Diffusion Probabilistic Models for Super Resolution Microscopy

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

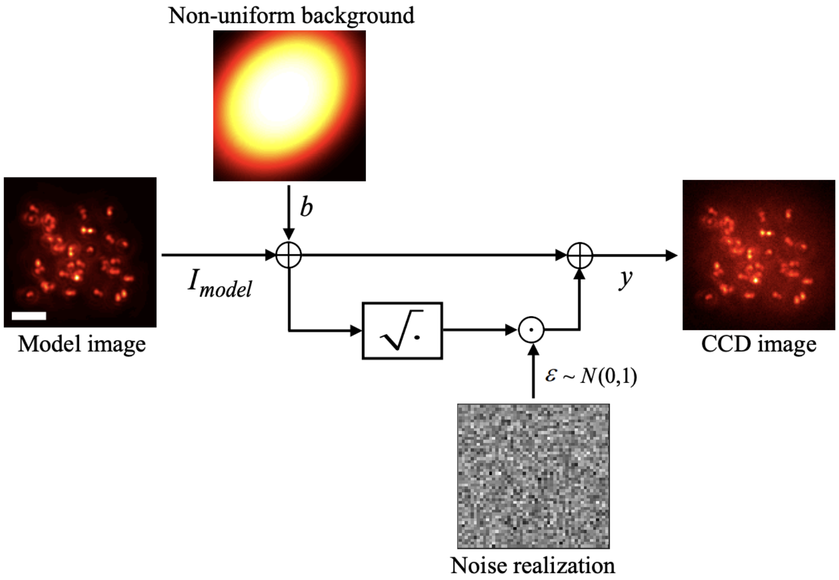
Single-molecule localization microscopy (SMLM) techniques are a mainstay of fluorescence microscopy and can be used to produce a pointillist representation of living cells at diffraction-unlimited precision. Classical SMLM approaches leverage the deactivation of fluorescent tags, followed by spontaneous or photoinduced reactivation, which can be used to estimate of the density of a tagged biomolecule in cellular compartments. Standard SMLM localization algorithms based on maximum likelihood estimators or least squares optimization require tight control of activation and reactivation to maintain sparse emitters, presenting a tradeoff between imaging speed and labeling density. Deep models have generalized SMLM to densely labeled structures, yet uncertainty quantification is still lacking. Recently, denoising diffusion probabilstic models (DDPMs) have been adapted conditional super resolution tasks, demonstrating promising results in detail reconstruction, while directly providing uncertainties in model predictions. Here, we adapt DDPM to the task of single molecule localization, and demonstrate that DDPM approaches the Cramer-Rao lower bound on localization uncertainty over a wide range of experimental conditions.

1 Introduction

Single molecule localization microscopy (SMLM) relies on the temporal resolution of fluorophores whose spatially overlapping point spread functions would otherwise render them unresolvable at the detector. Common strategies for the temporal separation of molecules involve transient intramolecular rearrangements to switch from dark to fluorescent states or the exploitation of non-emitting molecular radicals. Estimation of molecular coordinates in SMLM is acheived by modeling the optical impulse response of the imaging system. However, dense localization suffers from the curse of dimensionality - the parameter space volume grows exponentially with the number of molecules, which is often unknown a priori. Exploration of this high dimensional parameter space in dense SMLM is often intractable.

Previous approaches to this issue has been to predict super-resolution images from a sparse set of localizations with conditional generative adversarial networks (Ouyang 2018) or direct prediction of coordinates using deep neural networks (Nehme 2020; Speiser 2021). However, diffusion models are an appealing alternative because they infer a distribution of deconvolved images that are compatible with an observation. Although conditional VAEs and conditional GANs can provide a distribution of deconvolved images, both are known to suffer from mode collapse and produce insufficient diversity in their outputs. Diffusion models are a recently developed alternative to VAEs and GANs that excel at producing diverse samples and have been successfully applied to solve inverse problems. Here, we present a novel diffusion model for deconvolution in single molecule localization microscopy.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.



Denoising diffusion probabilistic models (DDPM) have emerged as powerful generative models, exceeding GANs and VAEs in a variety of generative modeling tasks. Nevertheless, learning diffusion models directly in data space can limit expressivity of the model (Vahdat 2021). Therefore, we build on previous approaches by using a CNN to compute a latent representation $\mathbf{z}_i$. A denoising diffusion probabilistic model (DDPM) is then used to model the distribution $P_\Phi(\mathbf{y}|\mathbf{z})$.

2 Denoising Diffusion Probabilistic Model for SMLM

We consider datasets $(\theta_i, \mathbf{x}_i, \mathbf{y}_i)_{i=1}^N$ of observed images $\mathbf{x}_i$ and kernel density estimate (KDE) images $\mathbf{y}_i$, given an underlying set of object coordinates θ_i. Observations $\mathbf{x}_i$ are generated from $\theta_i = (r_1, ..., r_N)$ under an image degradation model F. We aim to develop a framework for sampling from $p(\mathbf{y}_i|\mathbf{x}_i)$ and inference of θ_i, while fulfilling a resolution criterion under the condition $|r_i - r_j| \geq \epsilon; \forall(i,j)$.

2.1 Degradation Model

The central objective of single molecule localization microscopy is to infer a set of molecular coordinates θ from noisy, low resolution images $\mathbf{x}$. We define an abstract image stochastic degradation function F such that $\mathbf{x} = F(\theta)$. In the following paragraphs, we define such a function F.

In fluorescence microscopy, each pixel follows Poisson statistics, with expected value

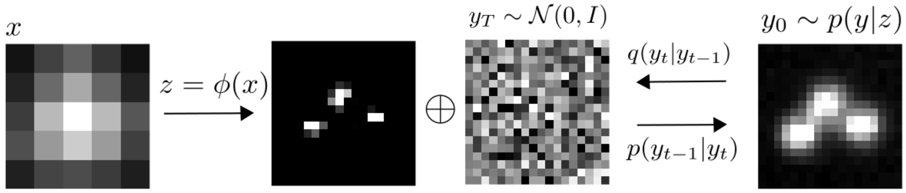
$$\omega = i_0 \int O(u)du \int O(v)dv \tag{1}$$

where $i_0 = \eta N_0 \Delta$. The optical impulse response $O(u,v)$ is often approximated as a 2D isotropic Gaussian with standard deviation σ (Zhang 2007). The parameter η is the photon detection probability of the sensor and Δ is the exposure time. N_0 represents the number of photons emitted.

For a fluorescent emitter located at $\theta = (u_0, v_0)$, we have that

$$\int O(u)du = \frac{1}{2}\left(\mathrm{erf}\left(\frac{u_k + \frac{1}{2} - u_0}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{u_k - \frac{1}{2} - u_0}{\sqrt{2}\sigma}\right)\right) \tag{2}$$

where we have used the common definition $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^t e^{-t^2}dt$. For the sake of generality, the number of photoelectrons at a pixel k, $\mathbf{s}_k$, is multiplied by a gain factor g_k $[\mathrm{ADU}/e^-]$, which is often



unity. The readout noise per pixel ζ_k can be Gaussian with some pixel-specific offset o_k and variance σ_k^2. Ultimately, we have a Poisson component of the signal, which scales with N_0 and may have Gaussian component, which does not. Therefore, in a single exposure, we measure:

$$\mathbf{x}_t = \mathbf{s}_t + \zeta \tag{3}$$

What we are after is the likelihood $p(\mathbf{x}_t|\theta)$ where θ are the molecular coordinates. Fundamental probability theory states that the distribution of $\mathbf{x}_k$ is the convolution of the distributions of $\mathbf{s}_k$ and ζ_k,

$$p(\mathbf{x}_t|\theta) = A \sum_{q=0}^{\infty} \frac{1}{q!} e^{-\omega_k} \omega_k^q \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(\mathbf{x}_k - g_k q - o_k)}{2\sigma_k^2}} \tag{4}$$

where $P(\zeta_k) = \mathcal{N}(o_k, \sigma_k^2)$ and $P(S_k) = \text{Poisson}(g_k \omega_k)$, A is some normalization constant. In practice, (4) is difficult to work with, so we look for an approximation. We will use a Poisson-Normal approximation for simplification. Consider,

$$\zeta_k - o_k + \sigma_k^2 \sim \mathcal{N}(\sigma_k^2, \sigma_k^2) \approx \text{Poisson}(\sigma_k^2) \tag{5}$$

Since $\mathbf{x}_k = \mathbf{s}_k + \zeta_k$, we transform $\mathbf{x}'_k = \mathbf{x}_k - o_k + \sigma_k^2$, which is distributed according to

$$\mathbf{x}'_k \sim \text{Poisson}(\omega'_k) \tag{6}$$

where $\omega'_k = g_k \omega_k + \sigma_k^2$. This result can be seen from the fact the the convolution of two Poisson distributions is also Poisson. The quality of this approximation will degrade with decreasing signal level, since the Poisson distribution does not retain its Gaussian shape at low expected counts. Nevertheless, the quality of the approximation can be predicted by the Komogonov distance between the convolution distribution (4).

2.2 Fisher Information Metric for Localization

Inversion of the degradation function F is generally intractable, particularly when fluorescent molecules are dense within the field of view. This difficulty arises because the parameter θ is typically of large and unknown dimension, rendering maximum likelihood estimation or Markov Chain Monte Carlo sampling computationally difficult. Previous solutions to this problem leverage convolutional neural networks (CNNs) to infer coordinates directly by learning a deterministic image transformation F^{-1}, which we refer to as a "localization map" (Nehme 2021). Such methods faithfully capture the information content in degraded images; however, such methods apply arbitrary thresholding to the CNN localization map, potentially creating erroneous localizations, and do not permit sampling.

We seek a generative approach, which casts localization as an image restoration problem, where a high resolution kernel density estimate $\mathbf{y}$ is reconstructed from a low resolution image $\mathbf{x}$. Building on previous efforts, we utilize a CNN learns a representation which compresses $\mathbf{x}$ while preserving the relevant information to the prediction of $\mathbf{y}$. We use the Fisher information as the information theoretic criteria (Chao 2016). The generative model (6) is also convenient for computing the Fisher information matrix (Smith 2010) and thus the Cramer-Rao lower bound, which bounds the variance of a statistical estimator of θ, from below. The Fisher information is

$$\mathcal{I}_{ij}(\theta) = \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_i}\frac{\partial \ell}{\partial \theta_j}\right) = \sum_k \frac{1}{\omega'_k}\frac{\partial \omega'_k}{\partial \theta_i}\frac{\partial \omega'_k}{\partial \theta_j} \tag{7}$$

where the log-likelihood is $\ell(\mathbf{x}_t|\theta)$.

3 Image Restoration Model

3.1 The Encoder Network

3.2 Optimization of the Encoder Network

3.3 Conditional Denoising Diffusion Model

Given datasets $(\theta_i, \mathbf{x}_i, \mathbf{y}_i)_{i=1}^{N}$ which represent samples drawn from an unknown conditional distribution $p(\mathbf{y}|\mathbf{x})$. This is a one-to-many mapping in which many target images may be consistent with an input image. The conditional DDPM model generates a target image y_0 in T refinement steps. Starting with a pure noise image $y_T \sim \mathcal{N}(0, I)$, the model iteratively refines the image through successive iterations according to learned conditional transition distributions $p(y_{t-1}|y_t, x)$ such that $y_0 \sim p(\mathbf{y}|\mathbf{x})$

3.4 Gaussian Diffusion Model

The *forward* process is the joint distribution $p_\theta(\mathbf{y}_{0:T})$, which is Markovian.

$$q(\mathbf{y}_t|\mathbf{y}_0) = \prod_{t=1}^{T} q(\mathbf{y}_t|\mathbf{y}_{t-1}) \quad q(\mathbf{y}_t|\mathbf{y}_{t-1}) = \mathcal{N}\left(\mathbf{y}_{t-1}, \sqrt{\alpha_t}\mathbf{y}_{t-1}, (1-\alpha_t)I\right) \tag{8}$$

We optimize a denoising model f_θ which takes as input an interpolated low-resolution input $\mathbf{y}$ and a noisy input $\mathbf{y}_T$.

$$p_\theta(\mathbf{y}_{0:T}) = p_\theta(\mathbf{y}_T)\prod_{t=1}^{T} p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) \quad p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) = \mathcal{N}\left(\mathbf{y}_{t-1}, \mu_\theta(\mathbf{y}_t, \gamma_t), \sigma_t^2 I\right) \tag{9}$$

where $\gamma_t = \prod_{i=1}^{t} \alpha_t$. Note that the model θ is not a function of t. The mean of the transition density reads

$$\mu_\theta(\mathbf{x}_t, \mathbf{y}, \gamma_t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\mathbf{x}_t, \gamma_t)\right) \tag{10}$$

Recall that the denoising model f_θ is trained to estimate ϵ, given any noisy image y_e including y_t. Thus, given y_t, we approximate y_0 by rearranging the terms in (5) as

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

Following the formulation of **?**, we substitute our estimate $\hat{y}_0$ into the posterior distribution of $q(y_{t-1}|y_0, y_t)$ in (4) to parameterize the mean of $p_\theta(y_{t-1}|y_t, x)$ as

$$\mu_\theta(x, y_t, \gamma_t) = \frac{1}{\sqrt{\alpha_t}}\left(y_t - (1-\alpha_t)\sqrt{1-\gamma_t}\, f_\theta(x, y_t, \gamma_t)\right),$$

and we set the variance of $p_\theta(y_{t-1}|y_t, x)$ to $(1-\alpha_t)$, a default given by the variance of the forward process **?**. Following this parameterization, each iteration of iterative refinement under our model takes the form,

$$y_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(y_t - (1-\alpha_t)\sqrt{1-\gamma_t}\, f_\theta(x, y_t, \gamma_t)\right) + \sqrt{1-\alpha_t}\epsilon_t,$$

where $\epsilon_t \sim \mathcal{N}(0, I)$. This resembles one step of Langevin dynamics with f_θ providing an estimate of the gradient of the data log-density. We justify the choice of the training objective in (6) for the probabilistic model outlined in (9) from a variational lower bound perspective and a denoising score-matching perspective in Appendix B.

3.5 Optimization of the Denoising Model

To help reverse the diffusion process, we take advantage of additional side information in the form of a source image x and optimize a neural denoising model f_θ that takes as input this source image x and a noisy target image y_e,

$$y_e = \sqrt{\gamma} y_0 + \sqrt{1 - \gamma} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

and aims to recover the noiseless target image y_0. This definition of a noisy target image y_e is compatible with the marginal distribution of noisy images at different steps of the forward diffusion process.

In addition to a source image x and a noisy target image y_e, the denoising model $f_\theta(x, y_e, \gamma)$ takes as input the sufficient statistics for the variance of the noise γ, and is trained to predict the noise vector ϵ. We make the denoising model aware of the level of noise through conditioning on a scalar γ, similar to **??**. The proposed objective function for training f_θ is

$$\mathbb{E}(x, y) \mathbb{E}_{\epsilon, \gamma} \left[f_\theta \left(x, \sqrt{\gamma} y_0 + \sqrt{1 - \gamma} \epsilon \,\middle|\, y_e, \gamma \right) - \epsilon \right],$$

where $\epsilon \sim \mathcal{N}(0, I)$, (x, y) is sampled from the training dataset, $p \in \{1, 2\}$, and $\gamma \sim p(\gamma)$. The distribution of γ has a big impact on the quality of the model and the generated outputs. We discuss our choice of $p(\gamma)$ in Section 2.4.

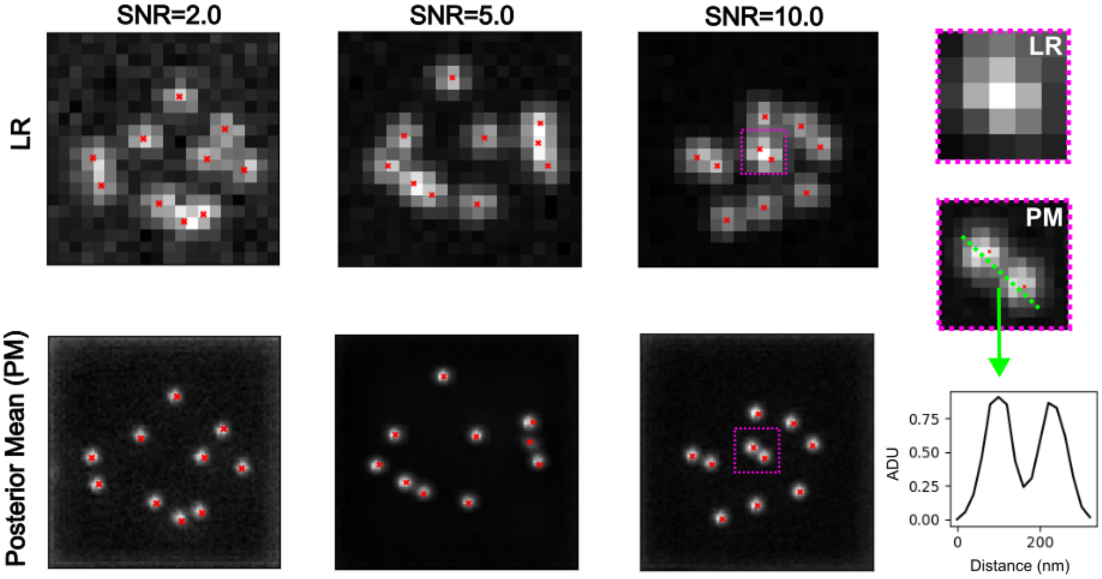
The SR3 architecture is similar to the U-Net found in DDPM **?**, with modifications adapted from **?**; we replace the original DDPM residual blocks with residual blocks from BigGAN **?**, and we re-scale skip connections by $\sqrt{\frac{1}{2}}$. We also increase the number of residual blocks, and the channel multipliers at different resolutions (see Appendix A for details). To condition the model on the input x, we up-sample the low-resolution image to the target resolution using bicubic interpolation. The result is concatenated with y_t along the channel dimension. We experimented with more sophisticated methods of conditioning, such as using FiLM **?**, but we found that the simple concatenation yielded similar generation quality.

For our training noise schedule, we follow **?**, and use a piecewise distribution for γ, $p(\gamma) = \frac{1}{T} \sum_{t=1}^{T} U(\gamma_{t-1}, \gamma_t)$. Specifically, during training, we first uniformly sample a time step $t \sim \{0, ..., T\}$ followed by sampling $\gamma \sim U(\gamma_{t-1}, \gamma_t)$. We set $T = 2000$ in all our experiments.

Prior work of diffusion models **??** require 1-2k diffusion steps during inference, making generation slow for large target resolution tasks. We adapt techniques from **?** to enable more efficient inference. Our model conditions on γ directly (vs t as in **?**), which allows us flexibility in choosing the number of diffusion steps, and the noise schedule during inference. This has been demonstrated to work well for speech synthesis **?**, but has not been explored for images. For efficient inference, we set the maximum inference budget to 100 diffusion steps, and hyper-parameter search over the inference noise schedule. This search is inexpensive as we only need to train the model once **?**. We use FID on held-out data to choose the best noise schedule, as we found PSNR did not correlate well with image quality.

4 Experiments

5 Related Work


SNR=2.0
SNR=5.0
SNR=10.0
LR
Posterior Mean (PM)
LR
PM
ADU
0.75
0.50
0.25
0.00
0
200
Distance (nm)