
Conditional Diffusion Probabilistic Models for Super Resolution Microscopy

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

17 1 Introduction

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

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

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

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

50 We seek a generative approach, which casts localization as an image restoration problem, where a
 51 high resolution kernel density estimate \mathbf{y} is reconstructed from a low resolution image \mathbf{x} . Building
 52 on previous efforts, we utilize a CNN learns a representation which compresses \mathbf{x} while preserving
 53 the relevant information to the prediction of \mathbf{y} .

54 2 Denoising Diffusion Probabilistic Model for SMLM

55 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)
 56 images \mathbf{y}_i , given an underlying set of object coordinates θ_i . Observations \mathbf{x}_i are generated from
 57 $\theta_i = (r_1, \dots, r_N)$ under an image degradation model F . We aim to develop a framework for
 58 sampling from $p(\mathbf{y}_i|\mathbf{x}_i)$ and inference of θ_i , while fulfilling a resolution criterion under the condition
 59 $|r_i - r_j| \geq \epsilon; \forall(i, j)$.

60 2.1 Degradation Model

61 The central objective of single molecule localization microscopy is to infer a set of molecular
 62 coordinates θ from noisy, low resolution images \mathbf{x} . We therefore begin by defining the likelihood on
 63 measured low-resolution images $p(\mathbf{x}|\theta)$. In fluorescence microscopy, each pixel is a Poisson random
 64 variable (Smith 2010; Nehme 2020; Chao 2016), with expected value

$$\omega = i_0 \int O(u) du \int O(v) dv \quad (1)$$

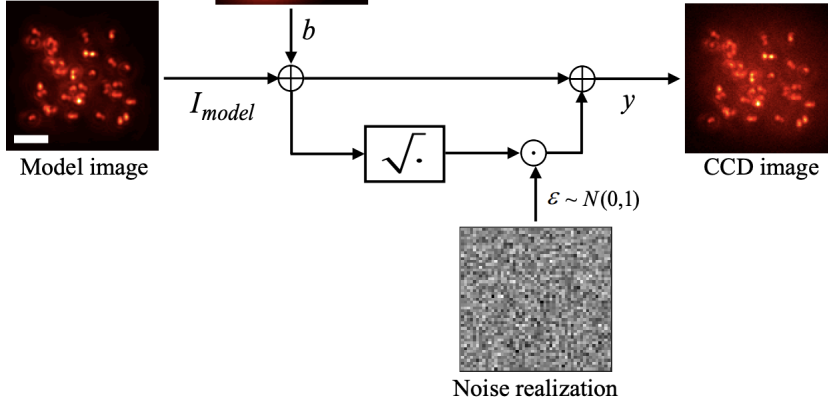
65 where $i_0 = \eta N_0 \Delta$. The scalar parameters η, Δ are the photon detection probability of the sensor and
 66 the exposure time, respectively. Without loss of generality, we assume $\eta = \Delta = 1$. Most importantly,
 67 N_0 represents the signal amplitude, which we assume maintains a fixed value. The optical impulse
 68 response $O(u, v)$ is often approximated as a 2D isotropic Gaussian with standard deviation σ (Zhang
 69 2007). This approximation has the convenient property, that the effects of pixelation can be expressed
 70 in terms of error functions. For example, given a fluorescent emitter located at $\theta = (u_0, v_0)$, we have
 71 that

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

72 where we have used the common definition $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt$. Our generative model also
 73 incorporates a normally distributed white noise per pixel ζ with offset o and variance σ^2 . Ultimately,
 74 we have a Poisson component of the signal, which scales with N_0 and a Gaussian component, which
 75 does not. Therefore, in a single exposure, we measure:

$$\mathbf{x} = \mathbf{s} + \zeta \quad (3)$$

76 The distribution of \mathbf{x} is the convolution of the distributions of \mathbf{s} and ζ ,



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

where $p(\zeta_k) = \mathcal{N}(o_k, \sigma_k^2)$ and $p(s_k) = \text{Poisson}(\omega_k)$, 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,

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

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

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

where $\omega'_k = \omega_k + \sigma_k^2$. This result can be seen from the fact the the convolution of two Poisson distributions is also Poisson. We then arrive at the following log likelihood

$$\ell(\mathbf{x}|\theta) = -\log \prod_k \frac{e^{-(\mu'_k)} (\mu'_k)^{n_k}}{n_k!} \approx \sum_k n_k \log n_k + \mu'_k - n_k \log (\mu'_k) \quad (7)$$

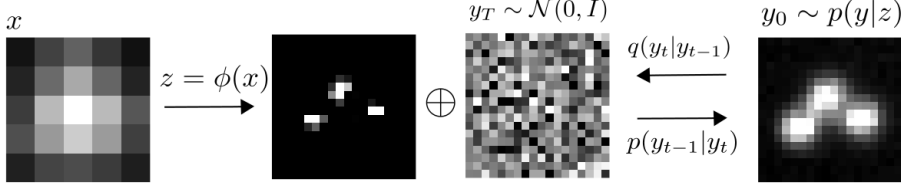
2.2 Fisher Information Metric for Quantitative Error Analysis

We use the Fisher information as an information theoretic criteria to assess the quality of the proposed algorithms, with respect to the root mean squared error (RMSE) of our predictions of θ . The generative model $\ell(\mathbf{x}|\theta)$ 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 i.e., $\text{var}(\hat{\theta}) \geq I^{-1}(\theta)$. It is shown in the appendix, that the Fisher information is straightforward to compute under the Poisson likelihood (7)

$$\mathcal{I}_{ij}(\theta) = \mathbb{E}_{\theta} \left(\frac{\partial \ell}{\partial \theta_i} \frac{\partial \ell}{\partial \theta_j} \right) = \sum_k \frac{1}{\omega'_k} \frac{\partial \omega'_k}{\partial \theta_i} \frac{\partial \omega'_k}{\partial \theta_j} \quad (8)$$

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 $y_0 \sim p(\mathbf{y}|\mathbf{x})$



97 3.1 Gaussian Diffusion

98 Diffusion models (Sohl-Dickstein 2015; Ho 2020) are a class of generative models inspired by
 99 nonequilibrium statistical physics, which slowly destroy structure in a data distribution $p(\mathbf{y}_0|\mathbf{x})$ via
 100 a fixed Markov chain referred to as the *forward process*. In essence, the forward process gradually
 101 adds Gaussian noise to the data according to a variance schedule $\beta_{0:T}$

$$q(\mathbf{y}_t|\mathbf{y}_0) = \prod_{t=1}^T q(\mathbf{y}_t|\mathbf{y}_{t-1}) \quad q(\mathbf{y}_t|\mathbf{y}_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{y}_{t-1}, \beta_t I\right) \quad (9)$$

102 An important property of the forward process is that it admits sampling x_t at an arbitrary timestep t
 103 in closed form (Ho 2020). Using the notation $\alpha_t := 1 - \beta_t$ and $\gamma_t := \prod_{s=1}^t \alpha_s$, we have

$$q(\mathbf{y}_t|\mathbf{y}_0) = \mathcal{N}(\sqrt{\gamma_t}\mathbf{y}_0, (1 - \gamma_t)I) \quad (10)$$

104 The usual procedure is then to learn a parametric representation of the *reverse process*, and therefore
 105 generate samples from $p(\mathbf{y}_0)$, starting from noise. Here, we are concerned with conditional diffusion
 106 models, which instead sample from a conditional distribution $p(\mathbf{y}_0|\mathbf{x})$. Formally, $p_\theta(\mathbf{y}_0|\mathbf{x}_0) =$
 107 $\int p_\theta(\mathbf{y}_{0:T}|\mathbf{x}_0)d\mathbf{x}_{1:T}$ where y_t is a latent representation with the same dimensionality of the data.
 108 $p_\theta(\mathbf{y}_{0:T}|\mathbf{x})$ is a Markov process, starting from a noise sample $p_\theta(y_T) = \mathcal{N}(0, I)$.

$$p_\theta(\mathbf{y}_{0:T}) = p_\theta(\mathbf{y}_T) \prod_{t=1}^T p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) \quad p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) = \mathcal{N}(\mu_\theta(\mathbf{y}_t), \beta_t I) \quad (11)$$

109 where we reuse the variance schedule of the forward process (Ho 2020). We seek to learn a denoising
 110 model μ_θ which computes the mean of the Gaussian transition density at each time step t . However,
 111 learning diffusion models directly in data space can limit expressivity of the model (Vahdat 2021).
 112 Since we are primarily interested in learning a restoration \mathbf{y} , we choose to define an encoder ϕ such
 113 that $\mathbf{z} = \phi(\mathbf{x}_0)$. The reverse process then becomes $p_\theta(\mathbf{y}_0|\mathbf{z} = \phi(\mathbf{x}_0)) = \int p_\theta(\mathbf{y}_{0:T}|\mathbf{z})d\mathbf{x}_{1:T}$. For all
 114 $t > 0$, the mean of the transition density is computed as

$$\mu_\theta(\mathbf{y}_t, \mathbf{x}, \gamma_t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{y}_t - \frac{(1 - \alpha_t)}{\sqrt{1 - \gamma_t}} f_\theta(\mathbf{y}, \mathbf{x}, \gamma_t) \right) \quad (12)$$

115 where f_θ is a neural network. Only at $t = 0$ is this mean directly a function of \mathbf{x} .

116 3.2 Optimization of the Denoising Model

117 Here, β_t are treated as hyperparameters, with a linear schedule from $\beta_0 = 10^{-4}$ to $\beta_T = 10^{-2}$ in
 118 T timesteps. To reverse the diffusion process, we follow (Saharia 2021), using a source encoding
 119 $\mathbf{z} = \phi(\mathbf{x})$ and optimize a neural denoising model f_θ that takes as input \mathbf{z} and a noisy target image
 120 $\mathbf{y}_t \sim q(\mathbf{y}_t|\mathbf{y}_0)$,

$$\mathbf{y}_t = \sqrt{\gamma_t}\mathbf{y}_0 + \sqrt{1 - \gamma_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (13)$$

121 This definition of a noisy target image \mathbf{y}_t is drawn from the marginal distribution of noisy images at
 122 a time step t of the forward diffusion process. In addition to a source image \mathbf{y}_0 and a noisy target

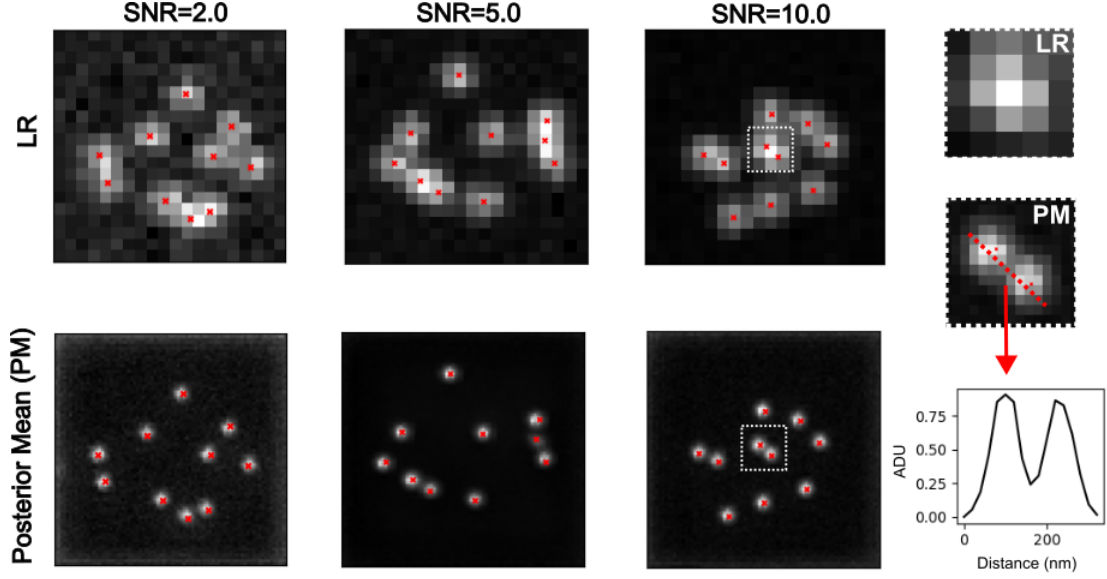


image \mathbf{y}_t , the denoising model f_θ takes as input the sufficient statistics for the variance of the noise γ , and is trained to predict the noise vector ϵ . We make the denoising model aware of the level of noise through conditioning on a scalar γ . The proposed objective function for training f_θ is

$$\mathbb{E}_{(\mathbf{z}, \mathbf{y}_0)} \mathbb{E}_{(\epsilon, \gamma)} \left[f_\theta \left(x, \sqrt{\gamma} \mathbf{y}_0 + \sqrt{1 - \gamma} \epsilon \mid \mathbf{y}_t, \gamma \right) - \epsilon \right], \quad (14)$$

where $\epsilon \sim \mathcal{N}(0, I)$, $(\mathbf{z}, \mathbf{y}_0)$ is sampled from the training dataset and $\gamma \sim p(\gamma)$. The distribution of γ has a big impact on the quality of the model and the generated outputs. For our training noise schedule, we use a piecewise distribution for γ , $p(\gamma) = \frac{1}{T} \sum_{t=1}^T U(\gamma_{t-1}, \gamma_t)$ (Nanxin 2021). Specifically, during training, we first uniformly sample a time step $t \sim \{0, \dots, T\}$ followed by sampling $\gamma \sim U(\gamma_{t-1}, \gamma_t)$. We set $T = 100$ in all our experiments.

4 Experiments

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, but we found that the simple concatenation yielded similar generation quality.

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.

5 Related Work

