Variational Diffusion Models

Diederik P. Kingma*, **Tim Salimans***, **Ben Poole, Jonathan Ho**
Google Research

Abstract

Diffusion-based generative models have demonstrated a capacity for perceptually impressive synthesis, but can they also be great likelihood-based models? We answer this in the affirmative, and introduce a family of diffusion-based generative models that obtain state-of-the-art likelihoods on standard image density estimation benchmarks. Unlike other diffusion-based models, our method allows for efficient optimization of the noise schedule jointly with the rest of the model. We show that the variational lower bound (VLB) simplifies to a remarkably short expression in terms of the signal-to-noise ratio of the diffused data, thereby improving our theoretical understanding of this model class. Using this insight, we prove an equivalence between several models proposed in the literature. In addition, we show that the continuous-time VLB is invariant to the noise schedule, except for the signal-to-noise ratio at its endpoints. This enables us to learn a noise schedule that minimizes the variance of the resulting VLB estimator, leading to faster optimization. Combining these advances with architectural improvements, we obtain state-of-the-art likelihoods on image density estimation benchmarks, outperforming autoregressive models that have dominated these benchmarks for many years, with often significantly faster optimization. In addition, we show how to use the model as part of a bits-back compression scheme, and demonstrate lossless compression rates close to the theoretical optimum.

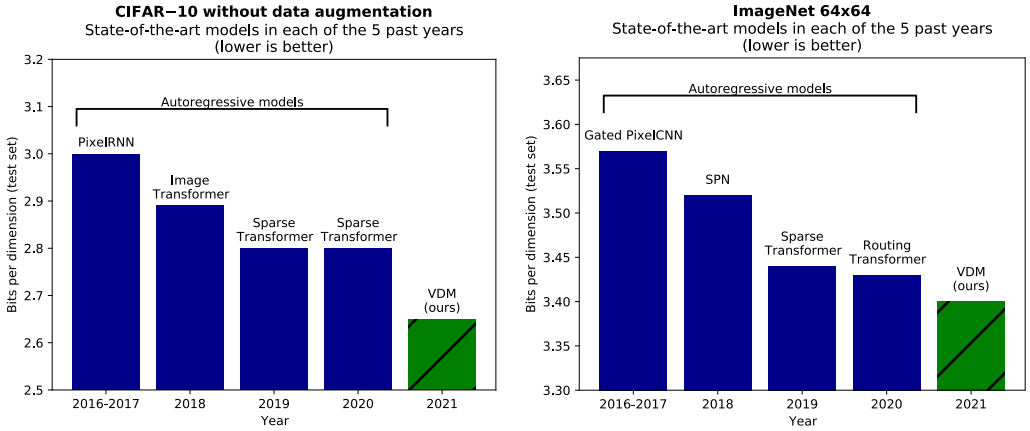
1 Introduction

Likelihood-based generative modeling is a central task in machine learning that is the basis for a wide range of applications ranging from speech synthesis [Oord et al., 2016], to translation [Sutskever et al., 2014], to compression [MacKay, 2003], to many others. Autoregressive models have long been the dominant model class on this task due to their tractable likelihood and expressivity, as shown in Figure 1. Diffusion models have recently shown impressive results in image [Ho et al., 2020, Song et al., 2021b, Nichol and Dhariwal, 2021] and audio generation [Kong et al., 2020, Chen et al., 2020] in terms of perceptual quality, but have yet to match autoregressive models on density estimation benchmarks. In this paper we make several technical contributions that allow diffusion models to challenge the dominance of autoregressive models in this domain. Our main contributions are as follows:

- We introduce a flexible family of diffusion-based generative models that achieve new state-of-the-art log-likelihoods on standard image density estimation benchmarks (CIFAR-10 and ImageNet). This is enabled by incorporating Fourier features into the diffusion model and using a learnable specification of the diffusion process, among other modeling contributions.
- We improve our theoretical understanding of density modeling using diffusion models by analyzing their variational lower bound (VLB), deriving a remarkably simple expression in terms of the signal-to-noise ratio of the diffusion process. This result delivers new insight

* Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.



(a) CIFAR-10 without data augmentation

(b) ImageNet 64x64

Figure 1: Autoregressive generative models were long dominant in standard image density estimation benchmarks. In contrast, we propose a family of diffusion-based generative models, *Variational Diffusion Models* (VDMs), that outperforms contemporary autoregressive models in these benchmarks. See Table 1 for more results and comparisons.

into the model class: for the continuous-time (infinite-depth) setting we prove a novel invariance of the generative model and its VLB to the specification of the diffusion process, and we show that various diffusion models from the literature are equivalent up to a trivial time-dependent rescaling of the data.

2 Related work

Our work builds on diffusion probabilistic models (DPMs) [Sohl-Dickstein et al., 2015], or *diffusion models* in short. DPMs can be viewed as a type of variational autoencoder (VAE) [Kingma and Welling, 2013, Rezende et al., 2014], whose structure and loss function allows for efficient training of arbitrarily deep models. Interest in diffusion models has recently reignited due to their impressive image generation results [Ho et al., 2020, Song and Ermon, 2020].

Ho et al. [2020] introduced a number of model innovations to the original DPM, with impressive results on image generation quality benchmarks. They showed that the VLB objective, for a diffusion model with discrete time and diffusion variances shared across input dimensions, is equivalent to multi-scale denoising score matching, up to particular weightings per noise scale. Further improvements were proposed by Nichol and Dhariwal [2021], resulting in better log-likelihood scores. Gao et al. [2020] show how diffusion can also be used to efficiently optimize energy-based models (EBMs) towards a close approximation of the log-likelihood objective, resulting in high-fidelity samples even after long MCMC chains.

Song and Ermon [2019] first proposed learning generative models through a multi-scale denoising score matching objective, with improved methods in Song and Ermon [2020]. This was later extended to continuous-time diffusion with novel sampling algorithms based on reversing the diffusion process [Song et al., 2021b].

Concurrent to our work, Song et al. [2021a], Huang et al. [2021], and Vahdat et al. [2021] also derived variational lower bounds to the data likelihood under a continuous-time diffusion model. Where we consider the infinitely deep limit of a standard VAE, Song et al. [2021a] and Vahdat et al. [2021] present different derivations based on stochastic differential equations. Huang et al. [2021] considers both perspectives and discusses the similarities between the two approaches. An advantage of our analysis compared to these other works is that we present an intuitive expression of the VLB in terms of the signal-to-noise ratio of the diffused data, which then leads to new results on the invariance of the generative model and its VLB to the specification of the diffusion process. We empirically compare to these works, as well as others, in Table 1.

Previous approaches to diffusion probabilistic models fixed the diffusion process, while we consider flexible learned diffusion processes. This is enabled by directly parameterizing the mean and variance of the marginal $q(\mathbf{z}_t|\mathbf{z}_0)$, where previous approaches instead parameterized the individual diffusion steps $q(\mathbf{z}_{t+\epsilon}|\mathbf{z}_t)$. In addition, our denoising models include several architecture changes, the most important of which is the use of Fourier features, which enable us to reach much better likelihoods than previous diffusion probabilistic models.

3 Model

We will focus on the most basic case of generative modeling, where we have a dataset of observations of $\mathbf{x}$, and the task is to estimate the marginal distribution $p(\mathbf{x})$. As with most generative models, the described methods can be extended to the case of multiple observed variables, and/or the task of estimating conditional densities $p(\mathbf{x}|\mathbf{y})$. The proposed latent-variable model consists of a diffusion process (Section 3.1) that we invert to obtain a hierarchical generative model (Section 3.2). We optimize the model parameters by maximizing the variational lower bound of the marginal log-likelihood (Section 4). In contrast with earlier DPMs, we optimize the forward time diffusion process jointly with the rest of the model. This turns the model into a type of VAE [Kingma and Welling, 2013, Rezende et al., 2014].

3.1 Forward time diffusion process

The starting point for our generative model is a diffusion process that begins with the data $\mathbf{x}$, and then samples a sequence of latent variables $\mathbf{z}_t$ given $\mathbf{x}$, where t runs forward in time from $t = 0$ to $t = 1$. The distribution of latent variable $\mathbf{z}_t$ conditioned on $\mathbf{x}$, for any $t \in [0, 1]$ is given by:

$$q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}\left(\alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I}\right), \tag{1}$$

where α_t and σ_t^2 are strictly positive scalar-valued functions of t. Both α_t and σ_t^2 are smooth, such that their derivatives with respect to time t are finite. Furthermore, their ratio α_t^2/σ_t^2 is strictly monotonically decreasing in t, such that $\alpha_t^2/\sigma_t^2 < \alpha_s^2/\sigma_s^2$ for any $t > s$.

The distribution of $\mathbf{z}_t$ given $\mathbf{z}_s$, for any $0 \leq s < t \leq 1$, is given by:

$$q(\mathbf{z}_t|\mathbf{z}_s) = \mathcal{N}\left(\alpha_{t|s}\mathbf{z}_s, \sigma_{t|s}^2 \mathbf{I}\right), \quad \text{where} \quad \alpha_{t|s} = \alpha_t/\alpha_s, \quad \text{and} \quad \sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2. \tag{2}$$

Furthermore we define $q(\mathbf{z}_u|\mathbf{z}_t, \mathbf{z}_s) = q(\mathbf{z}_u|\mathbf{z}_t)$ at any subsequent timesteps $0 \leq s < t < u \leq 1$, such that the joint distribution of latent variables $(\mathbf{z}_s, \mathbf{z}_t, \mathbf{z}_u)$ is first-order Markov.

Given the distributions above, it is relatively straightforward to verify through Bayes rule that the reverse time inference distribution $q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})$, for any $0 \leq s < t \leq 1$, is also Gaussian. This distribution is given in Appendix A. In Appendix J we provide an implementation that is numerically stable for small $t - s$.

As we show in Section 4, the continuous-time VLB objective that we will propose optimizing is surprisingly invariant to the choice of functions α_t and σ_t, which we refer to as the *noise schedule*. Their only impact on our objective is through their ratio $\mathrm{SNR}(t) = \alpha_t^2/\sigma_t^2$ at times $t = 0$ and $t = 1$, which we call the *signal-to-noise ratio* (SNR). In previous works the signal-to-noise ratio $\mathrm{SNR}(t)$ was a fixed function of time, but here we learn this function jointly with the rest of the model, as we explain in Section 4.4. We parameterize it as:

$$\mathrm{SNR}(t) = \exp(-\gamma_{\boldsymbol{\eta}}(t)) \tag{3}$$

where $\gamma_{\boldsymbol{\eta}}(t)$ is a monotonic neural network with parameters $\boldsymbol{\eta}$, as detailed in Appendix I. In our experiments, we choose to use *variance-preserving* diffusion processes as in [Sohl-Dickstein et al., 2015, Ho et al., 2020] where $\alpha_t = \sqrt{1 - \sigma_t^2}$. Written as a function of $\gamma_{\boldsymbol{\eta}}(t)$, this yields:

$$\alpha_t^2 = \mathrm{sigmoid}(-\gamma_{\boldsymbol{\eta}}(t)) \quad \text{and} \quad \sigma_t^2 = \mathrm{sigmoid}(\gamma_{\boldsymbol{\eta}}(t)). \tag{4}$$

The *variance-exploding* diffusion processes used by Song and Ermon [2019], Song et al. [2021b], can also be seen as special cases of the proposed model, using $\alpha_t^2 = 1$ and $\sigma_t^2 = \exp(\gamma_{\boldsymbol{\eta}}(t))$.

3.2 Reverse time generative model

We define our generative model by inverting the diffusion process of Section 3.1, yielding a hierarchical generative model that samples a sequence of latents $\mathbf{z}_t$, with time running backward from $t = 1$ to $t = 0$. We consider both the case where this sequence consists of a finite number of steps T, as well as a continuous time model corresponding to $T \to \infty$. We start by presenting the discrete-time case.

Given finite T, we discretize time uniformly into T segments of width $\tau = 1/T$. Defining $s(i) = (i-1)/T$ and $t(i) = i/T$, our hierarchical generative model for data $\mathbf{x}$ is then given by:

$$p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{z}_1) p(\mathbf{x}|\mathbf{z}_0) \prod_{i=1}^{T} p(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}). \tag{5}$$

With the variance preserving diffusion specification and sufficiently small $\mathrm{SNR}(1)$, we have that $q(\mathbf{z}_1|\mathbf{x}) \approx \mathcal{N}(\mathbf{z}_1; 0, \mathbf{I})$. We therefore model the marginal distribution of $\mathbf{z}_1$ as a spherical Gaussian:

$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1; 0, \mathbf{I}). \tag{6}$$

We wish to choose a model $p(\mathbf{x}|\mathbf{z}_0)$ that is close to the unknown $q(\mathbf{x}|\mathbf{z}_0)$. Let x_i and $z_{0,i}$ be the i-th elements of $\mathbf{x}, \mathbf{z}_0$, respectively. We then use a factorized distribution of the form:

$$p(\mathbf{x}|\mathbf{z}_0) = \prod_i p(x_i|z_{0,i}), \tag{7}$$

where we choose $p(x_i|z_{0,i}) \propto q(z_{0,i}|x_i)$, which is normalized by summing over all possible discrete values of x_i (256 in the case of 8-bit image data). With sufficiently large $\mathrm{SNR}(0)$, this becomes a very close approximation to the true $q(\mathbf{x}|\mathbf{z}_0)$, as the influence of the unknown data distribution $q(\mathbf{x})$ is overwhelmed by the likelihood $q(\mathbf{z}_0|\mathbf{x})$. Finally, we choose the conditional model distributions as

$$p(\mathbf{z}_s|\mathbf{z}_t) = q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x} = \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)), \tag{8}$$

i.e. the same as the reverse time inference model $q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})$, but with the original data $\mathbf{x}$ replaced by the output of a denoising model $\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)$ that predicts $\mathbf{x}$ from its noisy version $\mathbf{z}_t$. As we will show, this choice of $p(\mathbf{z}_s|\mathbf{z}_t)$ results in a surprisingly simple and numerically stable objective function. The derivation of the means and variances of $p(\mathbf{z}_s|\mathbf{z}_t)$ are given in Appendix A.

3.3 Parameterization of the denoising model with Fourier features

The parameterization of the denoising model and noise schedule is discussed in Appendix B. Prior work on diffusion models has mainly focused on the perceptual quality of generated samples, which emphasizes coarse scale patterns and global consistency of generated images. Here, we optimize for likelihood, which is sensitive to fine scale details and exact values of individual pixels. To capture the fine scale details of the data, we propose adding a set of *Fourier features* to the input of our denoising model: letting $z_{i,j,k}$ be the scalar value in the k-th channel in the (i, j) spatial position of network input $\mathbf{z}_t$, we add additional channels to the input of the denoising model of the form

$$f^n_{i,j,k} = \sin(z_{i,j,k} 2^n \pi), \quad \text{and} \quad g^n_{i,j,k} = \cos(z_{i,j,k} 2^n \pi). \tag{9}$$

These features are high frequency periodic functions that amplify small changes in the input data $\mathbf{z}_t$; see Appendix C for further details. Including these features in the input of our denoising model leads to large improvements in likelihood as demonstrated in Section 5 and Figure 5, especially when combined with a learnable SNR function. We did not observe such improvements when incorporating Fourier features into autoregressive models.

4 The variational lower bound

Similar to the original DPMs [Sohl-Dickstein et al., 2015], we optimize the parameters towards the variational lower bound of the marginal likelihood, also called the variational lower bound (VLB), which is given by

$$-\log p(\mathbf{x}) \leq -\mathrm{VLB}(\mathbf{x}) = \underbrace{D_{KL}(q(\mathbf{z}_1|\mathbf{x})||p(\mathbf{z}_1))}_{\text{Prior loss}} + \underbrace{\mathbb{E}_{q(\mathbf{z}_0|\mathbf{x})}\left[-\log p(\mathbf{x}|\mathbf{z}_0)\right]}_{\text{Reconstruction loss}} + \underbrace{\mathcal{L}_T(\mathbf{x})}_{\text{Diffusion loss}}. \tag{10}$$

The prior loss is a KL divergence between two Gaussians that can be computed in closed form; see Appendix F. The reconstruction loss can be evaluated and optimized using standard reparameterization gradients [Kingma and Welling, 2013]. The diffusion loss, $\mathcal{L}_T(\mathbf{x})$, is more complicated, and depends on the hyperparameter T that determines the depth of the generative model.

Unlike earlier DPMs, but similar to VAEs [Kingma and Welling, 2013, Rezende et al., 2014], we optimize the inference model parameters (that define the forward time diffusion process) jointly with the rest of the model.

4.1 Discrete-time loss

In the case of finite T, using $s(i) = (i-1)/T$, $t(i) = i/T$, the diffusion loss is:

$$\mathcal{L}_T(\mathbf{x}) = \sum_{i=1}^{T} \mathbb{E}_{q(\mathbf{z}_{t(i)}|\mathbf{x})} D_{KL}[q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x})||p(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)})]. \tag{11}$$

In appendix F we show that this expression simplifies considerably, yielding:

$$\mathcal{L}_T(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}), i\sim U\{1,T\}} \left[\frac{T}{2} \left(\mathrm{SNR}(s) - \mathrm{SNR}(t)\right) ||\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)||_2^2 \right], \tag{12}$$

where $U\{1, T\}$ is the uniform distribution on the integers $\{1, \ldots, T\}$, and $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$.

4.2 More steps is better

A natural question to ask is what the number of time segments T should be, and whether more segments is always better. In Appendix G we analyze the difference between the diffusion loss with T segments, $\mathcal{L}_T(\mathbf{x})$, and the diffusion loss with double the segments, $\mathcal{L}_{2T}(\mathbf{x})$, while keeping the SNR function fixed. We then find that if our trained denoising model $\hat{\mathbf{x}}_{\boldsymbol{\theta}}$ is sufficiently good, we have that $\mathcal{L}_{2T}(\mathbf{x}) < \mathcal{L}_T(\mathbf{x})$, i.e. that our VLB will be better for a larger number of time segments. Intuitively, the discrete time diffusion loss is an upper Riemann sum approximation of an integral of a strictly decreasing function, meaning that a finer approximation yields a lower diffusion loss. This result is illustrated in Figure 2.

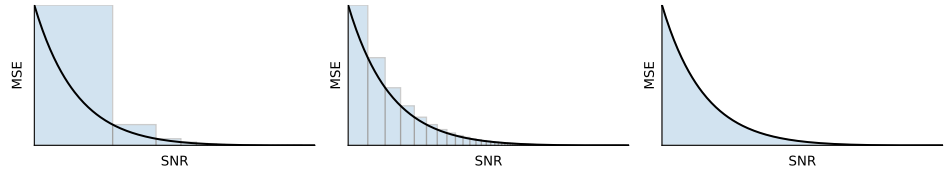


Figure 2: Illustration of the continuous-time (right) and discrete-time diffusion loss with few time segments (left) and more time segments (middle). The continuous-time loss (Equation 17) is an integral of mean squared error (MSE) over SNR, here visualized as a black curve. The black curve is strictly decreasing when the model is sufficiently well trained, so the discrete-time loss (Equation 12) is an upper Riemann sum approximation of this integral that decreases when segments are added.

4.3 Continuous-time loss

Since taking more time steps leads to a better VLB, we now take $T \to \infty$, effectively treating time t as continuous rather than discrete. The model for $p(\mathbf{z}_t)$ can be described as a continuous time diffusion process [Song et al., 2021b] governed by a stochastic differential equation; see Appendix D. In Appendix F we show that in this limit the diffusion loss $\mathcal{L}_T(\mathbf{x})$ simplifies further. Letting $\mathrm{SNR}'(t) = \frac{d\mathrm{SNR}(t)}{dt} = \frac{d\alpha_t^2/\sigma_t^2}{dt}$, we have:

$$\mathcal{L}_\infty(\mathbf{x}) = -\frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})} \int_0^1 \mathrm{SNR}'(t) \, \|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|_2^2 \, dt, \tag{13}$$

$$= -\frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}), t\sim\mathcal{U}(0,1)} \left[\mathrm{SNR}'(t) \, \|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|_2^2 \right]. \tag{14}$$

In terms of predicting the noise $\boldsymbol{\epsilon}$, this can equivalently be written as

$$\mathcal{L}_\infty(\mathbf{x}) = -\frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})} \int_0^1 \text{log-SNR}'(t) \left\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right\|_2^2 dt, \tag{15}$$

$$= -\frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}(0,1)} \left[\text{log-SNR}'(t) \left\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right\|_2^2\right], \tag{16}$$

where $\text{log-SNR}'(t) = \frac{d\log[\text{SNR}(t)]}{dt} = \frac{\text{SNR}'(t)}{\text{SNR}(t)}$.

4.3.1 Invariance to the noise schedule in continuous time

The signal-to-noise function $\text{SNR}(t) = \alpha_t^2/\sigma_t^2$ is invertible due to the monotonicity assumption in Section 3.1. Due to this invertibility, we can perform a change of variables, and make everything a function of $v \equiv \text{SNR}(t)$ instead of t, such that $t = \text{SNR}^{-1}(v)$. Let α_v and σ_v be the functions α_t and σ_t evaluated at $t = \text{SNR}^{-1}(v)$, and correspondingly let $\mathbf{z}_v = \alpha_v\mathbf{x} + \sigma_v\boldsymbol{\epsilon}$. Similarly, we rewrite our noise prediction model as $\tilde{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}, v) \equiv \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}, \text{SNR}^{-1}(v))$. With this change of variables, our continuous-time loss in Equation 13 can equivalently be written as:

$$\mathcal{L}_\infty(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})} \int_{\text{SNR}_{\text{min}}}^{\text{SNR}_{\text{max}}} \left\|\mathbf{x} - \tilde{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_v, v)\right\|_2^2 dv, \tag{17}$$

where instead of integrating w.r.t. time t we now integrate w.r.t. the signal-to-noise ratio v, and where $\text{SNR}_{\text{min}} = \text{SNR}(1)$ and $\text{SNR}_{\text{max}} = \text{SNR}(0)$.

What this equation shows us is that the only effect the functions $\alpha(t)$ and $\sigma(t)$ have on the diffusion loss is through the values $\text{SNR}(t) = \alpha_t^2/\sigma_t^2$ at endpoints $t = 0$ and $t = 1$. Given these values SNR_{max} and SNR_{min}, the diffusion loss is invariant to the shape of function $\text{SNR}(t)$ between $t = 0$ and $t = 1$.

We thus showed that the VLB is only impacted by the function $\text{SNR}(t)$ through its endpoints $\text{SNR}_{\text{min}}, \text{SNR}_{\text{max}}$. In Appendix E we show that, apart from these endpoints, the choice of α_t and σ_t actually does not matter at all in continuous time: this means that different diffusion processes proposed in the literature, such as the *variance preserving* process with $\sigma^2(t) = 1 - \alpha^2(t)$ [Ho et al., 2020], and the *variance exploding* process with $\alpha(t) = 1$ [Song and Ermon, 2019], actually define the same generative model in continuous time, up to a trivial re-scaling of the latents $\mathbf{z}$.

4.3.2 Practical stochastic approximation of the continuous-time loss

Since calculating the integral $\mathcal{L}_\infty(\mathbf{x})$; or its generalization $\mathcal{L}_\infty(\mathbf{x}, w)$, is not analytically tractable, we use its unbiased Monte Carlo estimator in practice. To do this, we construct the VLB in terms of predicting $\boldsymbol{\epsilon}$, which is equivalent to predicting $\mathbf{x}$, but easier to implement in a numerically stable way:

$$\mathcal{L}_\infty^{MC}(\mathbf{x}, w, \gamma) = \frac{1}{2}\gamma'(t)w(\gamma(t)) \left\|\boldsymbol{\epsilon} - \tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; \gamma(t))\right\|_2^2, \tag{18}$$

with $\gamma(t) = -\log \text{SNR}(t) = \log[\sigma_t^2/\alpha_t^2]$, and $\mathbf{z}_t = \alpha_t\mathbf{x} + \sigma_t\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), t \sim U[0, 1]$. For the models presented in this paper, we further use $w(v) = 1$ as corresponding to the (unweighted) VLB. The resulting VLB estimate can then be optimized using stochastic gradient descent as usual. The noise schedule $\gamma(t)$ influences the variance of our Monte Carlo estimator, which is why we jointly optimize it with the rest of the model as described in Section 4.4.

Here, we also found it helpful to sample time t using a low-discrepancy sampler. When processing a minibatch of k examples $\mathbf{x}^i$, $i \in \{1, \ldots, k\}$, we require k timesteps t^i sampled from a uniform distribution. Instead of sampling these timesteps independently, we sample a single uniform random number $u_0 \sim U[0, 1]$ and then set $t^i = \text{mod}(u_0 + i/k, 1)$. Each t^i now has the correct uniform marginal distribution, but the minibatch of timesteps covers the space in $[0, 1]$ more equally than when sampling independently, which we find to reduce the variance in our VLB estimate.

4.4 Learning the noise schedule

In previous work, the noise schedule $\text{SNR}(t)$ has a fixed form (see Appendix I, Fig. 4a). In contrast, as explained in Section 3.1, we learn this schedule through parameterization $\text{SNR}(t) = \exp(-\gamma_{\boldsymbol{\eta}}(t))$, with monotonic neural network $\gamma_{\boldsymbol{\eta}}(t)$ and parameters $\boldsymbol{\eta}$. In the discrete-time case, we simply optimize

$\boldsymbol{\eta}$ together with the rest of the parameters by maximizing the VLB. We find this to be especially beneficial when T is small, as we show in Section 5.

The continuous-time case is different: As we showed in Section 4.3.1, the continuous-time diffusion loss is invariant to $\mathrm{SNR}(t)$, except for its endpoints $\mathrm{SNR}_{\text{min}}, \mathrm{SNR}_{\text{max}}$. For this case, we therefore only optimize the VLB with respect to these endpoints, and not the parameters $\boldsymbol{\eta}$ of the schedule interpolating between them. The interpolating function $\gamma_{\boldsymbol{\eta}}$ does impact the variance of our stochastic estimate of it, given in Equation 18. We therefore propose to learn $\boldsymbol{\eta}$ by minimizing the variance, as detailed in Appendix H.

5 Experiments

We demonstrate our proposed class of diffusion models, which we call *Variational Diffusion Models* (VDMs), on the CIFAR-10 [Krizhevsky et al., 2009] dataset, and the downsampled ImageNet [Van Oord et al., 2016, Deng et al., 2009] dataset, where we focus on maximizing likelihood. The score models we use closely follow Ho et al. [2020], except that they process the data solely at the original resolution, without any internal downsampling or upsampling. Our score models are also deeper than those used by others in the literature. All reported models incorporate Fourier features (Appendix C) as well as a learnable diffusion specification (Section 4.4). For our result with data augmentation we used random flips, 90-degree rotations, and color channel swapping. Complete details on our model specifications can be found in Appendix B.

Model (Bits per dim on test set)	Type	CIFAR10 no data aug.	CIFAR10 data aug.	ImageNet 32x32	ImageNet 64x64
Previous work					
ResNet VAE with IAF [Kingma et al., 2016]	VAE	3.11			
Very Deep VAE [Child, 2020]	VAE	2.87		3.80	3.52
NVAE [Vahdat and Kautz, 2020]	VAE	2.91		3.92	
Glow [Kingma and Dhariwal, 2018]	Flow		$3.35^{(B)}$	4.09	3.81
Flow++ [Ho et al., 2019a]	Flow	3.08		3.86	3.69
PixelCNN [Van Oord et al., 2016]	AR	3.03		3.83	3.57
PixelCNN++ [Salimans et al., 2017]	AR	2.92			
Image Transformer [Parmar et al., 2018]	AR	2.90		3.77	
SPN [Menick and Kalchbrenner, 2018]	AR				3.52
Sparse Transformer [Child et al., 2019]	AR	2.80			3.44
Routing Transformer [Roy et al., 2021]	AR				3.43
Sparse Transformer + DistAug [Jun et al., 2020]	AR		$2.53^{(A)}$		
DDPM [Ho et al., 2020]	Diff		$3.69^{(C)}$		
EBM-DRL [Gao et al., 2020]	Diff		$3.18^{(C)}$		
Score SDE [Song et al., 2021b]	Diff	2.99			
Improved DDPM [Nichol and Dhariwal, 2021]	Diff	2.94			3.54
Concurrent work					
CR-NVAE [Sinha and Dieng, 2021]	VAE		$2.51^{(A)}$		
LSGM [Vahdat et al., 2021]	Diff	2.87			
ScoreFlow [Song et al., 2021a] (variational bound)	Diff		$2.90^{(C)}$	3.86	
ScoreFlow [Song et al., 2021a] (cont. norm. flow)	Diff	2.83	$2.80^{(C)}$	3.76	
Our work					
VDM (variational bound)	Diff	**2.65**	**$2.49^{(A)}$**	**3.72**	**3.40**

Table 1: Summary of our findings for density modeling tasks, in terms of bits per dimension (BPD) on the test set. Model types are autoregressive (AR), normalizing flows (Flow), variational autoencoders (VAE), or diffusion models (Diff). Our results were obtained using the continuous-time formulation of our model. CIFAR-10 data augmentations are: (A) extensive, (B) small translations, or (C) horizontal flips. The numbers for VDM are variational bounds, and can likely be improved by estimating the marginal likelihood through importance sampling, or through evaluation of the corresponding continuous normalizing flow as done by Song et al. [2021a].

5.1 Likelihood and samples

Table 1 shows our results on modeling the CIFAR-10 dataset, and the downsampled ImageNet dataset. We establish a new state-of-the-art in terms of test set likelihood on all considered benchmarks, by a significant margin. Our model for CIFAR-10 without data augmentation surpasses the previous best result of 2.80 about 10x faster than it takes the Sparse Transformer to reach this, in wall clock time on equivalent hardware.

Our CIFAR-10 model, whose hyper-parameters were tuned for likelihood, results in a FID (perceptual quality) score of 7.41. This would have been state-of-the-art until recently, but is worse than recent diffusion models that specifically target FID scores [Nichol and Dhariwal, 2021, Song et al., 2021b, Ho et al., 2020]. By instead using a weighted diffusion loss, with the weighting function $w(\text{SNR})$ used by Ho et al. [2020] and described in Appendix K, our FID score improves to 4.0. We did not pursue further tuning of the model to improve FID instead of likelihood.

A random sample of generated images from our model is provided in Figure 3. We provide additional samples from this model, as well as our other models for the other datasets, in Appendix M.

Figure 3: Non cherry-picked unconditional samples from our Imagenet 64x64 model, trained in continuous time and generated using $T = 1000$. The model's hyper-parameters and parameters are optimized w.r.t. the likelihood bound, so the model is not optimized for synthesis quality.

5.2 Ablations

Next, we investigate the relative importance of our contributions. In Table 2 we compare our discrete-time and continuous-time specifications of the diffusion model: When evaluating our model with a small number of steps, our discretely trained models perform better by learning the diffusion schedule to optimize the VLB. However, as argued theoretically in Section 4.2, we find experimentally that more steps T indeed gives better likelihood. When T grows large, our continuously trained model performs best, helped by training its diffusion schedule to minimize variance instead.

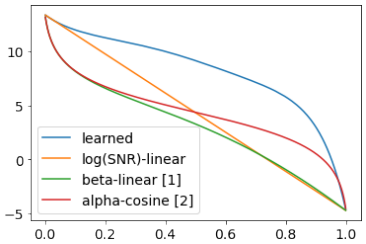
Minimizing the variance also helps the continuous time model to train faster, as shown in Figure 5. This effect is further examined in Table 4b, where we find dramatic variance reductions compared to our baselines in continuous time. Figure 4a shows how this effect is achieved: Compared to the other schedules, our learned schedule spends much more time in the high $\text{SNR}(t)$ / low σ_t^2 range.

In Figure 5 we further show training curves for our model including and excluding the Fourier features proposed in Appendix C: with Fourier features enabled our model achieves much better likelihood. For comparison we also implemented Fourier features in a PixelCNN++ model [Salimans et al., 2017], where we do not see a benefit. In addition, we find that learning the SNR is necessary to get the most out of including Fourier features: if we fix the SNR schedule to that used by Ho et al. [2020], the maximum log-SNR is fixed to approximately 8 (see figure 6), and test set negative likelihood stays above 4 bits per dim. When learning the SNR endpoints, our maximum log-SNR ends up at 13.3, which, combined with the inclusion of Fourier features, leads to the SOTA test set likelihoods reported in Table 1.

5.3 Lossless compression

For a fixed number of evaluation timesteps T_{eval}, our diffusion model in discrete time is a hierarchical latent variable model that can be turned into a lossless compression algorithm using bits-back coding [Hinton and Van Camp, 1993]. As a proof of concept of practical lossless compression,

Table 2 reports net codelengths on the CIFAR10 test set for various settings of T_{eval} using BB-ANS [Townsend et al., 2018], an implementation of bits-back coding based on asymmetric numeral systems [Duda, 2009]. Details of our implementation are given in Appendix N. We achieve state-of-the-art net codelengths, proving our model can be used as the basis of a lossless compression algorithm. However, for large T_{eval} a gap remains with the theoretically optimal codelength corresponding to the negative VLB, and compression becomes computationally expensive due to the large number of neural network forward passes required. Closing this gap with more efficient implementations of bits-back coding suitable for very deep models is an interesting avenue for future work.



(a) log SNR vs time t

SNR(t) schedule	Var(BPD)
Learned (ours)	**0.53**
log SNR-linear	6.35
β-Linear [1]	31.6
α-Cosine [2]	31.1

(b) Variance of VLB estimate

Figure 4: Our learned continuous-time variance-minimizing noise schedule SNR(t) for CIFAR-10, compared to its log-linear initialization and to schedules from the literature: [1] The β-Linear schedule from Ho et al. [2020], [2] The α-Cosine schedule from Nichol and Dhariwal [2021]. All schedules were scaled and shifted on the log scale such that the resulting $\text{SNR}_{\text{min}}, \text{SNR}_{\text{max}}$ were the equal to our learned endpoints, resulting in the same VLB estimate of 2.66. We report the variance of our VLB estimate per data point, computed on the test set, and conditional on the data: This does not include the noise due to sampling minibatches of data.

T_{train}	T_{eval}	BPD	Bits-Back Net BPD
10	10	4.31	
100	100	2.84	
250	250	2.73	
500	500	2.68	
1000	1000	2.67	
10000	10000	2.66	
∞	10	7.54	7.54
∞	100	2.90	2.91
∞	250	2.74	2.76
∞	500	2.69	2.72
∞	1000	2.67	2.72
∞	10000	**2.65**	
∞	∞	**2.65**	

Table 2: Discrete versus continuous-time training and evaluation with CIFAR-10, in terms of bits per dimension (BPD).

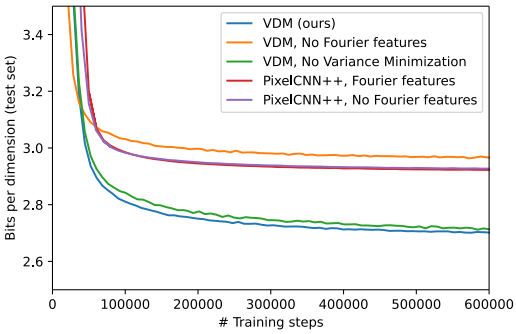


Figure 5: Test set likelihoods during training, with/without Fourier features, and with/without learning the noise schedule to minimize variance.

6 Conclusion

We presented state-of-the-art results on modeling the density of natural images using a new class of diffusion models that incorporates a learnable diffusion specification, Fourier features for fine-scale modeling, as well as other architectural innovations. In addition, we obtained new theoretical insight into likelihood-based generative modeling with diffusion models, showing a surprising invariance of the VLB to the forward time diffusion process in continuous time, as well as an equivalence between various diffusion processes from the literature previously thought to be different.

Our focus on density estimation comes at a slight cost to perceptual quality: uniting great performance on both likelihood and perceptual metrics is a challenge we are keen to address in future work.

Acknowledgments

We thank Yang Song, Kevin Murphy and Mohammad Norouzi for feedback on drafts of this paper.

References

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

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

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

Jarek Duda. Asymmetric numeral systems. *arXiv preprint arXiv:0902.0271*, 2009.

Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pages 5–13, 1993.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730. PMLR, 2019a.

Jonathan Ho, Evan Lohn, and Pieter Abbeel. Compression with flows via local bits-back coding. In *Advances in Neural Information Processing Systems*, pages 3874–3883, 2019b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. A variational perspective on diffusion-based generative models and score matching. *arXiv preprint arXiv:2106.02808*, 2021.

Heewoo Jun, Rewon Child, Mark Chen, John Schulman, Aditya Ramesh, Alec Radford, and Ilya Sutskever. Distribution augmentation for generative modeling. In *International Conference on Machine Learning*, pages 5006–5019. PMLR, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations*, 2013.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

Friso Kingma, Pieter Abbeel, and Jonathan Ho. Bit-swap: Recursive bits-back coding for lossless compression with hierarchical latent variables. In *International Conference on Machine Learning*, pages 3408–3417. PMLR, 2019.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. *arXiv preprint arXiv:1812.01608*, 2018.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

Danilo J Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Samarth Sinha and Adji B Dieng. Consistency regularization for variational auto-encoders. *arXiv preprint arXiv:2105.14859*, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *arXiv e-prints*, pages arXiv–2101, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling Through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021b.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020.

James Townsend, Thomas Bird, and David Barber. Practical lossless compression with latent variables using bits back coding. In *International Conference on Learning Representations*, 2018.

James Townsend, Thomas Bird, Julius Kunze, and David Barber. HiLLoC: lossless image compression with hierarchical latent variable models. In *International Conference on Learning Representations*, 2020.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *arXiv preprint arXiv:2106.05931*, 2021.

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756, 2016.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.