Diffusion Models for Image Super-Resolution

[1]Jemin Kava, [2]Dhwanil Raval, [3]Japan M. Mavani, [4]Dhaval R. Chandarana
[1]jeminkava915@gmail.com, [2]dhwanil.raval15@gmail.com, [3]jmmavani@gmiu.edu.in, [4]drchandarana@gmiu.edu.in

Abstract- Diffusion-based generative models have recently achieved remarkable success in single image super-resolution (ISR), producing high-fidelity, perceptually convincing high-resolution images from low-resolution inputs. These models, derived from Denoising Diffusion Probabilistic Models (DDPMs) and related formulations, address limitations of earlier CNN- and GAN-based approaches by generating rich textures and fine details aligned with human visual preferences. In this paper, we review the practical applications of diffusion models in ISR, focusing on their architectures, training strategies, and performance relative to traditional methods. We discuss prominent models such as DDPM-based upscalers, the SR3 approach for iterative refinement, and subsequent improvements. Experimental results from recent studies are analyzed, comparing diffusion-driven ISR with convolutional and GAN-based methods on standard benchmarks. Diffusion models consistently excel in perceptual quality (often achieving lower Fréchet Inception Distance and Learned Perceptual Similarity) and human evaluation fool rates, despite sometimes lower PSNR/SSIM metrics than optimized CNN/GAN methods. We include example comparisons, quantitative tables, and discuss the trade-offs in complexity and inference speed. The paper is organized as a standard academic report with sections covering the background, methodology, experimental evaluation, and a discussion on results and future directions in diffusion-based ISR.

I. INTRODUCTION

Image Super-Resolution (SR) is the task of reconstructing a high-resolution (HR) image from a given low-resolution (LR) input. This longstanding challenge is inherently ill-posed, as a single LR image can correspond to multiple plausible HR images differing in fine details or textures. SR has wide-ranging applications from enhancing everyday photography to improving satellite imagery and medical imaging. Traditional SR approaches included interpolation techniques (e.g. bicubic upsampling) and later, machine learning methods such as sparse coding and neighbor embedding. The advent of deep learning brought convolutional neural network (CNN) based SR models like SRCNN (2014) and its successors, which significantly improved reconstruction

fidelity by learning end-to-end mappings from LR to HR. These CNN-based methods optimized pixel-wise losses (e.g. mean squared error) to maximize Peak Signal-to-Noise Ratio (PSNR), but often produced overly smooth images lacking high-frequency details.

Generative models introduced new paradigms for SR by targeting perceptual quality. Generative Adversarial Networks (GANs), starting with SRGAN (Ledig et al. 2017), trained a generator to produce sharper, more realistic textures using an adversarial loss and perceptual loss (feature reconstruction). Enhanced GAN variants like ESRGAN further improved realism by architectural changes (Residual-in-Residual blocks, etc.) and improved training techniques. GAN-based SR methods demonstrated dramatic visual improvements, yielding much sharper outputs than pure CNN regression; however, GANs can suffer from training instabilities, mode collapse, and occasional unnatural artifacts. Careful regularization and tuning are required to keep GAN generators stable and prevent hallucinating details.

Diffusion models have recently emerged as a powerful alternative for image generation and have disrupted the SR field, closing the gap between algorithmic output and human perceptual preferences. Diffusion Models (DMs) generate images by iteratively denoising from pure noise, guided by a learned distribution. This approach has proven capable of producing extremely high-quality, detailed images that often exceed the realism of previous methods. Importantly, diffusion models avoid some pitfalls of GANs: their training objective (typically a simple denoising regression) is more stable and mode-covering, meaning DMs capture a wide range of plausible details rather than collapsing to a single solution. The introduction of diffusion-based SR (beginning around 2021) marked a significant shift, challenging the long-standing dominance of GANs. For instance, Saharia *et al.* (2021) proposed SR3 (Super-Resolution via Repeated Refinement), adapting a DDPM to conditional image generation for SR. SR3 demonstrated photo-realistic 8× upsampling on faces and natural images, achieving human evaluation fool rates of nearly 50% (i.e. human raters confuse the SR output for a real HR image about half the time), whereas prior state-of-the-art GAN outputs fooled humans only ~34% of the time. Indeed, human raters perceive diffusion-based SR results as more realistic than those produced by GANs. Diffusion models have thus ushered in a new era for SR, enabling generation of high-fidelity textures and details that align closely with human perceptual judgments.

Despite these successes, diffusion ISR methods are not without challenges. They demand high computational cost for training and inference due to the iterative sampling procedure, and typically employ large networks to achieve their quality gains. Early diffusion upsamplers could be orders of magnitude slower than single-pass CNN or GAN models. Other noted issues include color inconsistencies (e.g. slight color shifts in reconstructed images) and the inherent randomness in the generative process which can lead to variability in outputs. This paper provides a comprehensive review of diffusion models for image super-resolution, with emphasis on practical architectural and training insights, and a comparative evaluation against traditional CNN-based and GAN-based SR methods. In the following, we first survey related work in SR and diffusion modeling (Section Related Work). We then outline the methodology of diffusion-based SR, describing the typical architectures and training procedures (Section Methodology). In Section

Experiments, we summarize experimental setups and benchmarks commonly used to evaluate SR models. Section Results and Discussion presents quantitative and qualitative comparisons between diffusion models and prior approaches, including tables of performance metrics (PSNR, SSIM, FID, LPIPS) and visual examples. We discuss the implications of these results, current limitations, and improvements such as hybrid diffusion-GAN techniques. Finally, Section Conclusion concludes the paper and suggests future directions in diffusion-based super-resolution.

II.RELATED WORK

Deep CNN-based Super-Resolution: The modern era of SR began with deep learning models. SRCNN by Dong *et al.* (2014) was the first CNN for SR, demonstrating that a three-layer CNN could learn an end-to-end mapping from LR to HR and outperform earlier interpolation or dictionary-based methods. Subsequent networks like VDSR (very deep SR network), EDSR (enhanced deep SR, which removed batch norm to push PSNR higher) and RCAN (residual channel attention network) achieved progressively better reconstruction fidelity on benchmarks such as Set5, Set14 and DIV2K. These models optimized pixel-wise losses (L1 or L2), focusing on minimizing distortion metrics (PSNR/SSIM). While they achieved impressive numerical accuracy, the outputs tended to be overly smooth and lacked fine textures, a consequence of the objective which favors averaging to reduce pixel error. This over-smoothing prompted research into perceptual quality optimization.

GAN-based Super-Resolution: GANs introduced by Goodfellow *et al.* revolutionized image generation and were applied to SR to enhance perceptual realism. SRGAN first incorporated an adversarial loss (a discriminator trained to distinguish real HR images from generated images) along with a perceptual loss (measuring feature differences using a pre-trained network) to encourage the generator to produce sharper and more detailed images instead of optimizing only MSE. SRGAN's outputs were much more photo-realistic than previous CNN outputs, albeit with occasional artifacts. ESRGAN (Wang *et al.* 2018) improved upon SRGAN by introducing the Residual-in-Residual Dense Block (RRDB) architecture and a relativistic GAN loss, further improving detail generation and reducing artifacts. GAN-based SR methods dominated perceptual-quality SR for several years, and numerous variants addressed issues like stability and artifact reduction. However, GANs require careful balancing of generator and discriminator during training; otherwise one can observe phenomena like mode collapse or checkerboard artifacts. Additionally, GAN-trained SR models often sacrifice some fidelity (PSNR) in exchange for realism, as the generator may hallucinate details that do not exactly match the ground truth.

Diffusion Models for Super-Resolution: Diffusion models are a class of generative models that define a forward process of gradually adding noise to an image and a learned reverse process to remove noise and retrieve a clean image. Denoising Diffusion Probabilistic Models (DDPM) introduced by Ho *et al.* (2020) and score-based generative models by Song *et al.* (2020) demonstrated that iterative denoising approaches could generate images of excellent quality rivaling GANs. The adaptation of diffusion models to conditional image generation enabled their

use in ISR. In 2021, Saharia *et al.* introduced SR3, which conditions a diffusion model on the LR image to progressively super-resolve it. SR3's approach of iterative refinement showed outstanding results on face super-resolution (e.g. $16\times16 \to 128\times128$) and general natural images, significantly outperforming GAN baselines in human evaluations. Around the same time, Li *et al.* (2022) proposed SRDiff, another diffusion-based SR model, claiming to be one of the first diffusion models for general single-image SR. These works established diffusion as a compelling approach for SR, combining the strengths of classical regression (fidelity to input) with generative ability to synthesize realistic details. Subsequent research has produced numerous variants and enhancements. Nichol & Dhariwal (2021) introduced improved sampling techniques and classifier-guided diffusion, which can be adapted to further improve conditional generation quality. More recently, researchers have explored latent diffusion (performing the diffusion process in a lower-dimensional latent space to speed up inference) and cascaded diffusion (using multiple diffusion models in series for very large upscaling factors or high-resolution outputs). For example, StableSR leveraged a pre-trained text-to-image diffusion model (Stable Diffusion) as a prior for SR, achieving rich texture generation by guiding the SR process with the latent knowledge of a large diffusion model. Diffusion-based SR has also been applied in specialized domains such as face restoration, where DMs help recover facial details from heavily degraded inputs, and in remote sensing, where satellite images are upscaled with diffusion models to improve downstream analysis[1]. A comprehensive survey by Moser *et al.* (2024) catalogs these developments, noting that diffusion models have become a dominant paradigm in SR research, consistently ranking among state-of-the-art methods.

III.METHODOLOGY

Diffusion Process for SR: At the core of diffusion-based SR is a conditional denoising process. The model defines a forward diffusion (or degradation) process that adds Gaussian noise to a high-resolution image through T time steps, gradually destroying its structure. In SR, this forward process is conceptually applied to the *target* HR distribution. The reverse diffusion process is a Markov chain that starts from pure noise and iteratively denoises to recover an HR image, conditioned on the given LR image. Formally, let y be the desired HR image and x be the observed LR image. We define a forward noising sequence $y_0 = y$ (clean HR), $y_T \sim \mathcal{N}(0,I)$ (pure noise), and y_{t} is obtained by adding a small Gaussian noise to y_{t-1} at each step. The reverse model approximates $p_{\theta}(y_{t-1}|y_t, x)$ – the distribution of the denoised image at step $t-1$ given the noisy image at step t and the conditioning LR. The neural network (often a U-Net) is trained to predict either the noise ϵ added at each step or the clean image y_0 from a partially noised input. The training objective typically minimizes a reweighted variational bound or a simple mean-squared error between the network's predicted noise and the true noise added, as derived by Ho *et al.* (2020). In practice, this reduces to a straightforward L_2 loss between the model's output and the known noise, averaged over random timesteps and training samples. This training procedure is stable and

does not require an adversarial discriminator; it essentially teaches the model to *denoise* an image when a specific amount of noise has been applied, while utilizing the LR image as context.

Conditioning on the LR Image: A critical aspect of diffusion ISR models is how the low-resolution image is provided as conditioning input. A common strategy, as used in SR3, is to concatenate the LR image (usually upsampled to the HR size via simple interpolation) with the noisy image at each diffusion step, channel-wise, and feed this as input to the U-Net model. The U-Net then has access to the fixed LR guidance at every denoising step, ensuring the output remains consistent with the large-scale structure of the input. Another approach is to provide the LR image as additional input through an encoder branch. For example, Wang & Zhou (2024) augment the diffusion model with an LR encoder network that extracts features from the LR image; these features are injected into the denoising U-Net via cross-attention or concatenation at multiple layers. Such mechanisms aim to address the challenge of inadequate conditional information – i.e., ensuring the model fully utilizes the LR image to place generated details in correct correspondence with the input content. In the Diffusion Architecture for Large Scale Super-resolution (DiffALS) model for remote sensing, Li *et al.* fuse features from a pre-trained CNN on the LR image into the early layers of the U-Net denoiser, specifically adding the LR feature maps into the first two residual blocks of the contracting path. This helps the model align the generated high-frequency details with the structures present in the LR input. Overall, modern diffusion SR architectures extensively use conditioning mechanisms (concatenation, feature fusion, or cross-attention) to guide the denoising process with the LR image.

Network Architecture: The backbone of most diffusion-based SR models is a U-Net architecture inspired by the original DDPM work. The U-Net typically consists of a series of downsampling convolutional blocks, a bottleneck, and corresponding upsampling blocks with skip connections between matching levels. In SR applications, certain architectural modifications have been found beneficial. The SR3 model's architecture, for instance, adopts residual blocks from BigGAN in place of standard ResNet blocks, and uses group normalization in each block. Skip connection rescaling (by $1/\sqrt{2}$) is applied to stabilize training when layers are deep. SR3 also increases the model capacity (more feature channels and more residual blocks per scale) compared to the original DDPM, to effectively model the distribution of high-resolution images. Many diffusion SR models use position embeddings (usually sinusoidal) to encode the timestep t and incorporate this via addition to feature maps or through FiLM-like modulation in each residual block. This timestep embedding is crucial so that the network knows the current noise level and can modulate its denoising strength accordingly. Modern designs may also integrate attention mechanisms. For example, some approaches include self-attention layers at the lowest resolution features to capture global image context (as done in Imagen and Latent Diffusion models). Others, like the Enhanced Diffusion Model by Wang & Zhou (2024), use specialized blocks (ENAFBlocks) that combine efficient channel attention and gating mechanisms to improve performance in noise prediction. These architectural innovations aim to improve the model's capacity to remove noise while preserving fidelity to the conditioning image.

It's worth noting that diffusion SR models tend to be *parameter-heavy*. The SR3 model for $64\to256$ natural image SR had on the order of 155 million parameters, significantly larger than typical CNN or GAN SR models, which often have tens of millions. This large capacity is one factor in their success at generating detailed textures, but it also contributes to higher computational load.

Training and Sampling Strategies: Training a diffusion SR model involves drawing random timesteps $t \sim \text{Uniform}({1,\dots,T})$ each iteration, adding the corresponding amount of noise to a ground-truth HR image y, and training the network to predict the noise (or denoised image) given y_t and the LR condition x. DDPM-style training uses a simplified loss that is equivalent to optimizing the variational lower-bound and has been effective in practice. Some works introduce hybrid losses to further guide training. For instance, DiffALS adds an adversarial loss term by training a separate discriminator that assesses whether the denoising trajectory is realistic. In DiffALS, the discriminator (called a Noise Discriminator) looks at pairs of consecutive noisy images (y_{t}, y_{t-1}) and tries to distinguish real pairs (from the true diffusion of a real image) vs fake pairs (from the model's output). The generator (denoiser) then gets an additional loss for fooling this discriminator, which encourages it to produce noise removal steps that result in more realistic textures. Such adversarial augmentation can improve perceptual sharpness of diffusion outputs, combining the strengths of GANs with diffusion's robust framework. Another training consideration is classifier-free guidance, a technique wherein the diffusion model is trained sometimes with the condition (LR image) and sometimes without, and at inference one can interpolate between the conditional and unconditional predictions to trade off adherence to the LR input vs. output sharpness. Classifier-free guidance has been employed in some image-to-image diffusion contexts; in SR specifically, most works aim for strict fidelity to the LR input, so guided sampling is less common except in tasks like artistic upscaling or where some flexibility is desired. During inference (sampling), one starts from Gaussian noise and iteratively applies the learned denoiser T times to obtain an output. The number of diffusion steps T (often 50, 100, or even 1000) crucially impacts the runtime. Faster sampling schemes (e.g. DDIM deterministic sampling or progressive distillation) can reduce the needed steps without severely compromising quality – these have been explored in the literature to make diffusion more practical for ISR. Additionally, techniques like adaptive step size or early stopping have been investigated, but generally diffusion models remain slower than single-pass methods. We discuss this trade-off in the Results section.

IV.EXPERIMENTS

To evaluate diffusion-based SR models against traditional approaches, researchers have conducted experiments on standard SR benchmarks and some specialized datasets. Common evaluation datasets include DIV2K (a high-quality dataset of natural images with provided train/test splits for 4× SR), Urban100 (a set of urban scene images with rich textures/lines), and older sets like Set5, Set14, and BSD100 for classical SR evaluations. For face super-resolution, tests are often done on CelebA-HQ (high-quality celebrity face images) at tasks like $16^2 \to 128^2$ or $64^2 \to

256^2$. Domain-specific datasets also appear in literature, e.g. remote sensing images (satellite imagery upscaled for better resolution) and medical images (where resolution enhancement can aid diagnostics).

Metrics: Experiments typically report distortion metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), which gauge how closely the super-resolved image matches the ground truth pixel-wise. Higher PSNR/SSIM indicate better fidelity. However, as discussed, these metrics often do not correlate well with human visual preference, especially for high upscaling factors; a blurry but artifact-free image can score high PSNR, whereas a sharp, detailed image with slight texture variations from the true image can score lower. Therefore, perceptual quality metrics are critical. Many works report LPIPS (Learned Perceptual Image Patch Similarity), an error measured in a deep feature space (lower LPIPS means the output is perceptually closer to the reference). Another is the Fréchet Inception Distance (FID), which compares the statistics of a set of generated images to real images – lower FID indicates the output distribution is closer to the real distribution. Human evaluation studies (often in the form of A/B preference tests or confusion tests) are also conducted to directly measure perceptual quality. For example, a two-alternative forced choice (2AFC) test might ask human observers to choose which image is real between a model's output and a ground truth photo.

Baselines: In experiments, diffusion models are compared with both CNN-based SR methods (optimized for PSNR) and GAN-based SR methods (optimized for perceptual quality). A typical baseline for distortion-oriented SR is an EDSR or RRDB model trained with L_1 loss (sometimes referred to as a "Regression" model or RCAN, etc.). For perceptual baselines, SRGAN/ESRGAN or more recent variants like Real-ESRGAN (for real-world SR) are used. Notably, some studies include classical interpolation (bicubic) as a low-end baseline to highlight the improvements.

Training Setup: Diffusion SR models are exceedingly resource-intensive to train. For instance, SR3's training to super-resolve $64 \to 256$ images involved hundreds of GPU-hours and large batch sizes. Most experiments in literature train diffusion models with $T\approx1000$ noise steps and then often use a smaller number of inference steps with a sampler like DDIM. Some works have tried to reduce training cost by using fewer diffusion steps or by transfer learning from pre-trained generative models. For instance, StableSR fine-tuned a stable diffusion model (pre-trained on generative tasks) for SR, rather than training from scratch. Diffusion models can also be cascaded: train a model to do 2× or 4× SR, then apply it iteratively or in a chain to achieve 8×, 16×, etc., which is how SR3 handled 16× upscaling (by chaining a $16\to128$ and a $128\to512$ model).

In all experiments, visual comparison is key. Papers often include side-by-side image patches from different methods to qualitatively assess sharpness, noise, and artifact levels. We include such a comparison in the next section to illustrate the differences.

V.RESULTS AND DISCUSSION

Visual comparison of super-resolution outputs for a face image (top) and a natural image (leopard, bottom). Columns from left to right: the low-resolution input upscaled via Bicubic interpolation; output of a CNN "regression" model optimized for PSNR (producing a smooth but somewhat blurred result); output of a diffusion model (SR3) showing much finer detail; and the ground truth high-resolution image for reference. The diffusion-based SR result provides realistic textures (e.g. sharper eyes and fur) more closely resembling the ground truth, whereas the PSNR-oriented model, while faithful in overall color/structure, lacks high-frequency detail.

The qualitative examples above highlight the key strength of diffusion-based SR: the ability to synthesize realistic textures that traditional methods either blur out or cannot reconstruct. In the face example, the bicubic and CNN outputs are smooth and lose details in the eyes and skin, while the diffusion model restores crisp details (pores, eyelashes) that make the image look photo-realistic. Similarly, for the leopard, the diffusion output has clearly defined fur patterns compared to the smeared appearance from bicubic/CNN. These details contribute to much higher perceptual fidelity. As a result, images enhanced by diffusion models tend to fool human observers more often into thinking they are real. A human study by Saharia *et al.* reported nearly 50% confusion rate on 8× face SR for their SR3 model, versus only ~15–34% for GAN or CNN-based models. In other words, people were about as likely to mistake SR3's output for a real HR image as they were to identify the actual ground truth, a remarkable achievement for super-resolution.

Human evaluation of super-resolution models: the confusion rate (percentage of times raters mistook the SR output for the real photo) for different methods. Top: On $16\times16 \to 128\times128$ face SR (CelebA-HQ), SR3 achieved ~47% confusion, dramatically higher than GAN-based FSRGAN (8.5%), an optimization-based method PULSE, or a regression CNN. Bottom: On $64\times64 \to 256\times256$ natural image SR, SR3 reached ~39% confusion vs ~13% for a regression model. These results demonstrate the superior perceptual realism of diffusion-generated SR images.

Quantitatively, diffusion models often exhibit a trade-off between distortion metrics and perceptual metrics when compared to other methods. Traditional CNN models (e.g. RRDB with L_1 loss) typically achieve the highest PSNR/SSIM because they minimize pixel error, but their outputs can be perceptually inferior. Diffusion models, by contrast, prioritize producing realistic texture, which can introduce slight pixel discrepancies. For example, on the DIV2K benchmark (4× SR), a regression-based model might achieve PSNR around 28–29 dB, whereas diffusion models like SR3, SRDiff or others yield slightly lower PSNR (in the 26–27 dB range). In one experiment, an L_1-trained RRDB achieved 28.98 dB PSNR / 0.83 SSIM on DIV2K (with relatively high LPIPS 0.27), while SR3 obtained 26.17 dB / 0.68 SSIM (and LPIPS 0.24). However, the perceptual scores tell the other side of the story: SR3's FID was 33.87, dramatically better (lower) than RRDB's 78.55. Similarly, LPIPS for SR3 (0.24) was lower than the CNN's (0.27), indicating closer perceptual similarity to ground truth. This reflects the well-known observation that PSNR and SSIM do not fully reflect subjective image quality. Diffusion outputs may diverge from the exact ground truth pixels (lowering PSNR) but appear more realistic to the

eye (as evidenced by lower LPIPS/FID and human preference). Indeed, researchers caution against relying solely on PSNR/SSIM for high-factor SR, since these metrics favor overly smooth reconstructions over those with realistic texture.

It is informative to compare diffusion models with GAN-based SR in these metrics as well. GAN-based methods like ESRGAN often strike a middle ground: they improve perceptual scores over pure CNNs (with lower LPIPS than CNN, and better visual fidelity) but can still fall short of diffusion in realism. For instance, one study on $16\times$ face SR found a GAN (FSRGAN) had LPIPS ~0.135 and FID ~43.8, whereas diffusion (SR3) achieved LPIPS ~0.097 and FID ~35.1 – significantly better perceptual quality. Notably, the GAN's PSNR (23.01 dB) was slightly *lower* than SR3's (23.04 dB) in that case, indicating that both methods sacrifice some pixel fidelity for perceptual gains, but diffusion was able to produce more convincing details without further lowering PSNR. In general, diffusion models tend to deliver state-of-the-art perceptual performance on SR tasks; for example, SR3 and SRDiff outputs in studies are often ranked best or among the best by human evaluators[1]. Even in specialized domains like remote sensing, diffusion SR models produced visually sharper and more useful results for downstream tasks (like object segmentation) than GAN or CNN models, despite slightly lower PSNR[2]. Li *et al.* report that their diffusion model SRDiff outperformed a GAN (ESRGAN) and a CNN (NLSN) on segmentation accuracy when using the super-resolved images, suggesting the extra textures generated were genuinely beneficial and not just hallucinated nonsense[1]. This underscores that diffusion SR models are not merely creating "fake detail," but often adding plausible, context-appropriate detail that can improve real-world utility.

One important disadvantage of diffusion-based SR is the computational burden. The iterative denoising process is inherently slower than feed-forward upscaling. As a concrete example, in the $4\times$ SR experiment on DIV2K, the diffusion model SR3 took an average of 318.8 seconds per image (on a certain hardware setup), whereas a CNN like RRDB took only 1.6 seconds. Even more efficient diffusion models from recent work report on the order of several seconds per image for moderate resolution outputs. This large gap in inference speed and throughput has been a barrier for deploying diffusion SR in real-time or high-volume scenarios. Moreover, diffusion models are memory-hungry; their U-Net backbone operating at high resolution with many feature channels can consume a lot of GPU memory (e.g. hundreds of MB for a single image). Researchers are actively exploring solutions to this. Some approaches involve model distillation, compressing the multi-step sampling into a smaller number of steps (e.g. 10 steps) by training the model to jump larger intervals – this can dramatically accelerate inference at some cost of quality. Others use latent-space diffusion, where the model operates on a lower-dimensional representation of the image (such as a VAE latent code) and then uses a lightweight decoder to produce the final HR output; this was popularized by Latent Diffusion Models and applied in tasks like Stable Diffusion's built-in upscaler. Such latent approaches can speed up computation by working on smaller spatial sizes. Another line of improvement is architectural efficiency: Wang & Zhou's 2024 model introduced an adaptive sampling strategy to process large images in patches (to fit in memory) and carefully blend them. They also managed to reduce parameter count (their model

~18M vs earlier diffusion models 35–155M) and still beat prior diffusion methods in quality. These developments indicate that efficiency is being addressed, though a gap remains. Indeed, a survey noted that *computational cost and slow sampling* are among the top challenges facing diffusion models in SR.

It is also insightful to consider whether diffusion models truly represent a fundamentally better approach, or if their strong results partly come from using larger models/training for longer. A recent controlled study by Kuznedelev *et al.* (2024) titled *"Does Diffusion Beat GAN in Image Super Resolution?"* explored this question. They found that when GAN and diffusion models are given comparable model capacity and training data, their results on SR tasks can be very similar[3]. In other words, a well-tuned GAN can achieve on par perceptual quality with a diffusion model if it's scaled appropriately, suggesting that diffusion's edge in some comparisons may have come from heavier resource usage. Nevertheless, diffusion models have an advantage in their reliability and mode coverage – they more easily incorporate uncertainty and diversity in outputs (multiple samples can be generated for the same LR input, reflecting different plausible textures). GANs typically produce one deterministic output and may miss some modes of the solution space unless explicitly designed for diversity. Diffusion models also avoid typical GAN failure modes; one rarely sees training collapse with diffusion – at worst, more training just yields gradually better sample quality. This robustness is a practical benefit. Moreover, diffusion frameworks easily allow guidance (e.g. using classifier guidance or textual prompts to steer generation), which opens up new possibilities like text-conditioned SR or style-conditioned SR. While the baseline diffusion vs. GAN gap may narrow under fair conditions, diffusion remains highly attractive due to its flexibility and proven performance on extremely challenging cases (like 16× super-resolution or generating realistic human faces from tiny thumbnails).

In summary, our analysis finds that diffusion-based super-resolution methods deliver state-of-the-art perceptual quality, generating images that often appear indistinguishable from true high-resolution photos. They resolve textures and details that CNN methods cannot, and even GANs struggle to match the natural look of diffusion outputs. Quantitatively, they achieve lower FID and LPIPS (better perceptual scores) and high human preference, though their PSNR/SSIM might be slightly lower than ultra-fidelity methods. The choice between a diffusion model and a traditional model may thus depend on the application: for tasks where photorealism is paramount (e.g. enhancing images for human viewing, artistic applications, or downstream tasks tolerant to slight pixel shifts), diffusion models are clearly superior. On the other hand, if one requires pixel-accurate reconstruction (e.g. perhaps in some scientific imaging contexts or text super-resolution where exact shapes must be recovered), a CNN approach might still be relevant or one might combine diffusion with additional constraints. There are also hybrid approaches being explored, such as using a first stage CNN for coarse fidelity and a second stage diffusion to add details.

VI.CONCLUSION

Diffusion models have rapidly become a cornerstone of cutting-edge image super-resolution research. By leveraging a stochastic refinement process, they effectively address the historic challenge of the SR task's ill-posed nature: instead of producing a single blurry estimate, diffusion models can generate realistic high-frequency details consistent with the input. In this paper, we discussed how models like DDPM and score-based diffusion have been adapted for ISR, from the seminal SR3 model through to recent variants that integrate advanced conditioning and hybrid losses. We reviewed architecture designs (U-Net backbones with BigGAN-style blocks, attention mechanisms, and encoders for condition), and training insights that make diffusion SR feasible. Through experimental comparisons, we highlighted that diffusion-based approaches offer exceptional perceptual quality gains over traditional CNN or GAN methods – yielding images that humans often prefer for their fidelity to real-world textures – at the cost of increased computation. Diffusion SR models tend to slightly underperform classical models on PSNR/SSIM, emphasizing that those metrics capture only one aspect of quality. Nonetheless, in contexts where visual realism is the goal, diffusion models currently represent the state of the art.

Looking forward, there are several important directions to further advance diffusion-based ISR. Improving efficiency is paramount: research into faster samplers, model compression, and performing diffusion in compact domains (wavelet, latent spaces, etc.) is ongoing. Progress here will determine how widely diffusion SR gets adopted in real-time applications. Another direction is controllability – enabling user or algorithmic control over the output (for example, blending between a high-fidelity but low-detail result and a lower-fidelity but highly detailed result, or incorporating textual guidance to add plausible details like "increase the sharpness of text in image"). Diffusion frameworks are well-suited for such control via guidance techniques. Additionally, handling of real-world degradation (beyond simple bicubic down sampling) is a practical extension: recent works combine diffusion models with unsupervised degradation modeling to tackle real photographs where the downscale process is unknown. This remains challenging, but diffusion's ability to model uncertainty can be advantageous for capturing a distribution of possible clean images. We also foresee more domain-specific diffusion SR: as noted in the survey literature, applications in medical imaging, surveillance (e.g. enhancing faces or license plates), and scientific imaging (astronomy, microscopy) are emerging. Each of these domains can benefit from the high quality of diffusion-generated SR, possibly combined with domain knowledge (e.g. MRI physical models or anatomical priors for medical).

In conclusion, diffusion models have opened a new frontier for image super-resolution, achieving unprecedented levels of detail and realism. They complement and in many cases outperform GAN-based approaches, marking a shift in how researchers approach the SR problem. With continued research to mitigate their computational demands and enhance their control, diffusion-based ISR is poised to transition from research labs to practical deployment in the coming years, enabling anyone to enhance images in ways that were previously attainable only in imagination. The convergence of high-quality output and improved efficiency will ultimately determine the extent to which diffusion models revolutionize real-world super-resolution applications. The work

surveyed and presented here provides a foundation and inspiration for further innovations at this exciting intersection of generative modeling and image restoration.

REFERENCE

- [1] [2] Denoising Diffusion Probabilistic Model with Adversarial Learning for Remote Sensing Super-Resolution | MDPI https://www.mdpi.com/2072-4292/16/7/1219
- [3] Does Diffusion Beat GAN in Image Super Resolution?