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

Brian B. Moser, Arundhati S. Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel

Abstract— Diffusion models (DMs) have disrupted the image super-resolution (SR) field and further closed the gap between image quality and human perceptual preferences. They are easy to train and can produce very high-quality samples that exceed the realism of those produced by previous generative methods. Despite their promising results, they also come with new challenges that need further research: high computational demands, comparability, lack of explainability, color shifts, and more. Unfortunately, entry into this field is overwhelming because of the abundance of publications. To address this, we provide a unified recount of the theoretical foundations underlying DMs applied to image SR and offer a detailed analysis that underscores the unique characteristics and methodologies within this domain, distinct from broader existing reviews in the field. This article articulates a cohesive understanding of DM principles and explores current research avenues, including alternative input domains, conditioning techniques, guidance mechanisms, corruption spaces, and zero-shot learning approaches. By offering a detailed examination of the evolution and current trends in image SR through the lens of DMs, this article sheds light on the existing challenges and charts potential future directions, aiming to inspire further innovation in this rapidly advancing area.

Index Terms— Diffusion models (DMs), super-resolution (SR), survey.

I. INTRODUCTION

IN THE ever-evolving field of computer vision, the task of image super-resolution (SR)—enhancing low-resolution (LR) images into high-resolution (HR) counterparts—stands as a longstanding challenge due to its ill-posed nature. Multiple HR images are plausible for any given LR image, differing in aspects such as brightness and color [1]. Its applications span a broad spectrum, from everyday photography to refining satellite [2] and medical images [3]. Despite notable achievements of prior generative SR models, each comes with its own limitations. For example, the computational demands of autoregressive models (ARMs) often outweigh their utility, while normalizing flows (NFs) or variational autoencoders (VAEs) struggle to match quality expectations [4], [5], [6]. Although powerful, generative adversarial networks (GANs)

need careful regularization and optimization strategies to overcome instability issues [7].

The advent of diffusion models (DMs) marks a significant shift in image generation tasks, including SR, challenging the long-standing dominance of GANs [8], [9], [10], [11]. Applications like Dall-E and stable diffusion demonstrate that DMs have surpassed GANs in various aspects [12], [13], [14]. Their capability to generate high-quality images from LR inputs has shown immense promise in SR by closely aligning with the qualitative judgments of human evaluators [15]. In other words, human raters perceive SR images generated by DMs as more realistic than those produced by other generative models like GANs.

However, as the volume of publications expands, staying updated on the latest developments is becoming more challenging, particularly for those new to the field. DMs diverge fundamentally from prior generative models and pose new challenges while addressing the limitations of earlier models. Identifying coherent trends and potential research directions is challenging despite this rapid expansion. This article aims to demystify DMs, offers a comprehensive overview that bridges foundational concepts with the forefront of image SR, and critically analyzes current strengths and weaknesses.

This article builds upon the previous work *Hitchhiker's Guide to SR* [16], which gives a broad overview of the image SR field in general. Similar in spirit is the survey of Li et al. [17], which reviews DMs on the more general image restoration tasks like inpainting and dehazing. Both have overlapping topics, such as the foundations and types of DMs, namely, denoising diffusion probabilistic models (DDPMs) [8], score-based generative models (SGMs) [11], and stochastic differential equations (SDEs) [10]. Moreover, both surveys highlight the introduction of conditioning strategies and zero-shot diffusion as well as show potential research directions. However, this article covers all topics related to image SR and is, therefore, more detailed regarding recent developments specifically developed for image SR. Moreover, we explain SR-related challenges, such as color shifting and cascaded image SR. We also highlight the relationship of DMs with other generative SR models, namely, VAEs, GANs, and flow-based methods. In addition, we review frequency-based DMs, alternative corruption spaces, and diffusion-based image SR applications.

Concluding with a discussion on emerging trends and their potential for reshaping SR and DM development, this article sets the stage for future research. By offering clarity and direction in the rapidly evolving domain of DMs, we aim

to inspire and inform the next wave of research, fostering advancements that continue to push the boundaries of what is possible in image SR with DMs.

The structure of this article is organized as follows.

Section II—SR Basics: This section provides fundamental definitions and introduces standard datasets, methods, and metrics for assessing image quality commonly utilized in image SR publications.

Section III—DMs Basics: This section introduces the principles and various formulations of DMs, including DDPMs, SGMs, and SDEs. This section also explores how DMs relate to other generative models.

Section IV—Improvements for DMs: Common practices for enhancing DMs, focusing on efficient sampling techniques and improved likelihood estimation.

Section V—DMs for Image SR: This section 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 VII—Domain-Specific Applications: DM-based SR applications, namely, medical imaging, blind face restoration (BFR), atmospheric turbulence (AT) in face SR, and remote sensing.

Section VIII—Discussion and Future Work: Common problems of DMs for image SR and noteworthy research avenues for DMs specific to image SR.

Section IX—Conclusion: This section summarizes this article.

II. IMAGE SR

The goal of image SR is to transform one or more LR images into HR images. The domain can be broadly categorized into two areas [16]: single image super-resolution (SISR) and multi-image super-resolution (MISR). In SISR, a single LR image leads to a single HR image. In contrast, MISR methods use multiple LR images to produce one or many HR outputs. This section focuses on SISR and explores relevant datasets, established SR models, and techniques to assess image quality.

Given an LR image $\mathbf{x} \in \mathbb{R}^{\bar{w} \times \bar{h} \times c}$, the goal is to generate an HR counterpart $\mathbf{y} \in \mathbb{R}^{w \times h \times c}$ with $\bar{w} < w$ and $\bar{h} < h$. The relationship is represented by a degradation mapping

$$\mathbf{x} = \mathcal{D}(\mathbf{y}; \Theta) = ((\mathbf{y} \otimes \boldsymbol{k}) \downarrow_s + \boldsymbol{n})_{\text{JPEG}_q} \quad (1)$$

where $\mathcal{D}$ is a degradation map $\mathcal{D} : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}^{\bar{w} \times \bar{h} \times c}$ and Θ contains degradation parameters, including aspects such as blur $\boldsymbol{k}$, noise $\boldsymbol{n}$, scaling s, and compression quality q [18]. The degradation is typically unknown, posing the main challenge in determining the inverse mapping of $\mathcal{D}$ with parameters θ, usually embodied as SR model [19], [20]. It leads to an optimization task aimed at minimizing the difference between the predicted SR image $\hat{y}$ and the original HR image $\mathbf{y}$

$$\theta^* = \text{argmin}_\theta \mathcal{L}(\hat{y}, \mathbf{y}) + \lambda \phi(\theta) \quad (2)$$

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

The inherent complexity arises from the ill-posed nature of predicting θ, as several SR images can be valid for any given LR image, i.e., they can have similar loss values compared to the ground-truth image but are subjectively perceived differently due to many aspects such as brightness and coloring [1], [15], [21]. Traditional regression techniques, like standard convolutional neural networks (CNNs), are often adequate for lower magnifications but struggle to replicate high-frequency details required at higher magnifications (e.g., $s > 4$). To address this, SR models must hallucinate realistic details beyond interpolation, which typically falls under the umbrella topic of generative models, where DMs are now at the forefront.

A. Datasets

Several datasets offer a variety of images, resolutions, and content types. Typically, these datasets consist of LR and HR image pairs. However, some datasets contain only HR images, with LR images created by bicubic downsampling with antialiasing—a default setting for *imresize* in MATLAB [22]. One famous general SR train set is the Diverse 2K resolution (DIV2K) dataset [23], which includes various realistic images at different resolutions designed specifically for image SR. Classical test datasets for SR models trained on DIV2K are Set5 [24], Set14 [25], BSDS100 [26], Urban100 [27], and Manga109 [28] that cover a variety of scenes and images contents such as buildings and manga paintings. Flickr2K [23] and Flickr-Faces-HQ (FFHQ) [29] offer diverse sets of human-centric and scene-centric images from Flickr, respectively. While FFHQ is commonly employed for training models for face SR tasks, Flickr2K is usually used as a train data extension in combination with DIV2K. Another dataset for face SR is CelebA-HQ [30], which provides high-quality celebrity images and is typically used to evaluate FFHQ-trained SR models. For broader applications in CV, datasets such as ImageNet [31] and Visual Object Classes (VOC2012) [32] are favored. ImageNet offers an extensive range of images that help train models on various object classes, whereas VOC2012 is vital for object detection and segmentation. Both are valuable for multitask learning involving SR. More datasets can be found in the *Hitchhiker's Guide to SR* [16].

B. SR Models

The primary objective is to design an 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 (1)

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

where $\hat{y}$ is the predicted HR approximation of the LR image, and $\mathbf{x}$ and θ are the parameters of $\mathcal{M}$. The parameters θ are optimized using (2), i.e., minimizing the loss function $\mathcal{L}$ between the estimation $\hat{y}$ and the ground-truth HR image $\mathbf{y}$. Sections II-B1–II-B4 focuses on standard methods for designing an SR model, especially deep learning methods before we examine how DMs fulfill this role in detail.

1) Traditional Methods: Traditional methods for image SR define a range of methodologies, such as statistical [33], edge-based [34], [35], patch-based [36], [37], prediction-based [38], [39], and sparse representation techniques [40]. They fundamentally rely on image statistics and the information inherent in existing pixels to generate HR images. Despite their utility, a noteworthy drawback of these methods is the potential introduction of noise, blur, and visual artifacts [16].

2) Regression-Based Deep Learning: Image SR significantly evolved with advancements in deep learning and computational power. Typically, they employ a CNN for end-to-end mapping from LR to HR. Initial models, such as SRCNN [41], FSRCNN [42], and ESPCNN [43], utilized simple CNNs of diverse depth and feature maps sizes. Later models adapted concepts from the broader CV domain into SR models, e.g., ResNet led to SRResNet, where residual information was propagated to successive network layers [44]. Likewise, DenseNet [45] was adapted with SRDenseNet [46]. They employ dense blocks, where each layer receives additionally the features generated in all preceding layers. Recursive CNNs that recursively use the same module to learn representations were also inspired by other CV methods for regression-based SR methods in DRCN [47], DRRN [48], and CARN [49]. More recently, attention mechanisms have been incorporated to focus on regions of interest in images, predominantly via the channel and spatial attention mechanisms [16], [50], [51], [52]. All those methods have in common that they are regression-based. Commonly used loss functions are the L1 and L2 losses. As mentioned, they often produce satisfying results for lower magnifications but struggle to replicate the high-frequency details required at higher magnifications (e.g., $s > 4$). These limitations arise because these models primarily learn an averaged mapping (due to L1 and L2 losses) from LR to HR images, which tends to produce overly smooth textures lacking detail, especially noticeable in larger upscaling factors [16]. To address this, SR models must hallucinate realistic details beyond simple interpolation, a challenge typically tackled by generative models.

3) Generative Adversarial Networks: One of the most prominent generative models is the GAN. It uses 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 [44] or ESRGAN [53], 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 suffer from stabilization issues [7].

4) Flow-Based Methods: Flow-based methods employ optical flow algorithms to generate SR images [54]. 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 an 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 [54]. This facilitates flow-based methods to circumvent training instability but incurs a substantial computational cost.

C. Image Quality Assessment

Image quality is a multifaceted concept that addresses properties such as sharpness, contrast, and the absence of noise. Hence, a fair evaluation of SR models based on produced image quality forms a nontrivial task. This section presents the essential methods, especially for DMs, to assess image quality in the context of image SR, which falls 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) Peak Signal-to-Noise Ratio: 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 $\hat{y}$ and the HR image $\mathbf{y}$

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

Despite being one of the most popular IQA methods, it does not accurately match human perception [15]. 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 [16], whereas generative models tend to produce lower PSNR values [15].

2) SSIM Index: The structural similarity (SSIM), like the PSNR, is a popular evaluation method that focuses on the differences in structural features between images. It independently captures the SSIM 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

$$\mu_\mathbf{y} = \frac{1}{N_\mathbf{y}} \sum_{p \in \Omega_\mathbf{y}} \mathbf{y}_p \quad (5)$$

$$\sigma_\mathbf{y} = \frac{1}{N_\mathbf{y} - 1} \sum_{p \in \Omega_\mathbf{y}} [\mathbf{y}_p - \mu_\mathbf{y}]^2. \quad (6)$$

---

[1]More SR-related IQA methods can be found in [16].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

4

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

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

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

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 an HR image $\mathbf{y}$ and its approximation $\hat{y}$, the luminance ($\mathcal{C}_l$) and contrast ($\mathcal{C}_c$) comparisons are computed using $\mathcal{C}_l(\mathbf{y}, \hat{y}) = S(\mu_{\mathbf{y}}, \mu_{\hat{y}}, c_1)$ and $\mathcal{C}_c(\mathbf{y}, \hat{y}) = S(\sigma_{\mathbf{y}}, \sigma_{\hat{y}}, c_2)$, where $c_1, c_2 > 0$. The empirical covariance

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

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

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

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

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

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

3) Mean Opinion Score: The mean opinion score (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 [15]. 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.

4) Consistency: Consistency measures the degree of stability of nondeterministic SR methods, such as generative models such as 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 MSE.

5) Learned Perceptual Image Patch Similarity: Contrary to the pixel-based evaluation of PSNR and SSIM, the learned perceptual image patch similarity (LPIPS) utilizes a pretrained CNN φ, e.g., VGG [55] or AlexNet [56], 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 lth feature map respectively, and a scaling vector $\alpha_l \in \mathbb{R}^{C_l}$, the LPIPS metric is formulated as follows:

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

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 [16].

6) No-Reference Metrics: All IQA metrics discussed so far require a reference (ground-truth) image. However, there are cases where no reference images are available, e.g., in unsupervised settings. Fortunately, we can assess an image by measuring the distance of statistical features from those obtained from a collection of high-quality images of a similar domain, i.e., natural images. This can be opinion- and distortion-aware like BRISQUE [57] or opinion- and distortion-unaware like NIQE [58]. Another intriguing way to assess no-reference image quality is to exploit the visual-language pretrained CLIP model [59]. One example is CLIP-IQA, which calculates the cosine similarity of the encoded image with two prompts of opposing meaning, i.e., "good photography" and "bad photography" [60]. The resulting relative similarity metric for one or the other prompts determines the image quality. CLIP-IQA shows results comparable to those of BRISQUE without the hand-crafted features and surpasses other no-reference IQA methods like NIQE. Another way to exploit deep learning models is to train them to predict subjective scores using IQA datasets like TID2013 [61]. Examples are DeepQA [62], NIMA [63], or MUSIQ [64]. Others can be found in the learning-based perceptual quality section of the *Hitchhiker's Guide to SR* [16].

III. DMs BASICS

DMs have profoundly impacted the realm of generative AI, and many approaches that fall under the umbrella term DM have emerged. 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 Fig. 1. The forward and backward diffusion processes are distinguished as follows.

1) *Forward q:* degrade input data using noise iteratively, forward in time (i.e., t increases).
2) *Backward p:* denoise the degraded data, thereby reversing the noise iteratively, backward in time (i.e., t decreases).

The time step t increases during forward diffusion, whereas it propagates toward 0 during backward diffusion. Let $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ be a dataset of LR–HR image pairs. For each time step t, the random variable $\mathbf{z}_t$ describes the current state, a state between the image and corruption space. In the 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 $\mathbf{z}_t \sim q(\mathbf{z}_t \mid \mathbf{z}_{t-1})$. Conversely, in the backward diffusion, we assume $\mathbf{z}_{t-1} \sim p(\mathbf{z}_{t-1} \mid \mathbf{z}_t)$. 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 $\mathbf{z}_0 \sim q(\mathbf{x})$, which is then slowly injected with noise (additive). Vice versa, DMs remove noise therein by running a parameterized model $p_{\theta}(\mathbf{z}_{t-1} \mid \mathbf{z}_t)$ in the reverse time direction that approximates the ideal (but unattainable) denoised distribution $p(\mathbf{z}_{t-1} \mid \mathbf{z}_t)$.

The explicit implementation of the forward diffusion q and backward diffusion p, approximated by p_{θ}, is defined by the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

MOSER et al.: DIFFUSION MODELS, IMAGE SUPER-RESOLUTION, AND EVERYTHING: A SURVEY   5
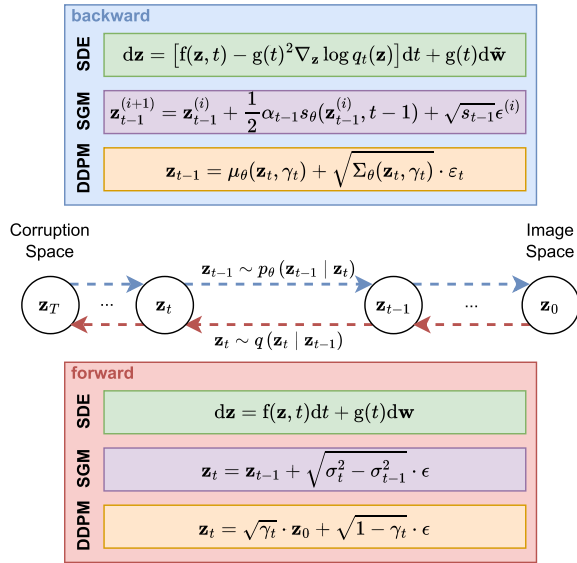

Fig. 1. Principle of DMs. The forward diffusion adds noise iteratively (red), which translates an image from the image space to the corruption space. The backward diffusion, the iterative refinement process, reverts the process (blue) back to the image space. Shown are three different implementations of DMs, namely, DDPMs, SGMs, and SDEs with their respect formulation of the forward and backward diffusions.

specific DM in use. There are three types: Two discrete forms, namely DDPMs and SGMs, and the continuous form by SDEs, which are shown in Fig. 1 and will be discussed next.

A. Denoising Diffusion Probabilistic Models

DDPMs [8] use two Markov chains to enact the forward and backward diffusions across a finite amount of discrete time steps.

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

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

where the hyperparameters $0 < \alpha_{1:T} < 1$ 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

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

where $\gamma_t = \prod_{i=1}^{t}(1-\alpha_i)$ [66]. Consequently, $\mathbf{z}_t$ can be directly sampled regardless of what ought to happen on previous time steps by

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

2) 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

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

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., an LR image) is using $\mu_\theta(\mathbf{z}_t, \mathbf{x}, \gamma_t)$ and $\Sigma_\theta(\mathbf{z}_t, \mathbf{x}, \gamma_t)$ instead.

3) 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

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

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

which leads to minimizing

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

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 (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.

B. Score-Based Generative Models

SGMs, much like DDPMs, utilize discrete diffusion processes but employ an alternative mathematical foundation. Instead of using probability density function $p(\mathbf{z})$ directly, Song and Ermon [10] 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.

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

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}\left(\mathbf{z}_t \mid \mathbf{z}_0, \sigma_t^2 \mathbf{I}\right). \qquad (19)$$

This equation results in a sequence of noisy data densities $q(\mathbf{z}_1), \ldots, q(\mathbf{z}_T)$ with $q(\mathbf{z}_t) = \int q(\mathbf{z}_t) q(\mathbf{z}_0) d\mathbf{z}_0$. 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.

2) 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 $s_\theta(\mathbf{z}_t, t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)$ [11].

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 $\mathbf{z}_t$. This is a key difference to DDPMs as $\mathbf{z}_t$ needs to be sampled iteratively, whereas DDPMs directly predict $\mathbf{z}_t$ from $\mathbf{z}_{t+1}$.

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 and Ermon [10]. Let N be the number of estimation iterations for $\mathbf{z}_t$ at time step t and $\alpha_t > 0$ be the corresponding step size, which determines how much the estimation moves from one estimate $\mathbf{z}_{t-1}^{(i)}$ toward $\mathbf{z}_{t-1}^{(i+1)}$. The initial state is $\mathbf{z}_{t-1}^{(N)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. 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 $i = 0, \ldots, N - 1$:

$$\epsilon^{(i)} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{20}$$

$$\mathbf{z}_{t-1}^{(i+1)} \leftarrow \mathbf{z}_{t-1}^{(i)} + \frac{1}{2}\alpha_{t-1} s_\theta\left(\mathbf{z}_{t-1}^{(i)}, t-1\right) + \sqrt{s_{t-1}}\epsilon^{(i)}. \tag{21}$$

This update rule guarantees that $\mathbf{z}_0^{(N)}$ converges to $q(\mathbf{z}_0)$ for $\alpha_t \to 0$ and $N \to \infty$ [67].

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

3) Optimization: Without specifically formulating the backward diffusion, we can train an 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 [68]

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T) \\ \mathbf{z}_0\sim q(\mathbf{z}_0) \\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t)\sigma_t^2\|\nabla_{\mathbf{z}_t}\log q(\mathbf{z}_t) - s_\theta(\mathbf{z}_t,t)\|^2\right] \\
& \stackrel{(i)}{=} \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T) \\ \mathbf{z}_0\sim q(\mathbf{z}_0) \\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t)\sigma_t^2\|\nabla_{\mathbf{z}_t}\log q(\mathbf{z}_t|\mathbf{z}_0) - s_\theta(\mathbf{z}_t,t)\|^2\right] + c \\
& \stackrel{(ii)}{=} \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T) \\ \mathbf{z}_0\sim q(\mathbf{z}_0) \\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t)\| -\frac{\mathbf{z}_t - \mathbf{z}_0}{\sigma_t} - \sigma_t s_\theta(\mathbf{z}_t,t)\|^2\right] + c \\
& \stackrel{(iii)}{=} \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(1,T) \\ \mathbf{z}_0\sim q(\mathbf{z}_0) \\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}} \left[\lambda(t)\|\epsilon + \sigma_t s_\theta(\mathbf{z}_t,t)\|^2\right] + c
\end{aligned}
\tag{22}
$$

where $\lambda(t) > 0$ is a weighting function, σ_t is the noise level added at time step t, (i) is derived by Vincent [68], (ii) is from (19), (iii) is 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 [69] or sliced score matching [70].

C. Stochastic Differential Equations

So far, we have discussed DMs that deal with finite time steps. A generalization to infinite continuous time steps is made by formulating these as solutions to SDEs, also known as Score SDEs [11]. In fact, we can view SGMs and DDPMs as discretizations of a continuous-time SDE. SDEs are not entirely bound to DMs, as they are a mathematical concept describing stochastic processes. As such, they fit perfectly to describe the processes we want to simulate in DMs. Like previously, data are perturbed in a general diffusion process but generalized to an infinite number of noise scales.

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

$$d\mathbf{z} = f(\mathbf{z}, t)dt + g(t)d\mathbf{w} \tag{23}$$

where f and g are the drift and diffusion functions, respectively, and $\mathbf{w}$ is the standard Wiener process (also known as Brownian motion). This generalized formulation allows uniform representation of both DDPMs and SGMs. The SDE for DDPMs is given by

$$d\mathbf{z} = -\frac{1}{2}\alpha(t)\mathbf{z}dt + \sqrt{\alpha(t)}d\mathbf{w} \tag{24}$$

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

$$d\mathbf{z} = \sqrt{\frac{d[\sigma(t)^2]}{dt}}d\mathbf{w} \tag{25}$$

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

2) Backward Diffusion: The reverse-time SDE is formulated by Anderson [71] as

$$d\mathbf{z} = \left[f(\mathbf{z}, t) - g(t)^2\nabla_{\mathbf{z}}\log q_t(\mathbf{z})\right]dt + g(t)d\tilde{\mathbf{w}} \tag{26}$$

where $\tilde{\mathbf{w}}$ is the standard Wiener process when time flows backward and dt an infinitesimal negative time step. Solutions to (26) 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. [11] and is

$$d\mathbf{z} = \left[f(\mathbf{z}, t) - \frac{1}{2}g(t)^2\nabla_{\mathbf{z}}\log q_t(\mathbf{z})\right]dt. \tag{27}$$

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

3) 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 (22) to continuous time as follows:

$$\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(0,T) \\ \mathbf{z}_0\sim q(\mathbf{z}_0) \\ \mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{z}_0)}} \left[\lambda(t)\|s_\theta(\mathbf{z}_t, t) - \nabla_{\mathbf{z}_t}\log q_t(\mathbf{z}_t | \mathbf{z}_0)\|^2\right] \tag{28}$$

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

D. Relation Between DMs

As highlighted in the SDE section, we can describe both variations, namely, SGMs, and DDPMs, with SDEs. We can also showcase this close relationship by reformulating the optimization targets. For DDPMs, we saw in (18) that

$$
\begin{aligned}
& \mathrm{KL}(q(\mathbf{z}_0, \ldots, \mathbf{z}_T)\|p_\theta(\mathbf{z}_0, \ldots, \mathbf{z}_T)) \\
& \stackrel{(ii)}{\geq} \mathbb{E}[-\log p_\theta(\mathbf{z}_0)] + c
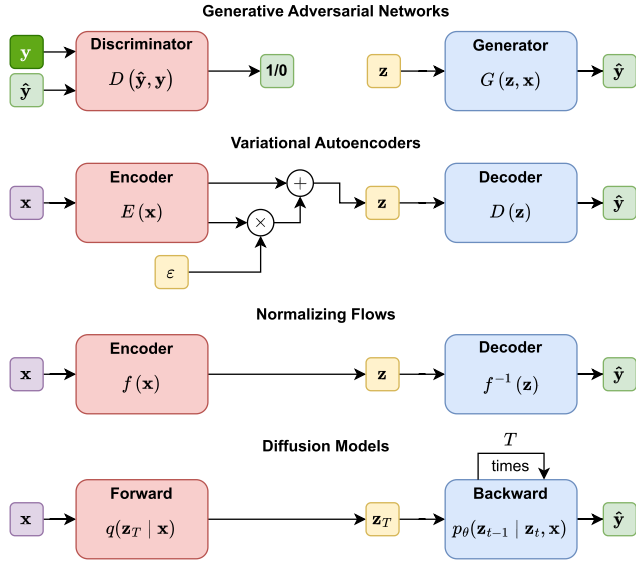\end{aligned}
$$

Fig. 2. Conceptual overview of generative models (GANs, VAEs, NFs, and DMs).

is minimized. By reweighting the VLB, as Ho et al. [8] recommend for improved sample quality, we can further derive

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

where $\lambda(t) > 0$ is a weighting function. If we now take the optimization target in (22) of SGMs, which was

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

the connection between DDPMs and SGMs becomes clear once we set $\epsilon_\theta(\mathbf{z}_t,t) = -\sigma_t s_\theta(\mathbf{z}_t,t)$. As the constant c is irrelevant for the optimization, we can see once again that there is a mathematical connection between DDPMs and SGMs.

E. Relation to Other Image SR Generative Models

Generative models in image SR differ primarily in how they approach the task of generating HR images from LR inputs and are illustrated in Fig. 2. These differences stem from the underlying architecture and training objectives. While they offer significant advantages, they come with an individual set of challenges, such as training stability and computational costs.

1) GAN: One prominent category of generative models is GANs [72], which have demonstrated the state-of-the-art performance in various vision-related tasks, including text-to-image (T2I) synthesis [7] and image SR [44]. GANs are known for their adversarial training, where a generator competes against a discriminator. Although DMs do not employ a discriminator, they utilize a similar adversarial training strategy by iteratively adding and removing noise to enable realistic data generation. However, approaches with GANs often suffer from nonconvergence, training instability, and high computational costs. They require careful hyperparameter tuning due to the interplay between the generator and the discriminator.

2) VAE: VAEs [73] 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 VLB 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.

3) ARM: ARMs treat images as sequences of pixels and generate each pixel based on the values of previously generated pixels in a sequential manner [6]. 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.

4) NF: NFs [74] 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 [75] 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 [76].

IV. IMPROVEMENTS FOR DMs

In the broader research community, there are several ways to improve DMs for image generation, as presented, for example, by Karras et al. [77]. This section, however, focuses on enhancements particularly interesting for image SR: efficient sampling and enhanced likelihood estimation.

A. Efficient Sampling

Efficient sampling refers to strategies that generate samples from noise more quickly, i.e., in fewer time steps, without compromising the quality of the produced image significantly. For instance, a DDPM takes about 20 h to sample $50\,000$ 32×32 images, in contrast to a GAN's less than one minute on a Nvidia 2080 Ti GPU; for larger 256×256 images, this extends to nearly 1000 h [78]. Fortunately, the independence between training and inference schedules is often leveraged in image SR. For example, a model may undergo training with 1000 time steps, but the subsequent inference phase may require only a fraction, i.e., 200 [15], [79]. However, the broader community of DM research has made further attempts focusing on either training-based or 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 process may be complete or partial. For example, Watson et al. [80] 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 [81] offers another approach, optimizing fast samplers for pretrained DMs by adjusting the kernel inception distance. Another technique is truncated diffusion, which improves speed by prematurely ending the forward diffusion process [82], [83]. This early termination results in outputs that are not purely Gaussian noise, presenting computational challenges. These challenges are addressed using proxy distributions from pretrained VAEs or GANs, which match the diffused data distribution and facilitate efficient backward diffusion. Lastly, 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) [84], [85]. As demonstrated by Salimans and Ho [86], this method progressively reduces the number of sampling steps, trading off a slight decrease in sample quality for increased speed. Similarly, Xiao et al. [87] 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. For image SR, an application for exploiting knowledge distillation can be found in AddSR [88]. Similarly, YONOS-SR [89] uses knowledge distillation, but instead of training faster samplers, they transfer different scaling task knowledge and use the training-free denoising diffusion implicit models (DDIMs) for efficient sampling, which is presented in the next section.

Training-free sampling methods aim to speed up sampling by minimizing the number of discretization steps while solving the SDE or probability flow ODE [90], [91]. DDIMs introduced by Song et al. [90] 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. Jolicoeur-Martineau et al. [91] 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 compromising sample quality. Building upon DDIM and Jolicoeur-Martineau et al. [91], the DPM-solver [92], inspired by the AnalyticalDPM [93], approximates the error prediction via Taylor expansion and thus achieves efficient sampling by analytically resolving the linear component of the ODE solution instead of relying on generic black-box ODE solvers. This method significantly reduces the sampling steps to 10–20. In a later work, Lu et al. [94] introduced an improved version with DPM-solver++ that essentially approximates the predicted image instead of the error. Lately, a more general formulation and extension of the DPM-solver++ was introduced by UniPC [95].

B. Improved Likelihood

Log-likelihood improvement is directly coupled with enhancing the performance of various applications and methods, including but not limited to compression [96], semi-supervised learning [97], 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. [98] 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 chosen weighting function sets an upper bound on the negative log-likelihood in the weighted score-matching objective. Upon minimization, this results in an elevation of the log-likelihood. Kingma et al. [99] explored methods that simultaneously train the noise schedule and diffusion parameters to maximize the VLB within variational DMs. Additionally, the improved denoising diffusion probabilistic models (iDDPMs) proposed by Nichol and Dhariwal [100] 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.

V. DMs FOR IMAGE SR

So far, we introduced the theoretical framework of DMs. This section reviews practical applications and recent advances in image SR. We will discuss concrete realizations of DMs, which are predominantly DDPMs. We then discuss guidance strategies to enhance conditioning usage, represent conditioning information in alternative state domains for DDPMs, and incorporate various conditioning methods. Additionally, we explore SR-specific research areas, including corruption spaces, color-shifting, and architectural designs. Fig. 3 provides a topological overview of this section.

A. Concrete Realization of DMs

While SGMs provide considerable design flexibility, the image SR trend leans toward DDPMs. DDPMs benefit from a straightforward implementation, which reduces the entry barrier. It is a significant advantage, as it allows quicker development cycles and replication of results. In addition, while the flexibility of SGMs is advantageous in creating customized solutions, it introduces design complexity due to the multitude of design variables that need to be considered. This poses a challenge in research settings, where rigorously evaluating the impact of each variable (e.g., different sampling algorithms) becomes cumbersome. Moreover, the growing DDPM literature contributes to their popularity. As more studies adopt DDPMs, a virtuous cycle is created, where familiarity and proven effectiveness encourage further adoption.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

MOSER et al.: DIFFUSION MODELS, IMAGE SUPER-RESOLUTION, AND EVERYTHING: A SURVEY
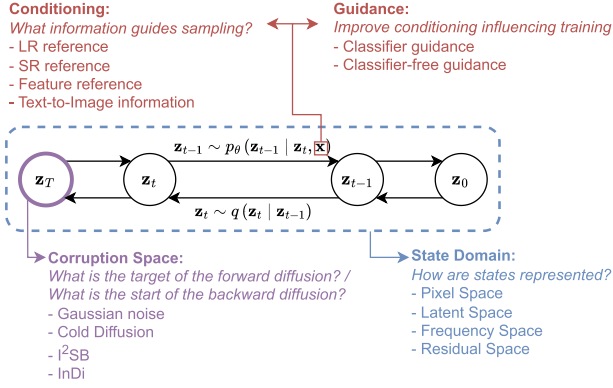
9

Fig. 3. Topology of this work. Conditioning (Section V-D) leads the backward diffusion, whereas guidance (Section V-B) is a training strategy to improve the incorporation of conditioning into DMs. The state domain (Section V-C) describes the representation of states $\mathbf{z}_t$. The corruption space (Section V-E) describes the target of the forward diffusion process or the start of the backward diffusion.

Among the pioneering DM efforts is SR3 [15], which concretely realizes DDPMs for image SR. Like typical for 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 (14) to approximate $\mathbf{z}_0$ as follows:

$$\mathbf{z}_t = \sqrt{\gamma_t} \cdot \hat{z}_0 + \sqrt{1 - \gamma_t} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t)$$
$$\iff \hat{z}_0 = \frac{1}{\sqrt{\gamma_t}} \cdot \left(\mathbf{z}_t - \sqrt{1 - \gamma_t} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t) \right). \quad (29)$$

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

$$\mu_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t) = \frac{1}{\sqrt{\alpha_t}} \left[\mathbf{z}_t - \frac{1 - \alpha_t}{\sqrt{1 - \gamma_t}} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t) \right]. \quad (30)$$

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

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left[\mathbf{z}_t - \frac{1 - \alpha_t}{\sqrt{1 - \gamma_t}} \cdot \varphi_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t) \right] + \sqrt{1 - \alpha_t} \cdot \epsilon_t. \quad (31)$$

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 [79], 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 images, i.e., the difference. Thus, it has an alternative state domain, which will be discussed next.

B. Guidance in Training

The backbone of diffusion-based image SR is the learning of conditional distributions [15], [101]. As such, the condition $\mathbf{x}$, e.g., the LR image, is integrated into the backward

diffusion, i.e., $p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{x})$ 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. A principle known as guidance can mitigate this issue by controlling the weighting of the conditioning information at the expense of sample diversity. It can be categorized into classifier and classifier-free guidance. To the authors' knowledge, while effectively used for improving DMs, they have not been applied to image SR.

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 [12]. This process is similar to low temperature or truncated sampling in BigGANs [102] and facilitates a tradeoff 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

$$\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t \mid \mathbf{x}) = \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t) + \lambda \nabla_{\mathbf{z}_t} \log q(\mathbf{x} \mid \mathbf{z}_t) \quad (32)$$

where $\lambda \in \mathbb{R}^+$ is a hyperparameter 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 pretrained image classification models lack.

2) Classifier-Free Guidance: Classifier-free guidance aims to achieve similar results without a classifier [103]. It modifies (32) into

$$\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{x}) = (1 - \lambda) \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t) + \lambda \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t \mid \mathbf{x}). \quad (33)$$

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 [104].

C. State Domains

So far, we have discussed methods that operate directly on the pixel space. This section introduces different methods that map the input into alternative state domains: latent, frequency, and residual space. Apart from particular challenges arising from the alternative state domain, these methods incur an additional step that maps the pixel domain into their own, as illustrated in Fig. 4.

1) Latent Space: Models such as SR3 [15] and SRDiff [79] 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,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

10

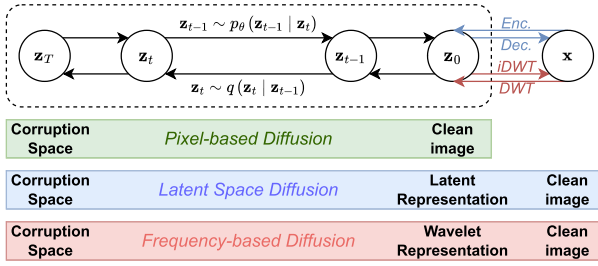IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

Fig. 4. Overview of state domains. The green bar shows the vanilla DM operating in pixel space. The blue bar shows the exploit of the latent space domain via autoencoders. The red bar shows the application of DMs in the wavelet domain.

one can move the diffusion process into the latent space of an autoencoder [105]. The first of this kind was the latent score-based generative models (LSGMs) by Vahdat et al. [106]. It is a regular SGM that operates in the latent space of a VAE and, by pretraining the VAE, achieves even faster sampling speeds. It yields comparable and better results than DMs operating in the pixel domain while being faster. Building upon LSGMs, Rombach et al. [13] introduced the latent diffusion model (LDM) [107], which also performs diffusion in a low-dimensional latent space of an autoencoder. In contrast to LSGM, LDM utilizes a DDPM and an autoencoder that is pretrained, like the VQ-GAN [5], and is not jointly trained with the denoising network. This approach significantly lowers resource requirements without compromising performance. Due to the decoupled training, it requires very little regularization of the latent space and allows the reuse of latent representations across multiple models. Improving upon LDMs is REFUSION (image REstoration with difFUSION models) by Luo et al. [108], which differs in two aspects. First, it uses a U-Net that contains skip connections from the encoder to the decoder, which provides the decoder with additional details. Moreover, it introduces nonlinear activation-free blocks (NAFBlocks) [109], replacing all nonlinear activations with an elementwise operation that splits feature channels into two parts and multiplies them to produce one output. Second, they train their U-Net with a latent-replacing training strategy, which partially replaces the latent representation with either the encoded LR or HR image for reconstruction training. Similarly, Chen et al. [110] improve the architectural aspects of LDMs and propose a two-stage strategy called the hierarchical integration diffusion model (HI-Diff). In the first stage, an encoder compresses the ground-truth image to a highly compact latent space representation, which has a much higher compression ratio than LDM. As a result, the computational burden of the DM, which refines multiscale latent representations, is much more reduced. The second stage is a vision transformer (ViT)-based autoencoder, which incorporates the latent representations of the first stage during the downsampling process via hierarchical integration modules (HIMs), a cross-attention fusion module.

2) Frequency Space: Wavelets provide a novel outlook on SR [16], [111]. The conversion from the spatial to the wavelet domain is lossless and offers significant advantages as 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, the conversion segregates high-frequency details into distinct channels, facilitating a more concentrated and intentional focus on high-frequency information, offering a higher degree of control [112]. 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 [113] or specifically target certain bands while the remaining bands are predicted via standard CNNs. For instance, WaveDM [114] modifies the low-frequency band, whereas WSGM [115] or ResDiff [116] conditions the high-frequency bands relative to the LR image. Altogether, the wavelet domain presents a promising avenue for future research. It provides the potential for significant performance acceleration while maintaining, if not enhancing, the quality of SR results.

3) Residual Space: SRDiff [79] was the first work that advocated for shifting the generation process into the residual space, i.e., 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 [16], [111]. Whang et al. [117] also employ residual predictions as a fundamental component of their predict-and-refine approach for image deblurring. However, unlike SRDiff, they provide an SR prediction with a CNN instead of the bilinear upsampled LR and predict the residuals between the SR prediction and the HR ground truth with their DM. An improvement is presented by ResDiff [116], which additionally incorporates the SR prediction and its high-frequency information during the backward diffusion for better guidance. In a different vein, Yue et al. [118] present ResShift. This technique constructs a Markov chain of transformations between HR and LR images by manipulating the residual between them. Thus, instead of just adding Gaussian noise with zero mean in the forward process, the residual is also added as the mean of the noise sampling during training. This novel approach substantially enhances sampling efficiency, i.e., only 15 sampling steps.

D. Conditioning DMs

DMs depend on conditioning information to guide the sampling process toward a reasonable HR prediction. One common strategy is to use the LR image during the backward diffusion. This section reviews various alternative methods for integrating conditioning information into backward diffusion.

1) Low-Resolution Reference: High-quality SR predictions can be achieved through a straightforward channel concatenation [119]. 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. [120] conditions the generative process of an unconditional LDM [13]. This approach offers the advantage of shorter training times, as it leverages a pretrained 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.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

MOSER et al.: DIFFUSION MODELS, IMAGE SUPER-RESOLUTION, AND EVERYTHING: A SURVEY
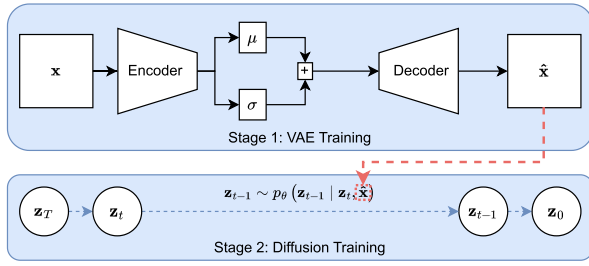
11

Fig. 5. Overview of DiffuseVAE. The two-stage approach employs a VAE (first stage), which generates variational prediction as a condition for the DM (second stage).

TABLE I

RESULTS FOR 4× SR OF GENERAL IMAGES ON DIV2K VAL. NOTE THAT EDSR, FxSR-PD, CAR, AND RRDB ARE THE REGRESSION-BASED METHODS THAT GENERALLY PRODUCE BETTER PSNR AND SSIM SCORES THAN GENERATIVE APPROACHES [15]

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Bicubic	26.70	0.77	0.409
EDSR	28.98	0.83	0.270
FxSR-PD	29.24	0.84	0.239
RRDB	29.44	0.84	0.253
CAR	32.82	0.88	-
RankSRGAN	26.55	0.75	0.128
ESRGAN	26.22	0.75	0.124
SRFlow	27.09	0.76	0.120
SRDiff	27.41	0.79	0.136
IDM	27.59	0.78	-
DiWa	28.09	0.78	0.104

TABLE II

PSNR AND SSIM COMPARISON ON CELEBA-HQ FACE SR $16\times16 \rightarrow 128\times 128$. CONSISTENCY MEASURES MSE $(\times10^{-5})$ BETWEEN LR INPUTS AND THE DOWNSAMPLED SR OUTPUTS

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

2) Super-Resolved Reference: An alternative to conditioning the denoising on the LR image involves learned priors from pretrained SR models to predict a reference image. For example, CDPMSR [121] conditions the denoising process with a predicted SR reference image obtained using existing and standard SR models. ResDiff [116], on the other hand, leverages a pretrained CNN to predict a low-frequency, content-rich image that includes partial high-frequency components. This image guides the noise toward the residual space, offering an alternative means of conditioning the generative process.

Pandey et al. [127] introduced an exciting idea of varying predicted conditions with DiffuseVAE as illustrated in Fig. 5. 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 NFs.

3) Feature Reference: Another avenue for conditioning involves relevant features extracted from pretrained networks. SRDiff [79] leverages a pretrained 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) [100] 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 Tables I and II. Recently, DeeDSR was introduced [128], which incorporates degradation-aware features extracted from the LR image to guide the diffusion process of an LDM [107].
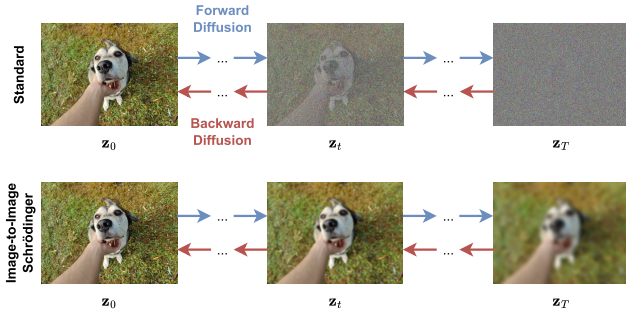
4) T2I Information: By incorporating conditioning information that goes beyond the LR image (e.g., its SR prediction, direct concatenation of the LR image, or its feature representation), one can add T2I information. The incorporation of T2I information proves advantageous as it allows the usage of pretrained 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 improving image synthesis and interpretation in SR tasks. Wang et al. [107] have put this concept into practice with StableSR. Central to StableSR is a time-aware encoder trained in tandem with a frozen stable DM, essentially an 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 [129], which enhances the versatility of StableSR in catering to diverse user needs and preferences. Likewise, Yang et al. [130] introduce a method known as pixel-aware stable diffusion (PASD). PASD takes conditioning a step further by incorporating text embeddings of the LR input using a CLIP text encoder [59] 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 III, demonstrating the impact of this text-based conditioning on image SR results. A similar concurrent work can be found

TABLE III
RESULTS FOR 4× SR OF GENERAL IMAGES ON RESIZED DIV2K VAL
(128 × 128 → 512 × 512)

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

Fig. 7. Example of color shifting produced by vanilla SR3 in a $64 \times 64 \to 256 \times 256$ setting when trained with a reduced batch size (8 instead of 256).

Fig. 6. Comparison of the standard corruption space and I²SB. Instead of injecting noise to the clean image (initial state $\mathbf{z}_0$), the final state $\mathbf{z}_T$ is the degraded image.

with SeeSR [131]. XPSR [132] extends this idea by fusing different levels of semantic text encodings (high-level: the content of the image; and low-level: the perception of overall quality, sharpness, noise level, and other distortions about the LR image).

E. Corruption Space

Karras et al. [77] 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 instead of pure Gaussian noise used during forward diffusion like soft score matching [135], i.e., the starting point for backward diffusion or the target for the forward diffusion $\mathbf{z}_T$. Soft score matching 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. Note that $\mathbf{z}_T$ may be represented differently due to alternative state domains (e.g., latent, frequency, or residual). Cold Diffusion [136] presents another ingenious way of modifying the corruption space for DDPMs. It shows that the generative capability is not strongly dependent on the choice of image degradation. It reveals that 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 (I²SB) goes in a similar direction but does not impose any assumptions on the underlying prior distributions [137]. 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, as illustrated in Fig. 6. 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 I²SB 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. Another avenue for alternative corruption space is presented by inversion by direct iteration (InDI) [138]. InDI delineates a direct mapping strategy, efficiently bridging the gap between the two quality spaces without the iterative refinement typically required by conventional diffusion processes. The intrinsic flexibility and the direct mapping capability of InDI propose intriguing possibilities for enhancing image quality, suggesting a potent avenue for research exploration. The potential integration of InDI's principles with those of conditional DMs could offer substantial advancements in the field of image SR. A detailed examination and discussion of InDI within the broader scope of diffusion-based image enhancement could yield valuable insights and contribute significantly to the ongoing development of generative models in image processing.

F. 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 [139]. An example with SR3 is shown in Fig. 7. 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 [107]. Mathematically, it gives the following equation:

$$\hat{z}_0 = \frac{\mathbf{z}_0^c - \mu_{\mathbf{z}_0}^c}{\sigma_{\mathbf{z}_0}^c} \cdot \sigma_{\mathbf{x}}^c + \mu_{\mathbf{x}}^c \tag{34}$$

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 cth channel of the predicted image $\mathbf{z}_0$ (or the input image $\mathbf{x}$), respectively. You only diffuse areas (YODA) [140], which targets diffusion on important image areas more frequently through time-dependent masks generated with DINO [141], 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.

G. 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 [102]. SR3 [15], for instance, employs residual blocks from BigGAN [102] and rescales skip connections by a factor of $(1/(2)^{1/2})$. SRDiff takes a similar approach [79] although it opts for vanilla residual blocks without the rescaling of skip connections and uses an LR encoder to incorporate the information of the LR image during the backward diffusion. Whang et al. [117] exploit an initial predictor to combine the strengths of deterministic image SR models and DMs. It has the advantage that the DM only needs to learn the residuals that the deterministic image SR model (initial predictor) fails to predict, thus 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 [113]. Moreover, wavelet SR models, such as DWSR [112]—a simple sequence of convolution layers of depth 10—are utilized for denoising prediction in the wavelet domain. In WaveDM [114], a deterministic U-Net predictor is used for the high-frequency band, while diffusion is applied in the low-frequency band.

LDMs proposed by Rombach et al. [13] use a VQ-GAN [5] autoencoder in the latent space. For DiffIR [142], multiple variations of state-of-the-art ViTs are employed [50], [143], [144]. Another common practice is pretraining deterministic components, as seen in models like SRDiff [79] or DiffIR [142]. 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.

VI. DIFFUSION-BASED ZERO-SHOT SR

Zero-shot image SR aims to develop methods that do not depend on prior image examples or training [16], [150]. Typically, these methods harness the inherent redundancy within a single image for improvement. They often leverage pretrained DMs for generation, incorporating LR images as conditions during the sampling process, in contrast to other conditioning methods discussed earlier [151]. 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. [17] 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 IV.

A. 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 [152]. 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 [140] applies a similar technique, but for image SR. YODA incorporates importance masks derived from DINO [141] to define the areas for diffusion during each time step, but it is not a zero-shot approach.

One zero-shot method is ILVR [120], 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) [153], which modifies the unified projection method to SR as follows:

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

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

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

B. Decomposition-Based

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

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

where $\mathbf{A}$ is the degradation operator and b is the contaminating noise. Among the earliest decomposition-based methods, we find SNIPS [145] and its subsequent work DDRM [146]. 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 [148] beyond image SR to tasks such as colorization, inpainting, and deblurring [148]. It leverages the range–null space decomposition methodology [154], [155] to tackle diverse IR challenges effectively. DDNM approaches the problem by reconfiguring (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 $\mathbf{x} \in \mathbb{R}^{d \times 1}$ as the linearized degraded image. Furthermore, it has to conform to the following two constraints:

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

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

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

TABLE IV
COMPARISON OF ZERO-SHOT METHODS. DATA IN BOLD REPRESENT THE BEST PERFORMANCE. SECOND-BEST IS UNDERLINED. VALUES DERIVED FROM LI ET AL. [17]

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

with $\mathbf{A}^\dagger \in \mathbb{R}^{D \times d}$ the pseudo-inverse that satisfies $\mathbf{AA}^\dagger\mathbf{A} \equiv \mathbf{A}$. 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 (39).

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

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

with $\epsilon_t = \epsilon_\theta(\mathbf{z}_t, t)$. To produce a $\mathbf{z}_0$ that fulfills the equation $\mathbf{Az}_0 \equiv \mathbf{x}$, 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

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

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

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

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=0}^t \alpha_i$, illustrated in Fig. 8.

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 $[(1/n^2) \dots (1/n^2)]$, functions to average each patch into a singular value. Similarly, we can construct its pseudo-inverse as $\mathbf{A}^\dagger \in \mathbb{R}^{n^2 \times 1} = [1 \dots 1]^\top$. 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 suboperations 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 [152]. This approach was leveraged to enhance the quality further.
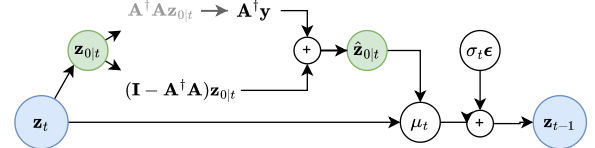

Fig. 8. Overview of DDNM [148]. It utilizes the range–null space decomposition to construct a general solution for multiple tasks, such as image SR, colorization, inpainting, and deblurring.

Given this approach's novelty, only a handful of subsequent studies extend and build upon it, such as the work presented in CDPMSR [121]. This research direction promises exciting possibilities 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.

C. Posterior Estimation

Most projection-based methods typically address the noise-less inverse problem. However, this assumption can weaken data consistency because the projection process can deviate the sample path from the data manifold [17]. To address this and enhance data consistency, some recent works [147], [156], [157] take a different approach by aiming to estimate the posterior distribution using the Bayes theorem

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

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

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

where $s_\theta(\mathbf{x}, t)$ is extracted from a pretrained 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 [156] and DPS [147] approximate the posterior $p_t(\mathbf{x}|\mathbf{z}_t)$ with $p_t(\mathbf{x}|\hat{z}_0(\mathbf{z}_t))$, where $\hat{z}_0(\mathbf{z}_t)$ is the expectation given $\mathbf{z}_t$ as $\hat{z}_0(\mathbf{z}_t) = \mathbb{E}[\mathbf{z}_0|\mathbf{z}_t]$ according to Tweedie's formula [147]. While MCG also relies on projection, which can

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

MOSER et al.: DIFFUSION MODELS, IMAGE SUPER-RESOLUTION, AND EVERYTHING: A SURVEY

15

be harmful to data consistency, DPS discards the projection step and estimates the posterior as

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{x} \mid \mathbf{z}_t) \approx \nabla_{\mathbf{z}_t} \log p\left(\mathbf{x} \mid \hat{z}_0(\mathbf{z}_t)\right)$$
$$\approx -\frac{1}{\sigma^2} \nabla_{\mathbf{z}_t} \|\mathbf{x} - H\left(\hat{z}_0(\mathbf{z}_t)\right)\|_2^2 \quad (46)$$

where H is a forward measurement operator. A further expansion of this formula to the unified form for the linear, nonlinear, differentiable inverse problem with Moore–Penrose pseudoinverse can be found in IIGDM [157].

A different approach to estimate $p_t(\mathbf{x}|\mathbf{z}_t)$ is demonstrated by GDP [149]. 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

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

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 $\hat{z}_0$ in the distance calculation, providing a pragmatic solution to the noise discrepancy issue.

VII. DOMAIN-SPECIFIC APPLICATIONS

SR3 [15] 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 [158]. This section highlights the applications of DMs to domain-specific SR tasks: medical imaging, special cases of face SR (BFR and ATs), and remote sensing.

A. 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. [159] propose a combined denoising and SR network referred to as regularized reverse diffusion denoiser + SR (R2D2+). 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. [153], 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. [160] address the lack of diffusion-based multicontrast MRI SR methods. They propose a disentangled conditional diffusion model (DisC-Diff) to leverage a multiconditional 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. An improvement of DisC-Diff by combining the DM with a transformer was introduced by Li et al. [161] with DiffMSR.

B. 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 BFR, where datasets typically contain supervised samples $(\mathbf{x}, \mathbf{y})$ with unknown degradation.

A solution to BFR was proposed by Yue and Loy [162] with DifFace that leverages the rich generative priors of pretrained 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 [163], [164], 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 pretrained DM, DifFace requires no full retraining for new and unknown degradations, unlike SR3.

A concurrent and better performing approach is Diff-BFR [165] that adopts a two-step approach to BFR: an 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 an LR 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. However, it has more parameters than SR3 and requires optimization to accelerate sampling.

Another method is DR2E [166], which employs two stages: degradation removal and enhancement modules. For degradation removal, they use a pretrained 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 is close in distribution. Theoretically, it produces visually clean intermediate results that are

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

16   IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

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.

C. AT in Face SR

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. [167] 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. [168] 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.

D. 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 [169]. To produce finer details and texture, Liu et al. [170] present DMs with a detail complement (DMDC) mechanism. They train their model similar to SR3 [15] 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. [171] design a new architecture for RS images that integrates ViTs with DMs as a two-stage approach for enhancement and super-resolution (TESR). In the first stage (SR stage), the SwinIR [50] 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. [172] 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. Recently, Khanna et al. [173] introduced DiffusionSat, which uses an LDM for RSSR and incorporates additional remote sensing conditioning information (e.g., longitude, latitude, cloud cover, etc.).

VIII. 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.

A. Color Shifting

Often, the most practical advancements come from a solid theoretical understanding. As discussed in Section V-F, 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 [139]. While well-defined diffusion methods [140] or color normalization [107] might mitigate this problem, a theoretical understanding of why it is emerging is necessary.

B. Computational Costs

In a study conducted by Ganguli et al. [174], it was observed that the computing power needed for large-scale AI experiments has surged by over 300 000 times in the last decade. 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 V-C. Examples of such approaches include LDMs [5], [13] and wavelet-based models [113], [115]. 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 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.

C. Efficient Sampling

A benefit of DMs is the possibility of decoupling training and inference schedules [175]. This allows for substantial enhancements in 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 has yet to be developed [176]. As outlined in Section IV-A, this research direction represents a promising avenue. We explored training-based sampling

methods for SR with AddSR [88] and YONOS-SR [89] but also introduced efficient DMs that need fewer sampling steps, like ResShift [118] and DiffIR [142]. An alternative is given by methods that use different corruption spaces, as discussed in Section V-E. Unlike sampling from pure Gaussian noise, notable works such as Luo et al. [108], I²SB [137], CCDF [153], or Cold Diffusion [136] define a process from the LR to the HR image directly. Additional techniques for decreasing computation time, such as knowledge distillation, alternative noise schedulers, or truncated diffusion, demand further investigation concerning image SR [84], [85], [87], [177].

D. Corruption Spaces

New approaches for corruption spaces allow a more direct approach for upsampling images from LR to HR. The significance of exploring different corruption spaces lies in addressing the inherent limitations and assumptions embedded within current DM frameworks, e.g., diversity and blurriness added during the forward diffusion process. The adaptability and efficiency demonstrated by novel approaches like InDI or I²SB, especially in handling diverse and complex corruption patterns spotlight the urgent need for future research.

E. Comparability

Comparing DMs in SR is complex because of the varied datasets used in different studies. They vary in resolution, content diversity, color distribution, and noise levels, all of which significantly influence model performance. A model may perform well with one dataset but poorly with another, complicating the assessment of its overall effectiveness. Establishing a standard benchmark with diverse, representative datasets and uniform evaluation metrics is essential for comparability. This approach would help identify models that consistently perform well across different conditions and tasks, thereby promoting faster progress in the field. Furthermore, evaluating the quality of SR images from generative models is still problematic. Although DMs often produce more photorealistic images, they typically score lower on standard metrics like PSNR and SSIM [16]. However, these models tend to receive more favorable assessments from human evaluators [15]. LPIPS [178] 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 [179], [180]. For instance, datasets with subjective ratings, like TID2013 [61], and neural networks, such as DeepQA [62] or NIMA [63], can be employed to predict human-like scoring of images and should be further explored.

F. Image Manipulation

Image manipulation can be particularly useful in multi-image SR for generating HR 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). SRDiff [79] proposed two potential extensions: 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 [181]. 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 VAEs or GANs, DMs have been found to offer less proficient latent representations [182]. Therefore, recent and ongoing research into the manipulation of latent representations in DMs is both in its early stages and greatly needed [183], [184], [185].

G. Cascaded Image Generation

Saharia et al. [15] presented cascaded image SR, in which multiple DDPMs are chained 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. Even though their approach underperforms compared to BigGAN [102] concerning cascaded generation, it still represents an exciting research opportunity.

IX. CONCLUSION

DMs revolutionized image SR by enhancing both technical image quality and human perceptual preferences. While traditional SR often focuses solely on pixel-level accuracy, DMs can generate HR images that are esthetically pleasing and realistic. Unlike previous generative models, they do not suffer typical convergence issues. This article explored the progress and diverse methods that have propelled DMs to the forefront of SR. Potential use cases, as discussed in our applications section, extend far beyond what was previously imagined. We introduced their foundational principles and compared them to other generative models. We explored conditioning strategies, from LR image guidance to text embeddings. 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 article provides a comprehensive guide to the current landscape and valuable insights into trends, challenges, and future directions. As we continue to explore and refine these models, the future of image SR looks more promising than ever.

REFERENCES

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

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

18   IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

[2] D. Valsesia and E. Magli, "Permutation invariance and uncertainty in multitemporal image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

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

[4] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11513–11522.

[5] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12868–12878.

[6] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1899–1908.

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

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, vol. 33, 2020.

[9] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, 2020.

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

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

[12] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. NeurIPS*, vol. 34, 2021.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

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

[16] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9862–9882, Aug. 2023.

[17] X. Li et al., "Diffusion models for image restoration and enhancement—A comprehensive survey," 2023, *arXiv:2308.09388*.

[18] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, "Blind image super-resolution: A survey and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5461–5480, May 2023.

[19] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4771–4780.

[20] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.

[21] S. Anwar and N. Barnes, "Densely residual Laplacian super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1192–1204, Mar. 2022.

[22] *MATLAB*, Mathworks, Inc., Natick, MA, USA, 2017.

[23] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.

[24] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, *Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding*. U.K.: The British Machine Vision Association (BMVA) Press, 2012.

[25] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Cham, Switzerland: Springer, 2010.

[26] 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 *Proc. 8th IEEE Int. Conf. Comput. Vision. ICCV*, vol. 2, Jul. 2001, pp. 416–423.

[27] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[28] Y. Matsui et al., "Sketch-based Manga retrieval using Manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.

[29] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

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

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2012). *The PASCAL Voc2012 Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[33] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[34] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, no. 2, pp. 1–11, Apr. 2011.

[35] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[36] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jul. 2004.

[37] W. Freeman, T. Jones, and E. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar. 2002.

[38] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[39] M. Irani and S. Peleg, "Improving resolution by image registration," *Graph. Models Image Process.*, vol. 53, no. 3, pp. 231–239, May 1991.

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

[41] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[42] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016.

[43] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[44] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[46] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4809–4817.

[47] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[48] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.

[49] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. ECCV*, 2018.

[50] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

MOSER et al.: DIFFUSION MODELS, IMAGE SUPER-RESOLUTION, AND EVERYTHING: A SURVEY
19

[51] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22367–22377.

[52] C.-C. Hsu, C.-M. Lee, and Y.-S. Chou, "DRCT: Saving image super-resolution away from information bottleneck," 2024, *arXiv:2404.00722*.

[53] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. ECCVW*, 2018.

[54] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "SRFlow: Learning the super-resolution space with normalizing flow," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 715–732.

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

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[57] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[58] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[59] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.

[60] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI*, 2023, vol. 37, no. 2.

[61] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[62] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1969–1977.

[63] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[64] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5128–5137.

[65] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, Apr. 2024.

[66] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. ICML*, 2015.

[67] G. Parisi, "Correlation functions and computer simulations," *Nucl. Phys. B*, vol. 180, no. 3, pp. 378–384, May 1981.

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

[69] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, no. 4, 2005.

[70] Y. Song, S. Garg, J. Shi, and S. Ermon, "Sliced score matching: A scalable approach to density and score estimation," in *Proc. 35th Uncertainty Artif. Intell. Conf.*, vol. 115. PMLR, 2020, pp. 574–584.

[71] B. D. O. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes their Appl.*, vol. 12, no. 3, pp. 313–326, May 1982.

[72] I. Goodfellow et al., "Generative adversarial nets," in *Proc. NeurIPS*, vol. 27, 2014.

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

[74] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. ICML*, 2015.

[75] Q. Zhang and Y. Chen, "Diffusion normalizing flow," in *Proc. NeurIPS*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[76] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *J. Mach. Learn. Res.*, vol. 22, no. 1, 2021.

[77] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. NeurIPS*, vol. 35, 2022.

[78] *Oxford VGGFace Implementation Using Keras Functional Framework V2+*. Accessed: Oct. 17, 2024. [Online]. Available: https://github.com/rcmalli/keras-vggface

[79] H. Li et al., "SRDiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, Mar. 2022.

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

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

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

[83] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3813–3824.

[84] C. Meng et al., "On distillation of guided diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14297–14306.

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

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

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

[88] R. Xie et al., "AddSR: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation," 2024, *arXiv:2404.01717*.

[89] M. Noroozi, I. Hadji, B. Martinez, A. Bulat, and G. Tzimiropoulos, "You only need one step: Fast super-resolution with stable diffusion via scale distillation," 2024, *arXiv:2401.17258*.

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

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

[92] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 5775–5787.

[93] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models," 2022, *arXiv:2201.06503*.

[94] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," 2022, *arXiv:2211.01095*.

[95] W. Zhao, L. Bai, Y. Rao, J. Zhou, and J. Lu, "UniPC: A unified predictor-corrector framework for fast sampling of diffusion models," in *Proc. NeurIPS*, vol. 36, 2024.

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

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

[98] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," in *Proc. NeurIPS*, vol. 34, 2021.

[99] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Proc. NeurIPS*, vol. 34, 2021.

[100] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. ICML*, 2021.

[101] 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, 2022.

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

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

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

[105] J. Kim and T.-K. Kim, "Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder," 2024, *arXiv:2403.10255*.

[106] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," in *Proc. NeurIPS*, vol. 34, 2021.

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

[108] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations," 2023, *arXiv:2301.11699*.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

20   IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

[109] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022.

[110] Z. Chen et al., "Hierarchical integration diffusion model for realistic image deblurring," 2023, *arXiv:2305.12966*.

[111] 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, Switzerland: Springer, 2023.

[112] T. Guo, H. S. Mousavi, T. H. Vu, and V. Monga, "Deep wavelet prediction for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1100–1109.

[113] B. B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, "Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2024, pp. 1–8.

[114] Y. Huang et al., "WaveDM: Wavelet-based diffusion models for image restoration," 2023, *arXiv:2305.13819*.

[115] F. Guth, S. Coste, V. De Bortoli, and S. Mallat, "Wavelet score-based generative modeling," in *Proc. NeurIPS*, vol. 35, 2022.

[116] S. Shang et al., "ResDiff: Combining CNN and diffusion model for image super-resolution," 2023, *arXiv:2303.08714*.

[117] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16272–16282.

[118] Z. Yue, J. Wang, and C. C. Loy, "ResShift: Efficient diffusion model for image super-resolution by residual shifting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024.

[119] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. NeurIPS*, vol. 35, 2022.

[120] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," 2021, *arXiv:2108.02938*.

[121] A. Niu et al., "CDPMSR: Conditional diffusion probabilistic models for single image super-resolution," 2023, *arXiv:2302.12831*.

[122] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[123] 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.

[124] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "RankSRGAN: Generative adversarial networks with ranker for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3096–3105.

[125] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2434–2442.

[126] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

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

[128] C. Bi, X. Luo, S. Shen, M. Zhang, H. Yue, and J. Yang, "DeeDSR: Towards real-world image super-resolution via degradation-aware stable diffusion," 2024, *arXiv:2404.00661*.

[129] S. Zhou, K. Chan, C. Li, and C. C. Loy, "Towards robust blind face restoration with codebook lookup transformer," in *Proc. NeurIPS*, vol. 35, 2022.

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

[131] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, "SeeSR: Towards semantics-aware real-world image super-resolution," 2023, *arXiv:2311.16518*.

[132] Y. Qu et al., "XPSR: Cross-modal priors for diffusion-based image super-resolution," 2024, *arXiv:2403.05049*.

[133] J. Liang, H. Zeng, and L. Zhang, "Details or artifacts: A locally discriminative learning approach to realistic image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5647–5656.

[134] C. Chen et al., "Real-world blind super-resolution via feature matching with implicit high-resolution priors," in *Proc. 30th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2022.

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

[136] A. Bansal et al., "Cold diffusion: Inverting arbitrary image transforms without noise," 2022, *arXiv:2208.09392*.

[137] G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar, "I^2SB: Image-to-image Schrödinger bridge," 2023, *arXiv:2302.05872*.

[138] M. Delbracio and P. Milanfar, "Inversion by direct iteration: An alternative to denoising diffusion for image restoration," 2023, *arXiv:2303.11435*.

[139] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11462–11471.

[140] B. B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, "Dynamic attention-guided diffusion for image super-resolution," 2023, *arXiv:2308.07977*.

[141] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[142] B. Xia et al., "DiffIR: Efficient diffusion model for image restoration," 2023, *arXiv:2303.09472*.

[143] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017.

[144] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[145] B. Kawar, G. Vaksman, and M. Elad, "SNIPS: Solving noisy inverse problems stochastically," in *Proc. NeurIPS*, vol. 34, 2021.

[146] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. NeurIPS*, vol. 35, 2022.

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

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

[149] B. Fei et al., "Generative diffusion prior for unified image restoration and enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9935–9946.

[150] A. Shocher, N. Cohen, and M. Irani, "Zero-shot super-resolution using deep internal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3118–3126.

[151] R. Li, X. Sheng, W. Li, and J. Zhang, "OmniSSR: Zero-shot omnidirectional image super-resolution using stable diffusion model," 2024, *arXiv:2404.10312*.

[152] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11451–11461.

[153] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12403–12412.

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

[155] Y. Wang, Y. Hu, J. Yu, and J. Zhang, "GAN prior based null-space learning for consistent super-resolution," in *Proc. AAAI*, 2023, vol. 37, no. 3.

[156] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," in *Proc. NeurIPS*, vol. 35, 2022.

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

[158] J. Lin et al., "Adaptive multi-modal fusion of spatially variant kernel refinement with diffusion model for blind image super-resolution," 2024, *arXiv:2403.05808*.

[159] H. Chung, E. S. Lee, and J. C. Ye, "MR image denoising and super-resolution using regularized reverse diffusion," *IEEE Trans. Med. Imag.*, vol. 42, no. 4, pp. 922–934, Apr. 2023.

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

[161] G. Li, C. Rao, J. Mo, Z. Zhang, W. Xing, and L. Zhao, "Rethinking diffusion model for multi-contrast MRI super-resolution," 2024, *arXiv:2404.04785*.

[162] Z. Yue and C. Change Loy, "DifFace: Blind face restoration with diffused error contraction," 2022, *arXiv:2212.06512*.

[163] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9164–9174.

[164] T. Yang, P. Ren, X. Xie, and L. Zhang, "GAN prior embedded network for blind face restoration in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 672–681.

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

[166] Z. Wang et al., "DR2: Diffusion-based robust degradation remover for blind face restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1704–1713.

[167] X. Wang, S. López-Tapia, and A. K. Katsaggelos, "Atmospheric turbulence correction via variational deep diffusion," in *Proc. IEEE 6th Int. Conf. Multimedia Inf. Process. Retr. (MIPR)*, Sep. 2023, pp. 1–4.

[168] N. G. Nair, K. Mei, and V. M. Patel, "AT-DDPM: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models," in *Proc. WACV*, 2023.

[169] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.

[170] 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 Sens.*, vol. 14, no. 19, p. 4834, Sep. 2022.

[171] 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 Sens.*, vol. 15, no. 9, p. 2346, Apr. 2023.

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

[173] S. Khanna et al., "DiffusionSat: A generative foundation model for satellite imagery," in *Proc. ICLR*, 2024.

[174] D. Ganguli et al., "Predictability and surprise in large generative models," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022.

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

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

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

[178] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[179] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1040–1049.

[180] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "DipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[181] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[182] B. Jing, G. Corso, R. Berlinghieri, and T. Jaakkola, "Subspace diffusion generative models," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022.

[183] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," in *Proc. ICLR*, 2023.

[184] Q. Wu et al., "Uncovering the disentanglement capability in text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1900–1910.

[185] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2416–2425.

Brian B. Moser received the M.Sc. degree in computer science from TU Kaiserslautern, Kaiserslautern, Germany, in 2021, where he is currently pursuing the Ph.D. degree.

He is also a Research Assistant with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His research interests include image super-resolution and deep learning.

Arundhati S. Shanbhag is currently pursuing the master's degree with TU Kaiserslautern, Germany.

She is also a Research Assistant with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. Her research interests include computer vision and deep learning.

Federico Raue received the M.Sc. degree in artificial intelligence from Katholieke Universiteit Leuven, Leuven, Belgium, in 2005, and the Ph.D. degree from TU Kaiserslautern, Germany, in 2018.

He is currently a Senior Researcher with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His research interests include meta-learning and multimodal machine learning.

Stanislav Frolov received the M.Sc. degree in electrical engineering from Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2017. He is currently pursuing the Ph.D. degree with TU Kaiserslautern, Germany.

He is also a Research Assistant with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His research interests include generative models and deep learning.

Sebastian Palacio is currently a Researcher of machine learning and the Head of the Multimedia Analysis and Data Mining Group, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His Ph.D. topic was about explainable AI with applications in computer vision. His research interests include adversarial attacks, multitask, curriculum, and self-supervised learning.

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