Image Super-Resolution via Iterative Refinement

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi

使用DiffusionModel（DDPM）改进的超分辨率方法SR3，通过迭代细化在生成过程中加入低分辨率图像，提高模型的定向生成能力。尽管训练稳定且能应用于去噪、去雾等领域，但论文主要贡献在于展示了一种新的思路而非仅限于超分辨率性能提升。

Abstract—We present SR3, an approach to image Super-Resolution via Repeated Refinement. SR3 adapts denoising diffusion probabilistic models (Ho et al. 2020), (Sohl-Dickstein et al. 2015) to image-to-image translation, and performs super-resolution through a stochastic iterative denoising process. Output images are initialized with pure Gaussian noise and iteratively refined using a U-Net architecture that is trained on denoising at various noise levels, conditioned on a low-resolution input image. SR3 exhibits strong performance on super-resolution tasks at different magnification factors, on faces and natural images. We conduct human evaluation on a standard 8× face super-resolution task on CelebA-HQ for which SR3 achieves a fool rate close to 50%, suggesting photo-realistic outputs, while GAN baselines do not exceed a fool rate of 34%. We evaluate SR3 on a 4× super-resolution task on ImageNet, where SR3 outperforms baselines in human evaluation and classification accuracy of a ResNet-50 classifier trained on high-resolution images. We further show the effectiveness of SR3 in cascaded image generation, where a generative model is chained with super-resolution models to synthesize high-resolution images with competitive FID scores on the class-conditional 256×256 ImageNet generation challenge.

Index Terms—Image super-resolution, diffusion models, deep generative models

1 Introduction

挑战

Single-image super-resolution is the process of generating a high-resolution image that is consistent with an input low-resolution image. It falls under the broad family of image-to-image translation tasks, including colorization, in-painting, and de-blurring. Like many such inverse problems, image super-resolution is challenging because multiple output images may be consistent with a single input image, and the conditional distribution of output images given the input typically does not conform well to simple parametric distributions, e.g., a multivariate Gaussian. Accordingly, while simple regression-based methods with feedforward convolutional nets may work for super-resolution at low magnification ratios, they often lack the high-fidelity details needed for high magnification ratios.

Deep generative models have seen success in learning complex empirical distributions of images (e.g., [3], [4]). Autoregressive models [5], [6], variational autoencoders (VAEs) [7], [8], Normalizing Flows (NFs) [9], [10], and GANs [11], [12], [13] have shown convincing image generation results and benefited conditional tasks such as image super-resolution [14], [15], [16], [17], [18]. However, existing techniques often suffer from various limitations; autoregressive models are prohibitively expensive for high-resolution image generation, NFs and VAEs often yield sub-optimal sample quality, and GANs require carefully designed regularization and optimization tricks to tame optimization instability and mode collapse [19], [20], [21], [22].

- *Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, and Mohammad Norouzi are with Google Research, Brain Team, Toronto, ON M5H0B3, Canada. E-mail: {chitwanit, wchan212}@gmail.com, {jonathanho, salimans, mnorouzi}@google.com.*
- *David J. Fleet is with Google Research, Brain Team, Toronto, ON M5H0B3, Canada, and also with the University of Toronto, Toronto, ON M5T2S8, Canada. E-mail: davidfleet@google.com.*

Manuscript received 23 August 2021; revised 13 June 2022; accepted 13 July 2022. Date of publication 12 September 2022; date of current version 6 March 2023.
(Corresponding author: David J. Fleet.)
Recommended for acceptance by C.C. Loy.
Digital Object Identifier no. 10.1109/TPAMI.2022.3204461

We propose SR3 (Super-Resolution via Repeated Refinement), a new approach to conditional image generation, inspired by recent work on Denoising Diffusion Probabilistic Models (DDPM) [1], [23], and denoising score matching [1], [24]. SR3 works by learning to transform a standard normal distribution into an empirical data distribution through a sequence of refinement steps, resembling Langevin dynamics. The key is a U-Net architecture [25] that is trained with a denoising objective to iteratively remove various levels of noise from an image. We adapt DDPMs to image-to-image translation by proposing a simple effective modification to the U-Net architecture. In contrast to GANs, which require inner-loop maximization, we minimize a well-defined loss function. Unlike autoregressive models, SR3 uses a constant number of inference steps regardless of output resolution.

创新点

SR3 models work well across a range of magnification factors and input resolutions (e.g., see Fig. 1), and they can be cascaded, e.g., going from 64×64 to 256×256, and then to 1024×1024. Cascading models allows one to independently train several small models rather than a single large model with a high magnification factor. We find that cascaded models enable more efficient inference, since directly generating a high-resolution image requires a larger number of costly iterative refinement steps for the same quality. We also show that one can cascade an unconditional generative model with SR3 models to unconditionally generate high-fidelity images. We conduct experiments on the general domain of natural images, as well as the domain of face images.

Automated image quality scores like PSNR and SSIM do not reflect human preference well when the input resolution is low and the magnification factor is large (e.g., [14], [15], [17], [26], [27], [28]). These quality scores often penalize synthetic high-frequency details, such as hair texture, because synthetic details do not perfectly align with the original details. We therefore resort to human evaluation to compare the quality of super-resolution methods. We adopt a 2-alternative forced-choice (2AFC) paradigm in which human subjects are shown a low-resolution input and are required to

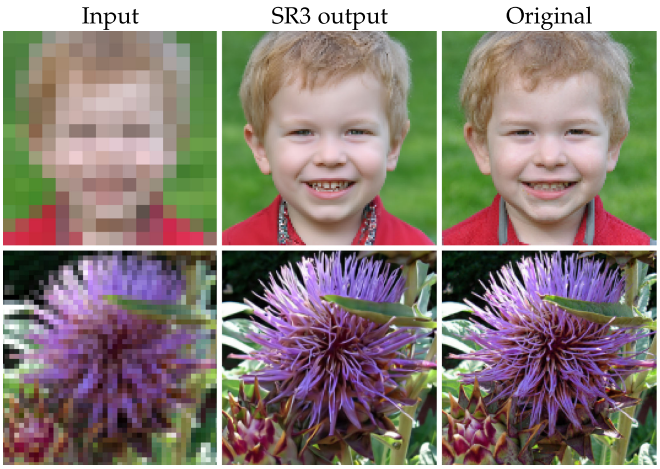


Fig. 1. Two representative SR3 outputs: (top) 8× face super-resolution at 16×16→128×128 pixels (bottom) 4× natural image super-resolution at 64×64→256×256 pixels.

select between a model output and a ground truth image (*cf.* [29]). With that data we calculate *fool rate* scores that capture both image quality and the consistency of model outputs with low-resolution inputs.

On a standard 8× face super-resolution task, SR3 achieves a human fool rate close to 50%, outperforming FSRGAN [14] and PULSE [17] with fool rates of at most 34%. On a 4× task on natural images, SR3 outperforms a ESRGAN [30] EnhanceNet [31] and SRFlow [32] on human evaluation, and a wide range of methods on classification accuracy of a ResNet-50 classifier trained on high-resolution images. To demonstrate unconditional and class-conditional generation we combine a 64×64 generative model with SR3 models to progressively generate 1024×1024 unconditional faces in 3 stages, and 256×256 class-conditional ImageNet samples in 2 stages, all with competitive FID scores.

2 Conditional Denoising Diffusion Model

We are given a dataset of input-output image pairs, denoted $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$, which represent samples drawn from an unknown distribution $p(\boldsymbol{x}, \boldsymbol{y})$. The conditional distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$ is a one-to-many mapping in which many target images may be consistent with a single source image. We are interested in learning a parametric approximation to $p(\boldsymbol{y} \mid \boldsymbol{x})$ through a *stochastic* iterative refinement process that maps a source image $\boldsymbol{x}$ to a target image $\boldsymbol{y} \in \mathbb{R}^d$. We approach this problem by adapting the denoising diffusion probabilistic (DDPM) model of [1], [23] to *conditional* image generation.

The conditional DDPM model generates a target image $\boldsymbol{y}_0$ in T refinement steps. Starting with a pure noise image $\boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the model iteratively refines the output image to attain a sequence $(\boldsymbol{y}_{T-1}, \boldsymbol{y}_{T-2}, \ldots, \boldsymbol{y}_0)$ according to learned conditional distributions $p_\theta(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{x})$ such that ultimately $\boldsymbol{y}_0 \sim p(\boldsymbol{y} \mid \boldsymbol{x})$ (see Fig. 2).

The distribution of intermediate images in the iterative refinement chain is defined in terms of a *forward* diffusion process that gradually adds Gaussian noise to the output via a fixed Markov chain, denoted $q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1})$. The goal of our model is to reverse the Gaussian diffusion process by iteratively recovering signal from noise through a reverse Markov chain conditioned on $\boldsymbol{x}$. We learn the reverse chain

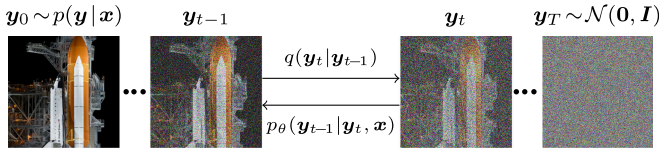


Fig. 2. The forward diffusion process q (left to right) gradually adds Gaussian noise to the target image. The reverse process p (right to left) iteratively denoises the target image, conditioned on a source image $\boldsymbol{x}$. (Source $\boldsymbol{x}$ is not shown.).

using a neural denoising model f_θ that takes as input a source image and a noisy target image and estimates the noise. In principle, each forward process step can be conditioned on $\boldsymbol{x}$ too, but we we find that a simple diffusion process that does not depend on $\boldsymbol{x}$ works reasonably well for super-resolution, and we leave extensions of the diffusion framework to future work.

2.1 Gaussian Diffusion Process

Following [1], [23], we define a *forward* Markovian diffusion process q that gradually adds Gaussian noise to a high-resolution image $\boldsymbol{y}_0$ over T iterations

$$q(\boldsymbol{y}_{1:T} \mid \boldsymbol{y}_0) = \prod_{t=1}^{T} q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}), \tag{1}$$

$$q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}) = \mathcal{N}(\boldsymbol{y}_t \mid \sqrt{\alpha_t}\,\boldsymbol{y}_{t-1}, (1-\alpha_t)\boldsymbol{I}), \tag{2}$$

where the scalar parameters $\alpha_{1:T}$ are hyper-parameters, subject to $0 < \alpha_t < 1$, which determine the variance of the noise added at each iteration. Note that $\boldsymbol{y}_{t-1}$ is attenuated by $\sqrt{\alpha_t}$ to ensure that the variance of the random variables remains bounded as $t \to \infty$. For instance, if the variance of $\boldsymbol{y}_{t-1}$ is 1, then the variance of $\boldsymbol{y}_t$ is also 1.

Importantly, one can characterize the distribution of $\boldsymbol{y}_t$ given $\boldsymbol{y}_0$ by marginalizing out the intermediate steps as

$$q(\boldsymbol{y}_t \mid \boldsymbol{y}_0) \;=\; \mathcal{N}(\boldsymbol{y}_t \mid \sqrt{\gamma_t}\,\boldsymbol{y}_0, (1-\gamma_t)\boldsymbol{I}), \tag{3}$$

where $\gamma_t = \prod_{i=1}^{t} \alpha_i$. Furthermore, with some algebraic manipulation and completing the square, one can derive the posterior distribution of $\boldsymbol{y}_{t-1}$ given $(\boldsymbol{y}_0, \boldsymbol{y}_t)$ as

$$\begin{aligned} q(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_0, \boldsymbol{y}_t) &= \mathcal{N}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \\ \boldsymbol{\mu} &= \frac{\sqrt{\gamma_{t-1}}\,(1-\alpha_t)}{1-\gamma_t}\,\boldsymbol{y}_0 + \frac{\sqrt{\alpha_t}\,(1-\gamma_{t-1})}{1-\gamma_t}\,\boldsymbol{y}_t \\ \sigma^2 &= \frac{(1-\gamma_{t-1})(1-\alpha_t)}{1-\gamma_t}. \end{aligned} \tag{4}$$

This posterior is helpful when parameterizing the reverse (generative) process and formulating a variational lower bound on the data log-likelihood of the reverse process.

2.2 Optimizing the Denoising Model

The key to inference with diffusion models (Section 2.3) is the denoising network. In our case it is conditioned on side information in the form of a source image $\boldsymbol{x}$. More formally, we optimize a neural denoising model f_θ that takes as input this source image $\boldsymbol{x}$ and a noisy target image $\widetilde{\boldsymbol{y}}$,

$$\widetilde{\boldsymbol{y}} = \sqrt{\gamma}\,\boldsymbol{y}_0 + \sqrt{1-\gamma}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{5}$$

and aims to recover the noiseless target image $\boldsymbol{y}_0$. This definition of a noisy target image $\widetilde{\boldsymbol{y}}$ is compatible with the marginal distribution of noisy images at different steps of the forward diffusion process (3).

Algorithm 1. Training a Denoising Model f_θ

1: **repeat**
2: $(\boldsymbol{x}, \boldsymbol{y}_0) \sim p(\boldsymbol{x}, \boldsymbol{y})$
3: $\gamma \sim p(\gamma)$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take a gradient descent step on $\nabla_\theta \| f_\theta(\boldsymbol{x}, \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \gamma) - \boldsymbol{\epsilon} \|_p^p$
6: **until** converged

Algorithm 2. Inference in T Iterative Refinement Steps

1: $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
4: $\boldsymbol{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)) + \sqrt{1-\alpha_t}\boldsymbol{z}$
5: **end for**
6: **return** $\boldsymbol{y}_0$

In addition to a source image $\boldsymbol{x}$ and a noisy target image $\widetilde{\boldsymbol{y}}$, the denoising model $f_\theta(\boldsymbol{x}, \widetilde{\boldsymbol{y}}, \gamma)$ takes as input the sufficient statistics for the variance of the noise γ, and is trained to predict the noise vector $\boldsymbol{\epsilon}$. Thus, the denoising model is aware of the level of noise through conditioning on γ, similar to [24], [33]. The proposed objective function for training f_θ is

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{\boldsymbol{\epsilon},\gamma}\left\| f_\theta(\boldsymbol{x}, \underbrace{\sqrt{\gamma}\,\boldsymbol{y}_0 + \sqrt{1-\gamma}\,\boldsymbol{\epsilon}}_{\widetilde{\boldsymbol{y}}}, \gamma) - \boldsymbol{\epsilon} \right\|_p^p, \tag{6}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $(\boldsymbol{x}, \boldsymbol{y})$ is sampled from the training dataset, $p \in \{1, 2\}$, and $\gamma \sim p(\gamma)$. The distribution of γ has a major impact on the quality of the model and the generated outputs. We discuss our choice of $p(\gamma)$ in Section 2.5.

Instead of regressing the output of f_θ to $\boldsymbol{\epsilon}$, as in (6), one can also regress the output of f_θ to $\boldsymbol{y}_0$. Given γ and $\widetilde{\boldsymbol{y}}$, the values of $\boldsymbol{\epsilon}$ and $\boldsymbol{y}_0$ can be derived from each other deterministically, but changing the regression target has an impact on the scale of the loss function. We expect both of these variants to work reasonably well if $p(\gamma)$ is modified to account for the scale of the loss function. Further investigation of the loss function for training the denoising model is an interesting area for future research (e.g., see [34]).

2.3 Inference via Iterative Refinement

Inference under our model is defined as a *reverse* Markovian process, which goes in the reverse direction of the forward diffusion process, starting from Gaussian noise $\boldsymbol{y}_T$

$$p_\theta(\boldsymbol{y}_{0:T}|\boldsymbol{x}) = p(\boldsymbol{y}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}) \tag{7}$$

$$p(\boldsymbol{y}_T) = \mathcal{N}(\boldsymbol{y}_T \,|\, \mathbf{0}, \boldsymbol{I}) \tag{8}$$

$$p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}_{t-1} \,|\, \mu_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t), \sigma_t^2 \boldsymbol{I}). \tag{9}$$

We define the inference process in terms of isotropic Gaussian conditional distributions, $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})$, which are learned. If the noise variance of the forward process steps are set as small as possible, i.e., $\alpha_{1:T} \approx 1$, the optimal reverse process $p(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})$ will be approximately Gaussian [23]. Accordingly, our choice of Gaussian conditionals in the inference process (9) can provide a reasonable fit to the true reverse process. Meanwhile, $1 - \gamma_T$ should be large enough so that $\boldsymbol{y}_T$ is approximately distributed according to the prior $p(\boldsymbol{y}_T) = \mathcal{N}(\boldsymbol{y}_T|\mathbf{0}, \boldsymbol{I})$, allowing the sampling process to start at pure Gaussian noise.

Recall that the denoising model f_θ is trained to estimate $\boldsymbol{\epsilon}$, given any noisy image $\widetilde{\boldsymbol{y}}$ including $\boldsymbol{y}_t$. Thus, given $\boldsymbol{y}_t$, we approximate $\boldsymbol{y}_0$ by rearranging the terms in (5) as

$$\hat{\boldsymbol{y}}_0 = \frac{1}{\sqrt{\gamma_t}}\left(\boldsymbol{y}_t - \sqrt{1-\gamma_t}\, f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)\right). \tag{10}$$

Following [1], we substitute our estimate $\hat{\boldsymbol{y}}_0$ into the posterior distribution of $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_0, \boldsymbol{y}_t)$ in (4) to parameterize the mean of $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})$ as

$$\mu_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)\right), \tag{11}$$

and we set the variance of $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})$ to $(1 - \alpha_t)$, a default given by the variance of the forward process [1].

Following this parameterization, each iteration of iterative refinement under our model takes the form,

$$\boldsymbol{y}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. This resembles one step of Langevin dynamics with f_θ providing an estimate of the gradient of the data log-density.

2.4 Justification of the Training Objective

Following Ho et al. [1], we justify the choice of the training objective in (6) for the probabilistic model outlined in (9) from a variational lower bound perspective. If the forward diffusion process is viewed as a fixed approximate posterior to the inference process, one can derive the following variational lower bound on the marginal log-likelihood

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_0)} \log p_\theta(\boldsymbol{y}_0|\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}_0}\mathbb{E}_{q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0)}\Bigg[\log p(\boldsymbol{y}_T) + \sum_{t\geq 1} \log \frac{p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})}{q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})}\Bigg]. \tag{12}$$

Given the particular parameterization of the inference process outlined above, one can show [1] that the negative variational lower bound can be expressed as the following simplified loss, up to a constant weighting of each term for each time step

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}_0,\boldsymbol{\epsilon}} \sum_{t=1}^{T} \frac{1}{T} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}, \sqrt{\gamma_t}\boldsymbol{y}_0 + \sqrt{1-\gamma_t}\boldsymbol{\epsilon}, \gamma_t) \right\|_2^2, \tag{13}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Note that this objective function corresponds to L_2 norm in (6), and a characterization of $p(\gamma)$ in terms of a uniform distribution over $\{\gamma_1, \ldots, \gamma_T\}$.

Our approach is also linked to denoising score matching [35], [36], [37], [38] for training unnormalized energy functions for density estimation. These methods learn a

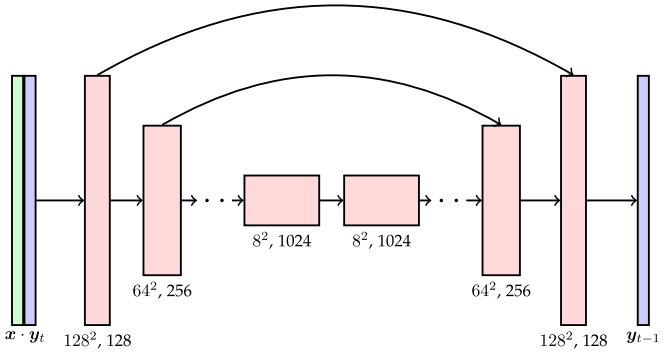


Fig. 3. Depiction of U-Net architecture of SR3. The low resolution input image $\boldsymbol{x}$ is up-sampled to the target resolution using bicubic interpolation, and concatenated with the noisy high resolution output image $\boldsymbol{y}_t$. We show the activation dimensions for a 16×16 → 128×128 super resolution model. We perform self-attention on 16×16 feature maps.

parametric score function to approximate the gradient of the empirical data log-density. To make sure the gradient of the data log-density is well-defined, one often replaces each data point with a Gaussian distribution with a small variance. Song and Ermon [39] advocate the use of a Multi-scale Guassian mixture as the target density, where each data point is perturbed with different amounts of Gaussian noise, so that Langevin dynamics starting from pure noise can still yield reasonable samples.

One can view our approach as a variant of denoising score matching in which the target density is given by a mixture of $q(\widetilde{\boldsymbol{y}}|\boldsymbol{y}_0, \gamma) = \mathcal{N}(\widetilde{\boldsymbol{y}} \mid \sqrt{\gamma}\boldsymbol{y}_0, 1-\gamma)$ for different values of $\boldsymbol{y}_0$ and γ. Accordingly, the gradient of data log-density is given by

$$\frac{\mathrm{d}\log q(\widetilde{\boldsymbol{y}} \mid \boldsymbol{y}_0, \gamma)}{\mathrm{d}\widetilde{\boldsymbol{y}}} = -\frac{\widetilde{\boldsymbol{y}} - \sqrt{\gamma}\boldsymbol{y}_0}{\sqrt{1-\gamma}} = -\boldsymbol{\epsilon}, \tag{14}$$

which is used as the regression target of our model.

2.5 SR3 Model Architecture and Noise Schedules

The SR3 architecture is similar to the U-Net in DDPM [1], with self-attention and modifications adapted from [40]; i.e., we replace the original DDPM residual blocks with residual blocks from BigGAN [41], and we re-scale skip connections by $\frac{1}{\sqrt{2}}$. We increase the number of residual blocks, and the channel multipliers at different resolutions.

To condition the model on the input $\boldsymbol{x}$, we up-sample the low-resolution image to the target resolution using bicubic interpolation. The result is concatenated with $\boldsymbol{y}_t$ along the channel dimension. We experimented with more sophisticated methods of conditioning, including FiLM [42], but found that the simple concatenation yielded similar generation quality.

For our training noise schedule, we follow [33], and use a piece-wise distribution for γ, $p(\gamma) = \sum_{t=1}^{T} \frac{1}{T} U(\gamma_{t-1}, \gamma_t)$. Specifically, during training, we first uniformly sample a time step $t \sim \{0, \ldots, T\}$ followed by sampling $\gamma \sim U(\gamma_{t-1}, \gamma_t)$. For all experiments we set $T = 2000$, and the γ_t are uniformly spaced. Larger values of T generally yield better models, but model performance is relatively insensitive to the exact values of these parameters, so we do no hyperparameter search during SR3 training.

TABLE 1
Task Specific U-Net Architecture Parameters

Task	Channel dim	Depth multipliers	ResNet blocks	#Params
$16^2 \rightarrow 128^2$	128	{1, 2, 4, 8, 8}	3	550M
$64^2 \rightarrow 256^2$	128	{1, 2, 4, 4, 8, 8}	3	625M
$64^2 \rightarrow 512^2$	64	{1, 2, 4, 8, 8, 16, 16}	3	625M
$256^2 \rightarrow 1024^2$	16	{1, 2, 4, 8, 16, 32, 32, 32}	2	150M

Channel dim is the dimension of the first layer, while the depth multipliers apply to the subsequent resolutions.

For sample generation (or inference), early diffusion models [1], [40] required 1-2 K diffusion steps, making generation slow, especially for high resolution images. For more efficient generation we instead adapt recent techniques [33]. In particular, by conditioning on γ directly (versus t in [1]), we have some flexibility in choosing number of diffusion steps and the noise schedule during sample generation. This worked well for speech synthesis [33], but has not been explored for images. In more detail, we set the maximum generation budget (to 100 diffusion steps unless stated otherwise), and assume a linear noise schedule, performing hyper-parameter search over the start and end noise levels. FID on held out data is used during the search to choose the best noise schedule, as we found PSNR did not correlate well with image quality. We also emphasize that this search is inexpensive as it does not require model retraining [33].

Fig. 3 depicts the architecture used for both SR3 and our regression baselines. This denoising U-Net takes as input a noisy high resolution image and a low-resolution conditioning image that has been interpolatd and up-sampled to the target resolution. Task dependent parameters are summarized in Table 1. These architectures have more parameters than many existing networks for image super-resolution, motivated in part by other domains where performance scales with model capacity and dataset size. As shown in Section 4.5, even a simple Regression model with a large architecture can perform surprisingly well.

3 Related Work

SR3 is inspired by recent work on deep generative models and recent learning-based approaches to super-resolution.

Generative Models. Autoregressive models (ARs) [43], [44] can model exact data log likelihood, capturing rich distributions. However, their sequential generation of pixels is expensive, limiting application to low-resolution images. Normalizing flows [9], [10], [45] improve on sampling speed while modelling the exact data likelihood, but the need for invertible parameterized transformations with a tractable Jacobian determinant limits their expressiveness. VAEs [7], [46] offer fast sampling, but tend to underperform GANs and ARs in image quality [8]. Generative Adversarial Networks (GANs) [11] are popular for class conditional image generation and super-resolution. Nevertheless, the inner-outer loop optimization often requires tricks to stabilize training [19], [20], and conditional tasks like super-resolution usually require an auxiliary consistency-based loss to avoid mode collapse [16]. Cascades of GAN models have been used to generate higher resolution images [47].

Score matching [35] models the gradient of the data log-density with respect to the image. Score matching on noisy data, called denoising score matching [36], is equivalent to training a denoising autoencoder, and to DDPMs [1]. Denoising score matching over multiple noise scales with Langevin dynamics sampling from the learned score functions has recently been shown to be effective for high quality unconditional image generation [1], [24]. These models have also been generalized to continuous time [40]. Denoising score matching and diffusion models have also found success in shape generation [48], and speech synthesis [33]. We extend this method to super-resolution, with a simple learning objective, a constant number of inference generation steps, and high quality generation.

Super-Resolution. Numerous super-resolution methods have been proposed [16], [30], [31], [32], [49], [50], [51], [52], [53]. Much of the early work on super-resolution is regression based and trained with an MSE loss [49], [52], [54], [55], [56]. As such, they effectively estimate the posterior mean, yielding blurry images when the posterior is multi-modal [16], [17], [31]. Our regression baseline defined below is also a one-step regression model trained with MSE (cf. [52], [56]), but with a large U-Net architecture. SR3, by comparison, relies on a series of iterative refinement steps, each of which is trained with a regression loss. This difference permits our iterative approach to capture richer distributions. Further, rather than estimating the posterior mean, SR3 generates samples from the target posterior.

Autoregressive models have been used successfully for super-resolution and cascaded up-sampling [15], [18], [57], [58]. Nevertheless, the expensive of inference limits their applicability to low-resolution images. SR3 can generate high-resolution images, e.g., 1024×1024, but with a constant number of refinement steps (often no more than 100).

GAN-based super-resolution methods have also found considerable success [12], [16], [17], [30], [31], [32], [59]. FSRGAN [14] and PULSE [17] in particular have demonstrated high quality face super-resolution results. However, many such GAN based methods are generally difficult to optimize, and often require auxiliary objective functions to ensure consistency with the low resolution inputs.

Normalizing flows have been used for super-resolution with a multi-scale approach [32], [60]. They are competitive with GAN models, and are capable of generating 1024×1024 images due in part to their efficient inference process. SR3 uses a series of reverse diffusion steps to transform a Gaussian distribution to an image distribution while flows require a deep and invertible network.

4 Experiments

We assess the effectiveness of SR3 in super-resolution on faces, natural images, and synthetic images obtained from a low-resolution generative model. The latter enables high-resolution image synthesis using cascaded model.

4.1 Datasets

数据集

We follow previous work [17], training face super-resolution models on Flickr-Faces-HQ (FFHQ) [61] and evaluating on CelebA-HQ [12]. For natural image super-resolution, we train on ImageNet 1K [62] and use the dev split for evaluation. We train unconditional face and class-conditional ImageNet generative models using DDPM on the same datasets discussed above. For training and testing, we use low-resolution images that are down-sampled using bicubic interpolation with anti-aliasing enabled. For ImageNet, we discard images where the shorter side is less than the target resolution. We use the largest central crop like [41], which is then resized to the target resolution using area resampling as our high resolution image.

4.2 Training Details

We train SR3 and regression models for 1 M training steps, with a batch size of 256; this typically takes about four days on 64 TPUv3 chips. Given the large model capacity and large datsets, the models often continue to improve well beyond 1 M steps. We choose a checkpoint for the regression baseline based on peak-PSNR on the held out set. We do not perform any checkpoint selection on SR3 models and simply select the latest checkpoint. Consistent with [1], we use the Adam optimizer with a linear warmup schedule over 10 K training steps, followed by a fixed learning rate of 1e-4 for SR3 models and 1e-5 for regression models. We use a dropout rate of 0.2 for 16×16 → 128×128 models super-resolution, but otherwise, we do not use dropout. We note that the baseline regression models are trained with the same architecture (see Section 2.5), and as shown below, provide a strong baseline for comparison.

4.3 Evaluation

We evaluate SR3 models on face and natural images:

- Face super-resolution at $16\times16\rightarrow128\times128$ and $64\times64\rightarrow512\times512$ trained on FFHQ and evaluated on CelebA-HQ.
- Natural image super-resolution at $64\times64\rightarrow256\times256$ and $56\times56\rightarrow224\times224$ pixels on ImageNet [62].
- Unconditional 1024×1024 face generation by a cascade of 3 models, and class-conditional 256×256 ImageNet image generation by a cascade of 2 models.

We compare SR3 with EnhanceNet [31], ESRGAN [30], SRFlow [32], FSRGAN [14] and PULSE [17]. We also compare to a Regression baseline that shares the same architecture and model capacity as SR3. Importantly, this enables one to directly assess the advantages of iterative refinement over a single step regression model, ablating the effects of model size, architecture, and training data. Performance is assessed qualitatively and quantitatively, using human evaluation, FID scores and the classification accuracy of a pre-trained model on super-resolution outputs.

评价指标

4.4 Qualitative Results

Fig. 4 compares SR3 and our Regression baseline for a 64×64 → 256×256 super-resolution task on a few ImageNet test images. As both models share the same architecture, this provides an indication of the difference between diffusion models and MSE regresion. In particular, as is common with regression models, the outputs are relatively blurry and lack fine-grained structure. The differences are most apparent in the enlarged patches in rows 2 and 4. In Fig. 5 we also show the diversity of SR3 outputs on the task of 16×16 → 256×256 super-resolution on two ImageNet test images. SR3 generates

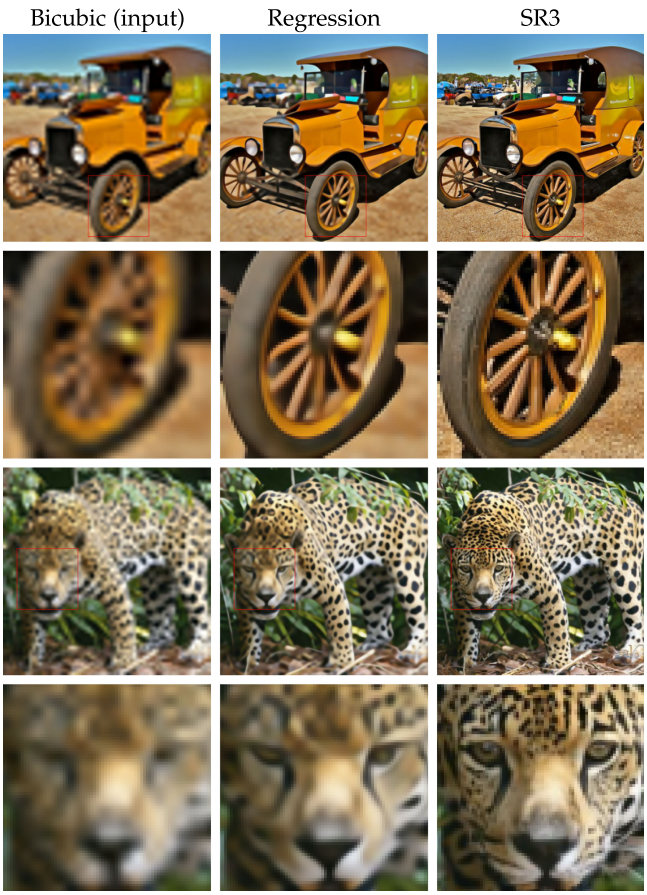


Fig. 4. Super-resolution results (64×64 → 256×256) for SR3 and Regression on ImageNet test images. Both models use the same architecture and training data. We display the full image and an enlarged patch to show fine-grained details.

diverse looking high-resolution outputs for the given low-resolution images.

Fig. 6 shows outputs of our face super-resolution models ($64\times64 \rightarrow 512\times512$) on three test images (provided by colleagues), again with selected patches enlarged. With the $8\times$ magnification factor one can clearly see the detailed structure inferred. Note that, because of the large magnification factor, there are many plausible outputs, so we do not expect the output to exactly match the original reference image (e.g., most evident in the enlarged patches).

Further qualitative comparisons are shown in Fig. 7, where SR3 is compared to SoTA GAN models [30], [31] and

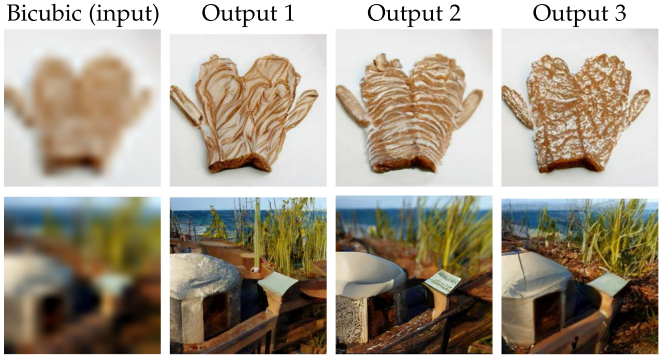


Fig. 5. Three samples from SR3 applied to ImageNet test images (16×16 → 256×256), demonstrating SR3 diversity.

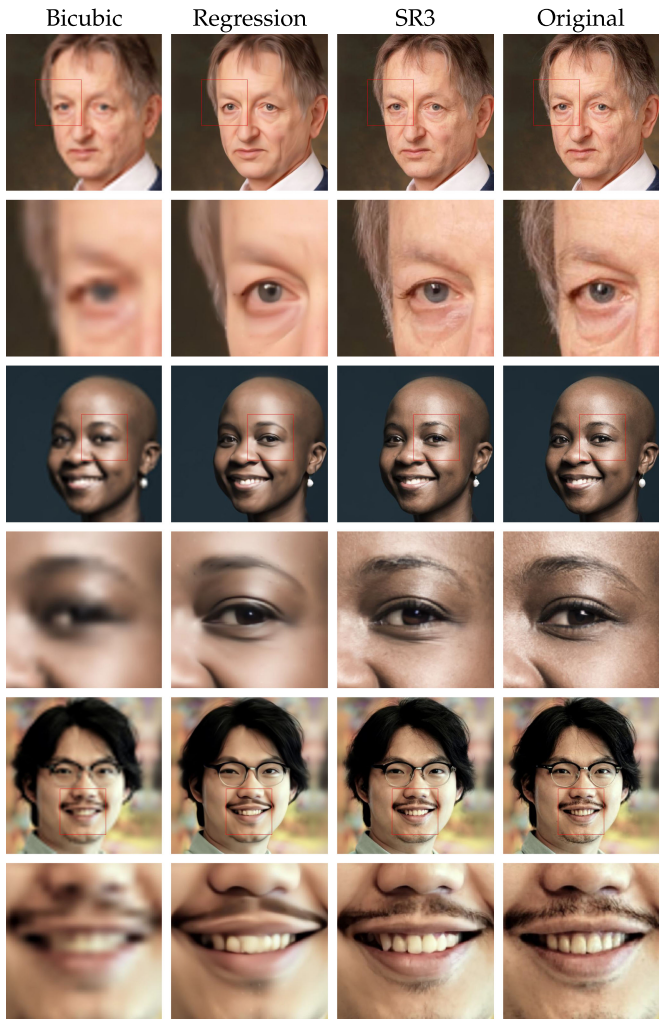


Fig. 6. Results of a SR3 model (64×64 → 512×512), trained on FFHQ, and applied to images outside of the training set.

a Normalizing Flow model [32]. While the GAN- and Flow-based methods produce sharp details, they also tend to generate artifacts in regions with fine-grained texture (e.g., see the face of the jaguar, and the structure of the dockyard). By comparison, SR3 produces sharp images with plausible details and minimal artifacts. As discussed above, while the high resolution details are realistic, they are not expected to perfectly match the original reference image.

4.5 Quantitative Evaluation

Table 2 shows the PSNR, SSIM [63] and Consistency scores for 16×16 → 128×128 face super-resolution. SR3 outperforms PULSE and FSRGAN on PSNR and SSIM while underperforming the regression baseline. Previous work [14], [15], [26] observed that these conventional automated evaluation measures do not correlate well with human perception when the input resolution is low and the magnification factor is large. This is not surprising because these metrics tend to penalize any synthetic high-frequency detail that is not perfectly aligned with the target image. Since generating perfectly aligned high-frequency details, e.g., the exact same hair strands in Fig. 6 and identical leopard spots in Fig. 7, is almost impossible. Thus, PSNR and SSIM tend to prefer MSE regression-based techniques which are conservative with

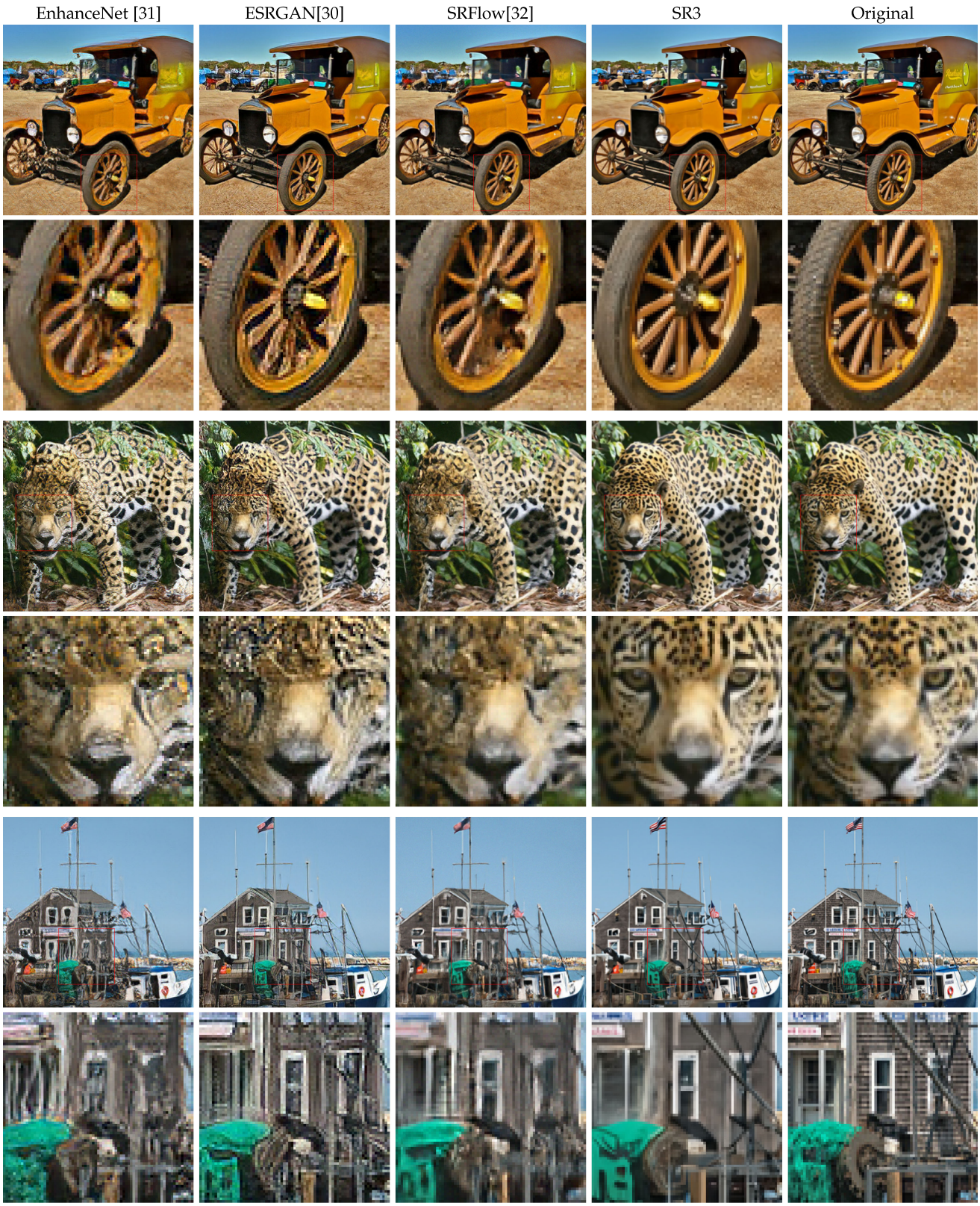


Fig. 7. SR3 and state-of-the-art methods on $4\times$ super-resolution ($64\times64 \rightarrow 256\times256$) applied to ImageNet test images. The outputs of EnhanceNet and ESRGAN are sharp, but include artifacts especially when inspecting enlarged patches. We found that ESRGAN trained on ImageNet-1 M produced similar artifacts. SR3 outputs seem to resemble the original images the most, but one can still find patches in the original images that contain more interesting texture than in SR3 outputs.

high-frequency details. This is further confirmed in Table 3 for ImageNet super-resolution ($64\times64 \rightarrow 256\times256$) where the outputs of SR3 achieve higher sample quality scores (FID and IS), but worse PSNR and SSIM than regression.

4.5.1 *Consistency With Low-Resolution Inputs*

It is important for super-resolution outputs to be consistent with low-resolution inputs. To measure this consistency, we compute MSE between the downsampled outputs and the

TABLE 2
PSNR & SSIM on 16×16 → 128×128 Face Super-Resolution

Metric	PULSE [17]	FSRGAN [14]	Regression	SR3
PSNR ↑	16.88	23.01	**23.96**	23.04
SSIM ↑	0.44	0.62	**0.69**	0.65
Consistency ↓	161.1	33.8	2.71	**2.68**

Consistency measures MSE ($\times 10^{-5}$) between the low-resolution inputs and the down-sampled super-resolution outputs. SR3 outperforms GAN baselines in all metrics, especially improving consistency by a big margin.

low-resolution inputs. Table 2 shows that SR3 achieves the best consistency error beating PULSE and FSRGAN by a significant margin, even slightly outperforming the regression baseline. This result demonstrates the key advantage of SR3 over the state-of-the-art GAN based methods. In fact, SR3 does not require any auxiliary objective function in order to ensure consistency with the low-resolution inputs.

4.5.2 Classification Accuracy on Super-Resolution Outputs

Table 4 compares the outputs of 4× natural image super-resolution models on object classification accuracy. Following [31], [64] we apply 4× super-resolution models to 56×56 center crops from the validation set of ImageNet. Then, we report classification accuracy of a pre-trained ResNet-50 [65] model. Since SR3 models are trained on the task of 64×64 → 256×256, we use bicubic interpolation to resize 56×56 inputs to 64×64, and then apply 4× super-resolution, followed by resizing to 224×224. We note that while SR3 and Regression were trained on ImageNet data (without labels), the remaining baselines in Table 4 were not.

SR3 outperforms existing methods by a significant margin on both top-1 and top-5 classification errors, suggesting higher perceptual quality. The strong performance of the Regression model can be attributed to the model capacity and architecture, and in part because it was trained on ImageNet data. The improvement of SR3 over Regression can be viewed as a direct indication of the power of the diffusion framework and iterative refinement, as both models use the same architecture. These results also reaffirm the limits of conventional reference-based metrics in super-resolution, like PSNR and SSIM, for which the baseline Regression model exhibits higher performance.

4.5.3 Human Evaluation (2AFC)

Direct human evaluation is one of the most desirable metrics for evaluating super-resolution models. While mean opinion score (MOS) is commonly used to measure image quality in this context, forced choice pairwise comparison has been found to be a more reliable method for such subjective quality assessments [69]. Furthermore, standard MOS studies do not capture consistency between low-resolution inputs and high-resolution outputs.

TABLE 3
Performance Comparison Between SR3 and Regression Baseline on Natural Image Super-Resolution Using Standard Metrics Computed on the ImageNet Validation Set

Model	FID ↓	IS ↑	PSNR ↑	SSIM ↑
Reference	1.9	240.8	-	-
Regression	15.2	121.1	**27.9**	**0.801**
SR3	**5.2**	**180.1**	26.4	0.762

TABLE 4
Comparison of ResNet-50 Classification Accuracy on 4× Super-Resolution Outputs of the First 1 K Images From the ImageNet Validation Set

Method	Top-1 Accuracy	Top-5 Accuracy
Baseline	0.748	0.920
DRCN [50]	0.523	0.758
PsyCo [66]	0.546	0.776
ENet-E [31]	0.551	0.786
FSRCNN [67]	0.563	0.804
RCAN [64]	0.607	0.833
DRLN [68]	0.655	0.879
Regression	0.617	0.827
SR3	**0.683**	**0.880**

Note: These existing baselines have not been trained on ImageNet.

We use a 2-alternative forced-choice (2AFC) paradigm to measure how well humans can discriminate true images from those generated from a model. In Task-1 subjects were shown a low resolution input in between two high-resolution images, one being the real image (ground truth), and the other generated from the model. Subjects were asked *"Which of the two images is a better high quality version of the low resolution image in the middle?"* This task takes into account both image quality and consistency with the low resolution input. Task-2 is similar to Task-1, except that the low-resolution image was not shown, so subjects only had to select the image that was more photo-realistic. They were asked *"Which image would you guess is from a camera?"* Subjects viewed images for 3 seconds before responding. The source code for human evaluation can be found here.[1]

The subject *fool rate* is the fraction of trials on which a subject selects the model output over ground truth. Our fool rates for each model are based on 50 subjects, each of whom were shown 50 of the 100 images in the test set. Fig. 9 shows the fool rates for Task-1 (top), and for Task-2 (bottom). In both experiments, the fool rate of SR3 is close to 50%, indicating that SR3 produces images that are both photo-realistic and faithful to the low-resolution inputs. We find similar fool rates over a wide range of viewing durations up to 12 seconds.

The fool rates for FSRGAN and PULSE in Task-1 are lower than the Regression baseline and SR3. The strength of SR3 over the Regression model reflects the benefits of iterative refinement in the diffusion model, since both models share the same architecture. We speculate that the PULSE optimization has failed to converge to high resolution images sufficiently close to the inputs. Indeed, when asked solely about image quality in Task-2 (Fig. 9 (bottom)), the PULSE fool rate increases significantly.

The fool rate for the Regression baseline is lower in Task-2 (Fig. 9 (bottom)) than Task-1. The regression model tends to generate images that are blurry, but nevertheless faithful

1. https://tinyurl.com/sr3-human-eval-code

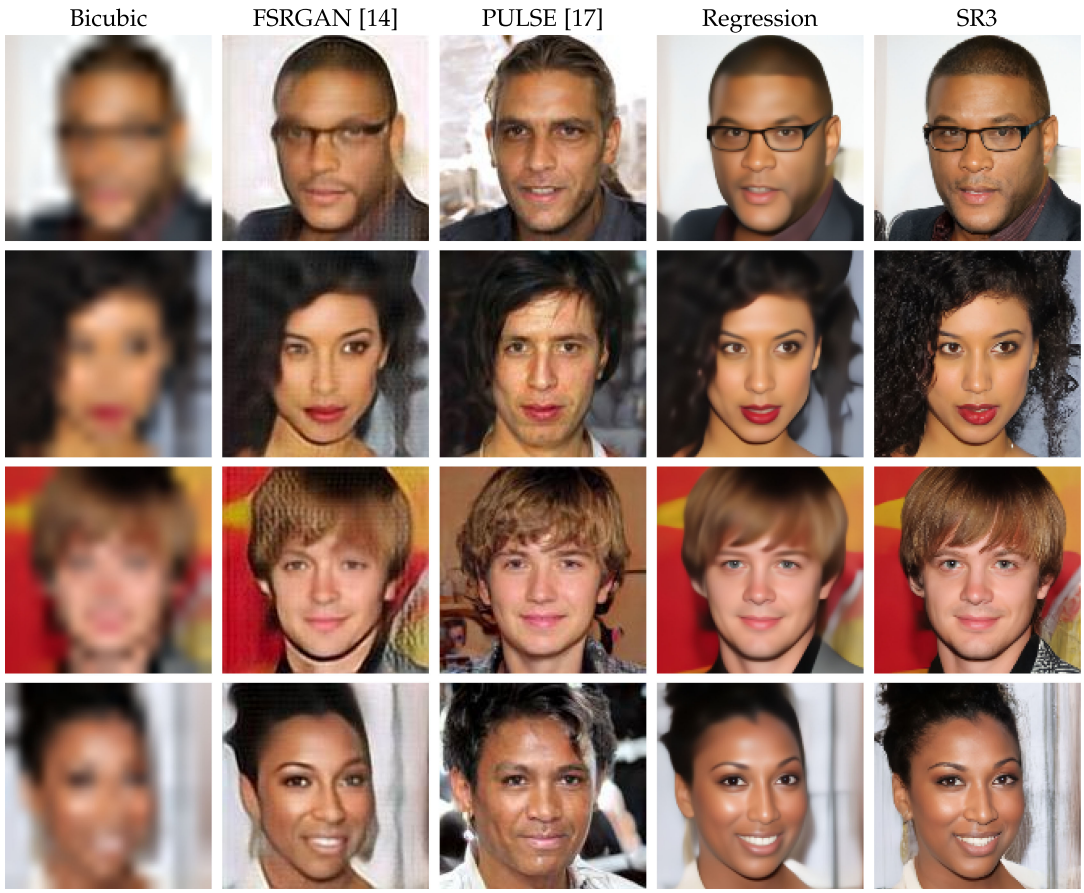


Fig. 8. Comparison on 4× face super-resolution (16×16 → 128×128). Reference images are removed for privacy concerns.

to the low resolution input. We speculate that in Task-1, given the inputs, subjects are influenced by consistency, while in Task-2, ignoring consistency, they instead focus on image sharpness. SR3 and Regression samples used for human evaluation are provided here[2].

The results of a similar study with natural images, comparing SR3 with Regression, GAN-based models [30], [31] and a Flow-based model [32] on a subset of the ImageNet validation set are shown in Fig. 10. In this study images were displayed for 6 seconds and the input images were not displayed (i.e., Task-2). We used somewhat longer display times because natural images are more complex and cluttered than the face images. We did not show the input image because inconsistency between inputs and model outputs did not appear to be problematic with the baselines used. From Fig. 10 one can see that SR3 outperforms baselines by a substantial margin, suggesting higher perceptual quality. The regression model is significantly weaker in this case, which we attribute to the longer viewing time which makes it easier to discern the image blur.

To further appreciate the experimental results, it is useful to visually compare outputs of different models on the same inputs, as in Fig. 8. FSRGAN exhibits distortion in face region and struggles with generating glasses properly (e.g., top row). It also fails to recover texture details in the hair region (see bottom row). PULSE often produces images that differ significantly from the input image, both in the shape of the face and the background, and sometimes in gender too (see bottom row) presumably due to failure of the optimization to find a sufficiently good minima. As noted above, our Regression baseline produces results consistent to the input, however they are typically quite blurry. By comparison, the SR3 results are consistent with the input and contain more detailed image structure.

In addition to the aggregate fool rate results in Fig. 9, it is also interesting to inspect images that attain highest and lowest fool rates for a given technique. This provides insight into the nature of the problems that models exhibit, as well as cases in which the model outputs are good enough to regularly confuse people.

Fig. 11 displays the outputs of PULSE [17] and SR3 with the lowest and highest fool rates for Task-1 (the conditional task). Notice that images from PULSE for which the fool rate is low have obvious distortions, and the fool rates are lower than 10%. For SR3, by comparison, the images with the lowest fool rates are still reasonably good, with much higher fool rates of 14% and 19%. It is interesting to see that the best fool rates for SR3 on Task-1 are 84% and 88%. The corresponding original images for these examples are somewhat noisy, and as a consequence, interestingly, many subjects prefer the SR3 outputs.

4.6 Generation Speed

As discussed in Section 2.5, diffusion models typically require a large number of refinement steps during sample

2. https://tinyurl.com/sr3-outputs

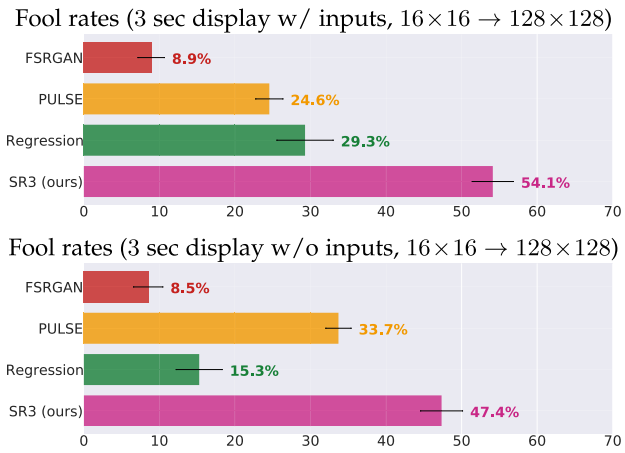


Fig. 9. Face super-resolution human fool rates (higher is better, for photo-realistic samples one would expect a fool rate close to 50%). Outputs of four models are compared to ground truth. (top) Task-1, subjects are shown low-resolution inputs. (bottom) Task-2, inputs are not shown.

generation, and are therefore expensive compared to GANs. For more efficient inference, given a generation budget, SR3 determines the noise schedule using hyper-parameter search. Fig. 12 shows the resulting trade-off between image quality (FID) and efficiency (number of diffusion steps) for a 64×64 → 256×256 models trained on ImageNet. FID is computed on model outputs for the entire ImageNet validation set, using the original validation set as the reference distribution. Fig. 12 also compares to a baseline Regression model (one-step generation). One can see that with just 4 refinement steps SR3 yields a significant drop in FID compared to the single step regression baseline. In practice, each refinement step takes about 1 ms.

4.7 Cascaded High-Resolution Image Synthesis

We also study *cascaded* image generation, where SR3 models at different scales are chained together with generative models, enabling high-resolution image synthesis. Cascaded generation allows one to train different models in parallel, and each model in the cascade solves a simpler task, requiring fewer parameters and less computation for training. Inference with cascaded models is also more efficient, especially for iterative refinement models. With cascaded generation we found it effective to use more refinement steps at low-resolutions, and fewer steps at higher resolutions. This was much more efficient than generating directly at high resolution without sacrificing image quality.

For cascaded face generation, as depicted in Fig. 14, we train a DDPM [1] model for unconditional 64×64 face images. Samples from this low-dimensional model are then

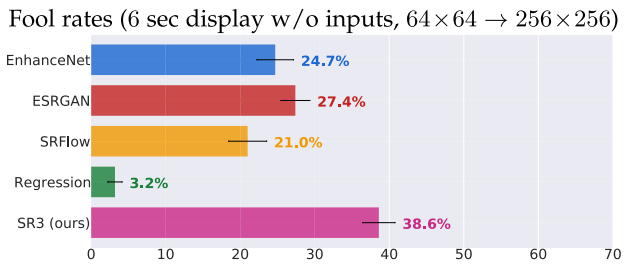


Fig. 10. ImageNet super-resolution fool rates. Model outputs are compared to ground truth with pair of images shown for 6 seconds.

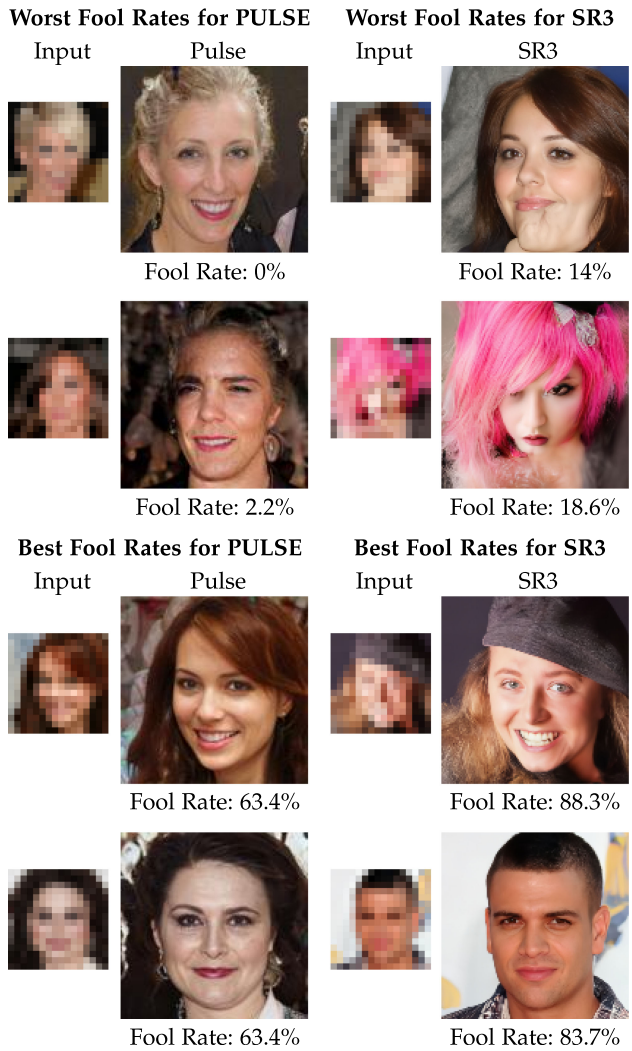


Fig. 11. Test cases with lowest and highest Fool rates for PULSE and SR3 in Task-1 (which compares models outputs to reference images, in the presence of low-resolution inputs). For privacy reasons, reference images are not shown.

fed to two 4× SR3 models, up-sampling to 256×256 and then to 1024×1024 pixels. A small set of synthetic high-resolution face samples is shown in Fig. 13.

We also trained a set of Improved DDPM [70] models on class-conditional 64×64 ImageNet data. Samples from these class-conditional models are then passed to a 4× SR3 model to produce 256×256 natural images. We note that the 4× SR3 model is not conditioned on the class label. Fig. 15 shows four samples for each of six classes. Fig. 16 shows a selected set of samples many different classes.

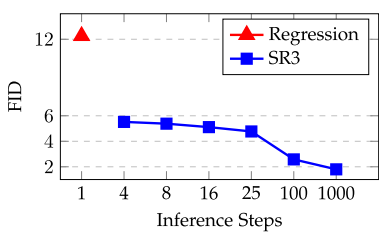


Fig. 12. FID score versus number of inference steps for 64×64 → 256×256 super-resolution. Regression (red) requires one step. All SR3 results (blue) are generated from the same denoising model but different inference noise schedules.

Fig. 13. Synthetic 1024×1024 faces, sampled from an unconditional 64×64 model, followed by two 4× SR3 models.

As a way to quantitatively evaluate sample quality, Table 5 reports FID scores for the resulting class-conditional ImageNet samples. Our 2-stage model improves on VQ-VAE-2 [71], is comparable to deep BigGANs [41] at truncation factor of 1.5 but underperforms them a truncation factor of 1.0. Unlike BigGAN, our diffusion models do not provide control of sample quality versus sample diversity; this remains an interesting avenue for future research. Nichol and Dhariwal [70] concurrently trained cascaded generation models using super-resolution conditioned on class labels (SR3 is not conditioned on class labels), and also observed a similar trend with improved FID scores. The effectiveness of cascaded image generation indicates that SR3 models are robust to the precise distribution of inputs (i.e., the specific form of anti-aliasing and downsampling).

4.8 Ablation Studies on Cascaded Models

Table 6 reports results of ablations on a $64\times64 \rightarrow 256\times256$ Imagenet SR3 model. First, to improve SR3 robustness, we experiment with data augmentation during training. That is, we trained SR3 with varying amounts of Gaussian Blurring noise added to the low resolution input image. No blurring is applied during inference. We find that this has a significant impact, improving the FID score roughly by 2 points. In addition to their efficiency, our initial cascade experiments gave lower FID scores than full resolution

Fig. 14. Cascaded generation with an unconditional model chained with two SR3 models.

Fig. 15. Class-conditional 256×256 ImageNet samples. Each row represents samples from a specific ImageNet class, from top to bottom: Goldfish, Red Fox, Balloon, Monarch Butterfly, Church, Fire Truck. For a given label, we sample a 64×64 image from a class-conditional diffusion model, and then apply a 4× SR3 model.

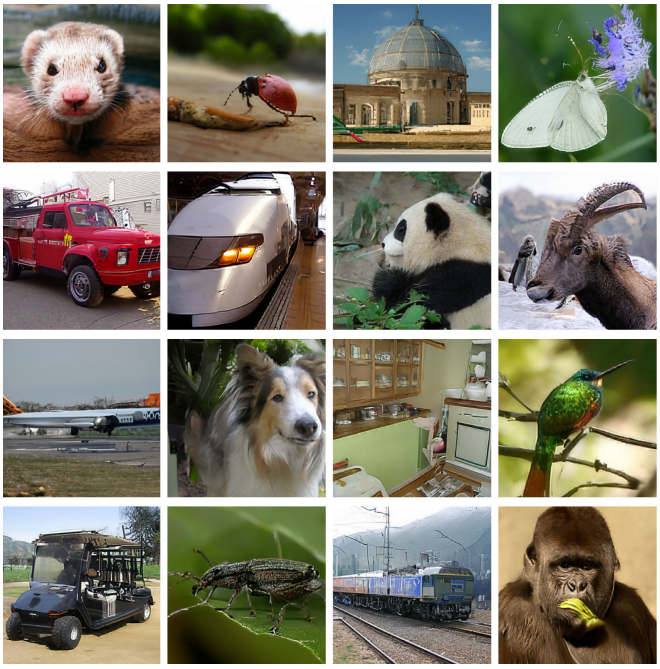


Fig. 16. Synthetic 256×256 ImageNet images. We draw a label at random, sample a 64×64 image from the corresponding class-conditional diffusion model, and then apply a 4× SR3 model.

models trained directly, which has been confirmed in more recent work [72].

We also explore the choice of L_p norm for the denoising objective (6). We find that L_1 norm gives slightly better FID scores than L_2. However, subsequent work shows that L_2 tends to generate greater diversity in SR3 outputs [34].

5 Discussion and Conclusion

SR3 leverages conditional diffusion models to address single image super-resolution. It initializes the output image with random Gaussian noise iteratively refines the conditioned on the low resolution input. We find that SR3 works well on natural images and faces images, with a wide range of magnification factors, or as part of a cascading pipeline to generate high resolution images. SR3 models outperform several GAN and Normalizing Flow baselines. Human studies, in which subjects are asked to discriminate model outputs from real images, yield SR3 fool rates close to 50% on faces and 40% on natural images, which indicates that SR3 produces high fidelity outputs. The success of SR3 is in part a function of large model capacity and the use of large training datasets, motivating further exploration of scaling in future super-resolution work.

TABLE 5
FID Scores for Class-Conditional, 256×256 ImageNet Generation

Model	FID-50 k
Prior Work	
VQ-VAE-2 [71]	38.1
BigGAN (Truncation 1.0) [41]	7.4
BigGAN (Truncation 1.5) [41]	11.8
Our Work	
SR3 (Two Stage)	11.3

TABLE 6
Ablations on SR3 for Class-Conditional 256×256 ImageNet

Model	FID-50 k
Training with Augmentation	
SR3	13.1
SR3 (w/ Gaussian Blur)	11.3
Objective L_p Norm	
SR3 (L_2)	11.8
SR3 (L_1)	11.3

Augmenting input images with Gaussian blur improves SR3 performance, and L_1 loss outperforms L_2. 未来研究方向

One practical issue with diffusion models is the computation cost of many refinement steps during inference. Our results indicate that one can trade sample quality for generation speed and achieve decent results in just 4 refinement steps. That said, recent and concurrent work proposes alternative approaches that can result in higher quality fast samplers for diffusion models [73], [74], [75], [76]. We further note that the use of self-attention, while powerful, also constrains the output dimension of our model; this will be addressed in future versions of SR3.

Finally, bias is an important issue with all generative models, including SR3. While in theory, our log-likelihood based objective is mode covering (e.g., unlike some GAN-based objectives), we do observe some indication of mode drop in SR3 outputs, e.g., the model consistently generates nearly the same image output during sampling (when conditioned on the same input). We also observe that the model generates very continuous skin texture in face super-resolution, dropping moles, pimples and piercings found in the reference image. We note that SR3 should be used in super-resolution products after further studies of its potential biases. Nevertheless, diffusion models like SR3 can be useful in reducing dataset bias by generating synthetic data from underrepresented groups.

Acknowledgments

Thanks to Jimmy Ba, Geoff Hinton and Shingai Manjengwa kindly provided their face images for testing, and to Ben Poole, Samy Bengio and the Google Brain team for discussions and technical assistance. The authors thank the authors of [17] for generously providing samples for human evaluation.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
[2] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.
[3] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[4] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[5] A. v. d. Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[6] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4797–4805.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2013.

[8] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 19667–19679.

[9] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," 2016, *arXiv:1605.08803*.

[10] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 10236–10245.

[11] I. J. Goodfellow et al., "Generative adversarial networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018.

[13] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[14] 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.*, 2018, pp. 2492–2501.

[15] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5449–5458.

[16] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 105–114.

[17] 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.*, 2020, pp. 2434–2442.

[18] N. Parmar et al., "Image transformer," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.

[19] M. Arjovsky, S. Chintala, and L. Bottou, "GAN Wasserstein," 2017, *arXiv:1701.07875*.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," 2017, *arXiv:1704.00028*.

[21] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," 2016, *arXiv:1611.02163*.

[22] S. Ravuri and O. Vinyals, "Classification accuracy score for conditional generative models," 2019, *arXiv:1905.10887*.

[23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[24] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 11918–11930.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[26] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 658–666.

[27] D. Berthelot, P. Milanfar, and I. Goodfellow, "Creating high resolution images with a latent adversarial generator," 2020, *arXiv:2003.02365*.

[28] Z. Kadkhodaie and E. P. Simoncelli, "Solving linear inverse problems using the prior implicit in a denoiser," 2021, *arXiv:2007.13640*.

[29] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[30] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 63–79.

[31] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4491–4500.

[32] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "SRFlow: Learning the super-resolution space with normalizing flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 715–732.

[33] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating Gradients for Waveform Generation," in *Proc. Int. Conf. Learn. Representations*, 2021.

[34] C. Saharia et al., "Palette: Image-to-image diffusion models," 2021, *arXiv:2111.05826*.

[35] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 695–709, 2005.

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

[37] M. Raphan and E. P. Simoncelli, "Least squares estimation without priors or supervision," *Neural Comput.*, vol. 23, pp. 374–420, 2011.

[38] S. Saremi, A. Mehrjou, B. Schölkopf, and A. Hyvärinen, "Deep energy estimator networks," 2018, *arXiv:1805.08306*.

[39] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," 2020, *arXiv:2006.09011*.

[40] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2021.

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

[42] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3942–3951.

[43] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1747–1756.

[44] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications," in *Proc. Int. Conf. Learn. Representations*, 2017.

[45] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1530–1538.

[46] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1278–1286.

[47] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[48] R. Cai et al., "Learning gradient fields for shape generation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 364–381.

[49] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

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

[51] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3147–3155.

[52] N. Ahn, B. Kang, and K.-A. Sohn, "Image super-resolution via progressive cascading residual network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 904–908.

[53] X. Li, C. Chen, S. Zhou, X. Lin, W. Zuo, and L. Zhang, "Blind face restoration via deep multi-scale component dictionaries," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 399–415.

[54] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 370–378.

[55] 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.

[56] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[57] J. Menick and N. Kalchbrenner, "Generating high fidelity images with subscale pixel networks and multidimensional upscaling," in *Proc. Int. Conf. Learn. Representations*, 2019.

[58] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4797–4805.

[59] L. Yang et al., "HiFaceGAN: Face renovation via collaborative suppression and replenishment," 2020, *arXiv:2005.05005*.

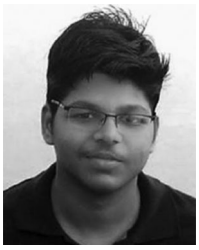
[60] J. J. Yu, K. G. Derpanis, and M. A. Brubaker, "Wavelet flow: Fast training of high resolution normalizing flows," 2020, *arXiv:2010.13821*.

[61] 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.*, 2019, pp. 4396–4405.

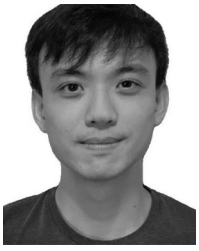
[62] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[63] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

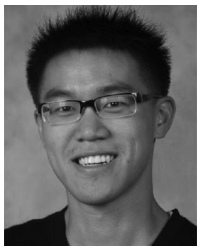
[64] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[66] E. Pérez-Pellitero, J. Salvador, J. Hidalgo, and B. Rosenhahn, "PSyCo: Manifold span reduction for super resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1837–1845.
[67] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.
[68] 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.
[69] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.
[70] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021, *arXiv:2102.09672*.
[71] A. Razavi, A. v. d. Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," 2019, *arXiv:1906.00446*.
[72] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *J. Mach. Learn. Res.*, vol. 23, no. 47, pp. 1–33, 2022.
[73] D. Watson, J. Ho, M. Norouzi, and W. Chan, "Learning to efficiently sample from diffusion probabilistic models," 2021, *arXiv:2106.03802*.
[74] 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*.
[75] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *Proc. Int. Conf. Learn. Representations*, 2022.
[76] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.



Chitwan Saharia received the bachelor's degree in computer science from IIT Bombay, in 2019. He is a software engineer with Google Research. His research interests include deep generative models, and natural language understanding.



Jonathan Ho received the PhD degree in computer science from UC Berkeley, in 2020. He is a research scientist with Google Research. His research interests include deep generative models, data compression, unsupervised learning, and reinforcement learning.



William Chan received the BASc degree in computer engineering from the University of Waterloo, in 2011, and the PhD degree in electrical and computer engineering from Carnegie Mellon University, in 2016. He is currently a research scientist with Google Brain Toronto. His current research crosses the fields of machine learning, deep learning, sequence understanding/generation with applications to natural language processing, speech recognition, speech synthesis and computer vision.

Tim Salimans is a senior research scientist on the Google Brain team in Amsterdam. He works on generative modeling, semi-supervised and unsupervised deep learning, and reinforcement learning. He is most well known for work on GANs, showing how they can be used for semi-supervised learning and how they can be evaluated using the Inception score, and for his work on VAEs and other applications of variational inference using reparameterization.

David J. Fleet received the PhD degree from the University of Toronto, in 1991. He is professor of computer science with the University of Toronto, faculty member of the Vector Institute, and a research scientist with the Brain Team of Google Research. His research interests include computer vision, image processing, machine learning and visual neuroscience. He was awarded an Alfred P. Sloan Research Fellowship, in 1996, and has won paper awards with ICCV 1999, CVPR 2001, UIST 2003, and BMVC 2009. In 2010 he was awarded the Koenderink Prize for his work with Michael Black and Hedvig Sidenbladh on human pose tracking. He has served as area chair for numerous vision and machine learning conference, and as program co-chair for CVPR 2003 and ECCV 2014. He currently serves on the TPAMI Advisory Board.

Mohammad Norouzi received the PhD degree in computer science from the University of Toronto, in 2015. He is a staff research scientist with the Google Research Brain Team in Toronto. His research interests include deep generative models, sequence models, self-supervised learning, and reinforcement learning.