficult to gauge its overall effectiveness. Thus, the performance of a DM on a specific dataset does not necessarily reflect its general effectiveness.

To address this comparability issue, it is crucial to establish a standard benchmark consisting of diverse, representative datasets and a standardized set of evaluation metrics, not only for DM-based methods. This benchmark should cover various degradation types and levels, allowing a fair and comprehensive comparison of DMs in image SR. A well-structured benchmark can facilitate the identification of models that perform well across various tasks and conditions, thus accelerating the advancement of the field.

Also, assessing the quality of SR images produced by generative models remains a challenging and unresolved issue. While DMs underperform on standard metrics (PSNR and SSIM), the images generated are more photo-realistic, unlike the results obtained from classical image SR works [14]. Consequently, generative models like DMs tend to show lower scores using these metrics compared to pixel-based methods, despite often being rated more favorably by human evaluators [42]. LPIPS [153] performs better reflecting this perception, but the domain of image SR has to adapt to more diverse metrics, such as predictors that reflect human ratings directly [154], [155]. For instance, datasets with subjective ratings, like TID2013 [156], and neural networks, such as DeepQA [157] or NIMA [158], can be employed to predict human-like scoring of images and should be further explored in the context of diffusion generated SR images.

7.5 Image Manipulation

Li *et al.* [74], the authors of SRDiff, proposed two potential extensions of SR DMs: content fusion and latent space interpolation. Content fusion involves the combination of content from two source images. For instance, they replace the eyes in one source image with the face from another image before conducting diffusion in the image space like CutMix [159]. The backward diffusion successfully creates a smooth transition between both images. In the latent space interpolation model, the latent space of two SR predictions is linearly interpolated to generate a new image. While these extensions have yielded remarkable results, unlike other generative models such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs), DMs have been found to offer less proficient latent representations [160]. Therefore, recent and ongoing research into the manipulation of latent representations in diffusion models is both in its early stages and greatly needed [161], [162], [163]. This can be particularly useful in multi-image SR for generating high-resolution images that blend characteristics from multiple sources, potentially improving the quality and diversity of the output (e.g., satellite imagery for SR predictions with flexible daylights).

7.6 Cascaded Image Generation

Saharia *et al.* [42] presented the concept of cascaded image generation, in which multiple DDPMs are chained together across different scales. This strategy was applied to unconditional and class-conditional generation, cascading a model synthesizing 64×64 images with SR3 models generating 1024×1024 unconditional faces and 256×256 class-conditional natural images. The cascading approach allows several simpler models to be trained simultaneously, improving computational efficiency due to faster training times and reduced parameter counts. Furthermore, they implemented cascading for inference, using more refinement steps at lower and fewer steps at higher resolutions. They found this more efficient than generating SR images directly while achieving satisfactory FID scores. Even though their model's underperformance compared to BigGAN [110] concerning cascaded generation, the approach still represents an exciting research opportunity.

8 Conclusion

In image Super-Resolution (SR), integrating Diffusion Models (DMs) has ushered in a new era of innovation. DMs, as exemplified by SR3, have the potential to bridge the gap between technical image quality and human perceptual preferences. Unlike traditional SR methods, which often focus solely on pixel-level accuracy, DMs have the unique ability to generate images that are high in resolution, aesthetically pleasing, and realistic to the human eye. In contrast to previous generative models, they do not suffer typical training convergence issues like mode collapse.

Throughout this survey, we have explored the remarkable progress and diverse methodologies that have propelled SR using DMs to the forefront of the field. We began by establishing the foundational principles of DMs, highlighting their relationship to other generative approaches, elucidating their core concepts and their application to image SR. The survey unveiled the myriad of conditioning strategies, from LR image guidance to text embeddings and pre-trained networks, showcasing their crucial role in enhancing image quality. Zero-shot SR, a particularly intriguing paradigm, was also a subject, as well as corruption spaces and image SR-specific topics like color shifting and architectural designs.

In conclusion, this survey unveils the profound impact of DMs on image SR. It provides a comprehensive guide to the current landscape and valuable insights into trends, challenges, and future directions. DMs do not just close a significant portion of the gap in SR technology; they set a new benchmark, pushing the boundaries of image quality and realism. As we continue to explore and refine these models, the future of image SR looks more promising than ever, with potential applications, as discussed in our applications section, that extend far beyond what was previously imagined. For future research, this survey highlights important issues, such as color shifting, computational costs, efficient sampling, and comparability, while also discussing interesting avenues like image manipulation and cascaded image generation.

ACKNOWLEDGMENTS

This work was supported by the BMBF project XAINES (Grant 01IW20005) and SustainML (Horizon Europe grant agreement No 101070408).

REFERENCES

[1] W. Sun and Z. Chen, "Learned image downscaling for upscaling using content adaptive resampler," *IEEE TIP*, vol. 29, pp. 4027–4040, 2020.

[2] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[3] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, vol. 2. IEEE, 2001, pp. 416–423.

[4] D. Valsesia and E. Magli, "Permutation invariance and uncertainty in multitemporal image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[5] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu, "A comprehensive review of deep learning-based single image super-resolution," *PeerJ Computer Science*, vol. 7, p. e621, 2021.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[8] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *NeurIPS*, vol. 32, 2019.

[9] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[10] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *CVPR*, 2022, pp. 11 523–11 532.

[11] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021, pp. 12 873–12 883.

[12] B. Guo, X. Zhang, H. Wu, Y. Wang, Y. Zhang, and Y.-F. Wang, "Lar-sr: A local autoregressive model for image super-resolution," in *CVPR*, 2022, pp. 1909–1918.

[13] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review," *Neural Networks*, vol. 144, pp. 187–209, 2021.

[14] B. B. Moser, F. Raue, S. Frolov, S. Palacio, J. Hees, and A. Dengel, "Hitchhiker's guide to super-resolution: Introduction and recent advances," *IEEE TPAMI*, pp. 1–21, 2023.

[15] *MATLAB version 9.3.0.713579 (R2017b)*, The Mathworks, Inc., Natick, Massachusetts, 2017.

[16] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *CVPRW*, 2017, pp. 1122–1131.

[17] ——, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *CVPRW*, July 2017.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[21] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[22] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE TPAMI*, vol. 32, no. 6, pp. 1127–1133, 2010.

[23] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, no. 2, apr 2011.

[24] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *CVPR*, 2008, pp. 1–8.

[25] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1. IEEE, 2004, pp. I–I.

[26] W. Freeman, T. Jones, and E. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[27] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[28] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Models and Image Processing*, vol. 53, no. 3, pp. 231–239, 1991.

[29] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, vol. 19, no. 11, pp. 2861–2873, 2010.

[30] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE TPAMI*, vol. 38, no. 2, pp. 295–307, 2015.

[31] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV*. Springer, 2016, pp. 391–407.

[32] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016, pp. 1874–1883.

[33] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[35] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *ICCV*, 2017, pp. 4799–4807.

[36] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018.

[37] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *CVPR*, 2016, pp. 1637–1645.

[38] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *CVPR*, 2017, pp. 3147–3155.

[39] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *ECCV*, 2018, pp. 252–268.

[40] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.

[41] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Srflow: Learning the super-resolution space with normalizing flow," in *ECCV*. Springer, 2020.

[42] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE TPAMI*, vol. 45, no. 4, pp. 4713–4726, 2023.

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

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[45] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[46] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*. PMLR, 2015, pp. 2256–2265.

[47] G. Parisi, "Correlation functions and computer simulations," *Nuclear Physics B*, vol. 180, no. 3, pp. 378–384, 1981.

[48] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[49] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[50] Y. Song, S. Garg, J. Shi, and S. Ermon, "Sliced score matching: A scalable approach to density and score estimation," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 574–584.

[51] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[52] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[54] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[55] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *ICML*. PMLR, 2015, pp. 1530–1538.

[56] Q. Zhang and Y. Chen, "Diffusion normalizing flow," in *NeurIPS*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 16 280–16 291.

[57] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 2617–2680, 2021.

[58] "Oxford vggface implementation using keras functional framework v2+," https://github.com/rcmalli/keras-vggface.

[59] D. Watson, J. Ho, M. Norouzi, and W. Chan, "Learning to efficiently sample from diffusion probabilistic models," *arXiv preprint arXiv:2106.03802*, 2021.

[60] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," *arXiv preprint arXiv:2202.05830*, 2022.

[61] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," in *CVPR*, 2023, pp. 14 297–14 306.

[62] E. Luhman and T. Luhman, "Knowledge distillation in iterative generative models for improved sampling speed," *arXiv preprint arXiv:2101.02388*, 2021.

[63] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[64] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *arXiv preprint arXiv:2112.07804*, 2021.

[65] Z. Lyu, X. Xu, C. Yang, D. Lin, and B. Dai, "Accelerating diffusion models via early stop of the diffusion process," *arXiv preprint arXiv:2205.12524*, 2022.

[66] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *CVPR*, 2023, pp. 3836–3847.

[67] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[68] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman, and I. Mitliagkas, "Gotta go fast when generating data with score-based models," *arXiv preprint arXiv:2105.14080*, 2021.

[69] J. Ho, E. Lohn, and P. Abbeel, "Compression with flows via local bits-back coding," *NeurIPS*, vol. 32, 2019.

[70] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good semi-supervised learning that requires a bad gan," *NeurIPS*, vol. 30, 2017.

[71] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," *NeurIPS*, vol. 34, pp. 1415–1428, 2021.

[72] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *NeurIPS*, vol. 34, pp. 21 696–21 707, 2021.

[73] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*. PMLR, 2021, pp. 8162–8171.

[74] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[75] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[76] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," *NeurIPS*, vol. 34, pp. 11 287–11 302, 2021.

[77] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *CVPR*, 2023, pp. 1680–1691.

[78] ——, "Image restoration with mean-reverting stochastic differential equations," *arXiv preprint arXiv:2301.11699*, 2023.

[79] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *ECCV*. Springer, 2022, pp. 17–33.

[80] Z. Chen, Y. Zhang, D. Liu, B. Xia, J. Gu, L. Kong, and X. Yuan, "Hierarchical integration diffusion model for realistic image deblurring," *arXiv preprint arXiv:2305.12966*, 2023.

[81] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *CVPR*, 2020, pp. 2437–2445.

[82] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *CVPR*, 2018, pp. 2492–2501.

[83] B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, "Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution," 2023.

[84] B. B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, "Dwa: Differential wavelet amplifier for image super-resolution," in *Artificial Neural Networks and Machine Learning – ICANN 2023*, L. Iliadis, A. Papaleonidas, P. Angelov, and C. Jayne, Eds. Cham: Springer Nature Switzerland, 2023, pp. 232–243.

[85] T. Guo, H. Seyed Mousavi, T. Huu Vu, and V. Monga, "Deep wavelet prediction for image super-resolution," in *CVPRW*, 2017, pp. 104–113.

[86] Y. Huang, J. Huang, J. Liu, Y. Dong, J. Lv, and S. Chen, "Wavedm: Wavelet-based diffusion models for image restoration," *arXiv preprint arXiv:2305.13819*, 2023.

[87] F. Guth, S. Coste, V. De Bortoli, and S. Mallat, "Wavelet score-based generative modeling," *NeurIPS*, vol. 35, pp. 478–491, 2022.

[88] S. Shang, Z. Shan, G. Liu, and J. Zhang, "Resdiff: Combining cnn and diffusion model for image super-resolution," *arXiv preprint arXiv:2303.08714*, 2023.

[89] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *CVPR*, 2022, pp. 16 293–16 303.

[90] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting," 2023.

[91] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017, pp. 136–144.

[92] S. H. Park, Y. S. Moon, and N. I. Cho, "Flexible style image super-resolution using conditional objective," *IEEE Access*, vol. 10, pp. 9774–9792, 2022.

[93] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," in *CVPR*, 2019, pp. 3096–3105.

[94] A. Lugmayr, M. Danelljan, L. V. Gool, and R. Timofte, "Srflow: Learning the super-resolution space with normalizing flow," in *ECCV*. Springer, 2020, pp. 715–732.

[95] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, vol. 35, pp. 36 479–36 494, 2022.

[96] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," 2021.

[97] A. Niu, K. Zhang, T. X. Pham, J. Sun, Y. Zhu, I. S. Kweon, and Y. Zhang, "Cdpmsr: Conditional diffusion probabilistic models for single image super-resolution," 2023.

[98] K. Pandey, A. Mukherjee, P. Rai, and A. Kumar, "Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents," *arXiv preprint arXiv:2201.00308*, 2022.

[99] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *arXiv preprint arXiv:2305.07015*, 2023.

[100] S. Zhou, K. Chan, C. Li, and C. C. Loy, "Towards robust blind face restoration with codebook lookup transformer," *NeurIPS*, vol. 35, pp. 30 599–30 611, 2022.

[101] T. Yang, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," *arXiv preprint arXiv:2308.14469*, 2023.

[102] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning

transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[103] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *ICCV*, 2021, pp. 4791–4800.

[104] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *CVPR*, 2021, pp. 1905–1914.

[105] J. Liang, H. Zeng, and L. Zhang, "Details or artifacts: A locally discriminative learning approach to realistic image super-resolution," in *CVPR*, 2022, pp. 5657–5666.

[106] C. Chen, X. Shi, Y. Qin, X. Li, X. Han, T. Yang, and S. Guo, "Real-world blind super-resolution via feature matching with implicit high-resolution priors," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1329–1338.

[107] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *CVPR*, 2021, pp. 1833–1844.

[108] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation." *J. Mach. Learn. Res.*, vol. 23, no. 47, pp. 1–33, 2022.

[109] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *NeurIPS*, vol. 34, pp. 8780–8794, 2021.

[110] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[111] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[112] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022.

[113] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, and Z. Chen, "Diffusion models for image restoration and enhancement–a comprehensive survey," *arXiv preprint arXiv:2308.09388*, 2023.

[114] B. Kawar, G. Vaksman, and M. Elad, "Snips: Solving noisy inverse problems stochastically," *NeurIPS*, vol. 34, pp. 21 757–21 769, 2021.

[115] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *NeurIPS*, vol. 35, pp. 23 593–23 606, 2022.

[116] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," *arXiv preprint arXiv:2209.14687*, 2022.

[117] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *arXiv preprint arXiv:2212.00490*, 2022.

[118] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *CVPR*, 2023, pp. 9935–9946.

[119] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *CVPR*, 2018, pp. 3118–3126.

[120] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *CVPR*, 2022, pp. 11 461–11 471.

[121] B. B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, "Yoda: You only diffuse areas. an area-masked diffusion approach for image super-resolution," *arXiv preprint arXiv:2308.07977*, 2023.

[122] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *CVPR*, 2021, pp. 9650–9660.

[123] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *CVPR*, 2022, pp. 12 413–12 422.

[124] J. Schwab, S. Antholzer, and M. Haltmeier, "Deep null space learning for inverse problems: convergence analysis and rates," *Inverse Problems*, vol. 35, no. 2, p. 025008, 2019.

[125] Y. Wang, Y. Hu, J. Yu, and J. Zhang, "Gan prior based null-space learning for consistent super-resolution," in *AAAI*, vol. 37, no. 3, 2023, pp. 2724–2732.

[126] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," *NeurIPS*, vol. 35, pp. 25 683–25 696, 2022.

[127] J. Song, A. Vahdat, M. Mardani, and J. Kautz, "Pseudoinverse-guided diffusion models for inverse problems," in *ICLR*, 2022.

[128] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, vol. 35, pp. 26 565–26 577, 2022.

[129] G. Daras, M. Delbracio, H. Talebi, A. G. Dimakis, and P. Milanfar, "Soft diffusion: Score matching for general corruptions," *arXiv preprint arXiv:2209.05442*, 2022.

[130] A. Bansal, E. Borgnia, H.-M. Chu, J. S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein, "Cold diffusion: Inverting arbitrary image transforms without noise," *arXiv preprint arXiv:2208.09392*, 2022.

[131] G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar, "I ^{2} sb: Image-to-image schrödinger bridge," *arXiv preprint arXiv:2302.05872*, 2023.

[132] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *CVPR*, 2022, pp. 11 472–11 481.

[133] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," *arXiv preprint arXiv:2303.09472*, 2023.

[134] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[135] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[136] H. Chung, E. S. Lee, and J. C. Ye, "Mr image denoising and super-resolution using regularized reverse diffusion," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 922–934, 2022.

[137] Y. Mao, L. Jiang, X. Chen, and C. Li, "Disc-diff: Disentangled conditional diffusion model for multi-contrast mri super-resolution," *arXiv preprint arXiv:2303.13933*, 2023.

[138] Z. Yue and C. C. Loy, "Difface: Blind face restoration with diffused error contraction," *arXiv preprint arXiv:2212.06512*, 2022.

[139] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *CVPR*, 2021, pp. 9168–9178.

[140] T. Yang, P. Ren, X. Xie, and L. Zhang, "Gan prior embedded network for blind face restoration in the wild," in *CVPR*, 2021, pp. 672–681.

[141] X. Qiu, C. Han, Z. Zhang, B. Li, T. Guo, and X. Nie, "Diffbfr: Bootstrapping diffusion model towards blind face restoration," *arXiv preprint arXiv:2305.04517*, 2023.

[142] Z. Wang, Z. Zhang, X. Zhang, H. Zheng, M. Zhou, Y. Zhang, and Y. Wang, "Dr2: Diffusion-based robust degradation remover for blind face restoration," in *CVPR*, 2023, pp. 1704–1713.

[143] X. Wang, S. López-Tapia, and A. K. Katsaggelos, "Atmospheric turbulence correction via variational deep diffusion," in *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2023, pp. 1–4.

[144] N. G. Nair, K. Mei, and V. M. Patel, "At-ddpm: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models," in *WACV*, 2023, pp. 3434–3443.

[145] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, and B. Lu, "Diffusion model with detail complement for super-resolution of remote sensing," *Remote Sensing*, vol. 14, no. 19, 2022.

[146] A. M. Ali, B. Benjdira, A. Koubaa, W. Boulila, and W. El-Shafai, "Tesr: Two-stage approach for enhancement and super-resolution of remote sensing images," *Remote Sensing*, vol. 15, no. 9, 2023.

[147] M. Xu, J. Ma, and Y. Zhu, "Dual-diffusion: Dual conditional denoising diffusion probabilistic models for blind super-resolution reconstruction in rsis," *arXiv preprint arXiv:2305.12170*, 2023.

[148] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage *et al.*, "Predictability and surprise in large generative models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1747–1764.

[149] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[150] Z. Cheng, "Sampler scheduler for diffusion models," *arXiv preprint arXiv:2311.06845*, 2023.

[151] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *arXiv preprint arXiv:2112.07804*, 2021.

[152] T. Chen, "On the importance of noise scheduling for diffusion models," *arXiv preprint arXiv:2301.10972*, 2023.

[153] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.
[154] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *ICCV*, 2017, pp. 1040–1049.
[155] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE TIP*, vol. 26, no. 8, pp. 3951–3964, 2017.
[156] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.
[157] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *CVPR*, 2017, pp. 1676–1684.
[158] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE TIP*, vol. 27, no. 8, pp. 3998–4011, 2018.
[159] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *CVPR*, 2019, pp. 6023–6032.
[160] B. Jing, G. Corso, R. Berlinghieri, and T. Jaakkola, "Subspace diffusion generative models," in *ECCV*. Springer, 2022, pp. 274–289.
[161] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," in *ICLR*, 2023.
[162] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang, "Uncovering the disentanglement capability in text-to-image diffusion models," in *CVPR*, 2023, pp. 1900–1910.
[163] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *CVPR*, 2022, pp. 2426–2435.

Brian B. Moser is a Ph.D. student at the TU Kaiserslautern and a research assistant at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern. He received the M.Sc. degree in computer science from the TU Kaiserslautern in 2021. His research interests include image super-resolution and deep learning.

Arundhati S. Shanbhag is a Master's student at the TU Kaiserslautern and research assistant at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern. Her research interests include computer vision and deep learning.

Federico Raue is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern. He received his Ph.D. at TU Kaiserslautern in 2018 and his M.Sc. in Artificial Intelligence from Katholieke Universiteit Leuven in 2005. His research interests include meta-learning and multimodal machine learning.

Stanislav Frolov is a Ph.D. student at the TU Kaiserslautern and a research assistant at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern. He received the M.Sc. degree in electrical engineering from the Karlsruhe Institute of Technology in 2017. His research interests include generative models and deep learning.

Sebastian Palacio is a researcher in machine learning and head of the multimedia analysis and data mining group at the German Research Center for Artificial Intelligence (DFKI). His Ph.D. topic was about explainable AI with applications in computer vision. Other research interests include adversarial attacks, multi-task, curriculum, and self-supervised learning.

Andreas Dengel is a Professor at the Department of Computer Science at TU Kaiserslautern and Executive Director of the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern, Head of the Smart Data and Knowledge Services research area at DFKI and of the DFKI Deep Learning Competence Center. His research focuses on machine learning, pattern recognition, quantified learning, data mining, semantic technologies, and document analysis.