Conditional Diffusion Models for Uncertainty Estimation in Super Resolution Microscopy

Anonymous Author(s)
Affiliation
Address
email

Abstract

The field of deep generative modeling for image translation has experienced a boom in research in the past few years, with various tools developed to estimate different types of uncertainty. Yet, many powerful models in image translation are selected and trained under a simple reconstruction loss and therefore are unable report uncertainty in their outputs. Uncertainty quantification is a necessary modeling component in many high-risk applications and in the sciences. In order to quantify uncertainty in otherwise deterministic models, we propose a hybrid generative modeling framework based denoising diffusion probabilistic models (DDPMs). Specifically, our model learns a distribution on the true image latent in the input conditioned on the network output, in order to represent the posterior on reconstructions. Then, we apply this framework to the task of single molecule localization in fluorescence microscopy, and demonstrate that blending the DeepSTORM architecture with a DDPM permits uncertainty quantification of kernel density estimates (KDEs) regressed by DeepSTORM. Our results suggest the proposed solution is an interesting addition to the modeling toolkit for fluorescence microscopists and the field of deep image translation in general.

1 Introduction

Deep learning has attracted tremendous attention from researchers in fields such as physics, biology, and manufacturing, to name a few (Baldi et al., 2014; Anjos et al., 2015; Bergmann et al., 2014). Tools such as neural networks (NNs), dropout, convolutional neural networks (convnets), and others are used extensively. In particular, the application of deep image translation in single-molecule localization microscopy (SMLM) has received considerable interest (Ouyang 2018; Nehme 2020; Speiser 2021). SMLM techniques are a mainstay of fluorescence microscopy and can be used to produce a pointillist representation of living cells at diffraction-unlimited precision. As this technology enables increasingly precise measurements of the cellular environment, there is an increasing need for machine learning methods to report uncertainty for quality control.

In previous applications of deep models to localization microscopy, super-resolution images can be recovered from a sparse set of localizations with conditional generative adversarial networks (Ouyang 2018) or kernel density estimation can be performed using convolutional networks (Nehme 2020; Speiser 2021). Here, we focus on the latter class of models which perform single molecule localization using neural networks. In this approach, one estimates molecular coordinates by predicting kernel density estimates (KDEs) $\mathbf{y}$, which are latent in the raw data $\mathbf{x}$, using a convolutional neural network. Importantly, inferences in SMLM are often necessarily made on a single measurement, thus common measures of model performance are based on localization errors computed over ensembles of simulated images. However, this choice precludes computation of aleatoric uncertainty at test time under a fixed model, and may result in the application of models to out of distribution datasets.

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

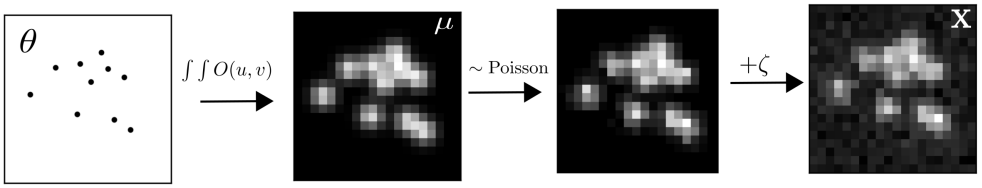

Figure 1: Generative model of single molecule localization microscopy images

Bayesian probability theory offers us mathematically grounded tools to reason about model uncertainty, but these usually come with a prohibitive computational cost. Recent approaches to avoiding this intractiblity have been deterministic uncertainty quantification (Amersfoort 2020), ensembling (Lakshminarayanan et al., 2017) or Monte Carlo sampling (Gal and Ghahramani, 2016). Here, we report a method which models estimates uncertainty in KDE predictions by learning a distribution on the true image latent in the input conditioned on the network output, in order to represent the posterior on reconstructions. Our approach preserves image structure and produces pixel-wise uncertainties, which can be used for out of distribution sample detection or filtering. Our approach could be readily integrated with existing localization performance measures to address both model accuracy on training data and precision on datasets produced by experiments.

2 Background

2.1 Image Likelihood and Localization Error

The central objective of single molecule localization microscopy is to infer a set of molecular coordinates from measured low resolution images $\mathbf{x}$. The likelihood on measured low-resolution images $p(\mathbf{x}|\theta)$ is taken to be a convolution of Gaussian and Poisson PDFs

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

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.

$$\mathbf{x}'_k \sim \text{Poisson}(\omega'_k) \tag{2}$$

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 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)

$$\mathcal{I}_{ij}(\theta) = \mathop{\mathbb{E}}_{\theta}\left(\frac{\partial\ell}{\partial\theta_i}\frac{\partial\ell}{\partial\theta_j}\right) = \sum_k \frac{1}{\omega'_k}\frac{\partial\omega'_k}{\partial\theta_i}\frac{\partial\omega'_k}{\partial\theta_j} \tag{3}$$

2.2 Kernel density estimation with deep networks

Direct optimization of the log-likelihood in (7) from observations $\mathbf{x}$ alone is challenging when fluorescent emitters are dense within the field of view and fluorescent signals significantly overlap. Convolutional neural networks (CNN) have recently been used in fluorescence microscopy to extract

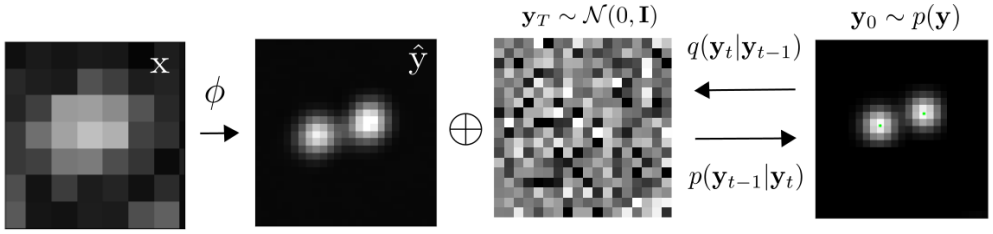

Figure 2: Conditional diffusion model for sampling kernel density estimates

parameters describing fluorescent emitters such as color, emitter orientation, z coordinate, background signal. For localization tasks, CNNs typically employ upsampling layers to reconstruct Bernoulli probabilities of emitter occupancy or kernel density estimates with higher resolution than experimental measurements. Kernel density estimates are the most common data structure used in SMLM, and can be easlily generated from molecular coordinates using the optical impulse response (2). While ground-truth data from experiments to train the neural network are typically not available, synthetic training data can also be generated by (2) and simulation from the Poisson likelihood (7).

The DeepSTORM CNN, initially proposed in [1] for 3D localization, can be viewed as a deep kernel density estimator, reconstructing kernel density estimates $\mathbf{y}$ from low-resolution inputs $\mathbf{x}$. We utilize a simplified form of the original architecture for 2D localization, which we denote ϕ hereafter, which consists of three main modules: a multi-scale context aggregation module, an upsampling module, and a prediction module. For context aggregation, the architecture utilizes dilated convolutions to increase the receptive field of each layer. The upsampling module is then composed of two consecutive 2x resize-convolutions, computed by nearest-neighbor interpolation, to increase the lateral resolution by a factor of 4. For a common sCMOS camera, each pixel has a lateral size of approximately 108 nanometers, giving approximately 27 nanometer pixels in the KDE. The terminal prediction module contains three additional convolutional blocks for refinement of the upsampled image, followed by an element-wise HardTanh.

3 Denoising Diffusion Probabilistic Model for SMLM

We consider datasets $(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}_i)_{i=1}^N$ of observed images $\mathbf{x}_i$ true kernel density estimate (KDE) images $\mathbf{y}_i$, and KDE estimates $\hat{\mathbf{y}}_i = \phi(\mathbf{x}_i)$. Observations $\mathbf{x}_i$ are generated under the image degradation model. We aim to develop a framework for sampling from $p(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y})$.

4 Conditional Denoising Diffusion Model

Point estimates $\hat{\mathbf{y}}_i$ produced by the DeepSTORM architecture lack uncertainty quantification. To address this, we propose a DDPM to model the conditional distribution $p(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y})$. Consider the factorization $p(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y})p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{x}|\mathbf{y}, \hat{\mathbf{y}})p(\mathbf{y}|\hat{\mathbf{y}})p(\hat{\mathbf{y}})$. Given that $\mathbf{x}$ is conditionally independent of $\hat{\mathbf{y}}$, we find

$$p_\Psi(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\hat{\mathbf{y}})$$

Evidently, the DDPM Ψ can be trained on pairs $(\mathbf{y}_i, \hat{\mathbf{y}}_i)_{i=1}^N$. The conditional DDPM generates a target KDE $\mathbf{y}_0$ in T refinement steps. Starting with a pure noise image $\mathbf{y}_T \sim \mathcal{N}(0, \mathbf{I})$, the model iteratively refines the KDE through successive iterations according to learned conditional transition distributions $p(\mathbf{y}_{t-1}|\mathbf{y}_t,)$ such that $\mathbf{y}_0 \sim p(\mathbf{y}|\hat{\mathbf{y}})$

4.1 Gaussian Diffusion

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

$$q(\mathbf{y}_t|\mathbf{y}_0) = \prod_{t=1}^{T} q(\mathbf{y}_t|\mathbf{y}_{t-1}) \quad q(\mathbf{y}_t|\mathbf{y}_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{y}_{t-1}, \beta_t I\right) \tag{4}$$

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

$$q(\mathbf{y}_t|\mathbf{y}_0) = \mathcal{N}\left(\sqrt{\gamma_t}\mathbf{y}_0, (1-\gamma_t)I\right) \tag{5}$$

The usual procedure is then to learn a parametric representation of the *reverse process*, and therefore generate samples from $p(\mathbf{y}_0)$, starting from noise. Formally, $p_\theta(\mathbf{y}_0|\hat{\mathbf{y}}) = \int p_\theta(\mathbf{y}_{0:T}|\hat{\mathbf{y}}) d\hat{\mathbf{y}}_{1:T}$ where $\mathbf{y}_t$ is a latent representation with the same dimensionality of the data. $p_\theta(\mathbf{y}_{0:T}|\hat{\mathbf{y}})$ is a Markov process, starting from a noise sample $p_\theta(\mathbf{y}_T) = \mathcal{N}(0, \mathbf{I})$.

$$p_\theta(\mathbf{y}_{0:T}) = p_\theta(\mathbf{y}_T) \prod_{t=1}^{T} p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) \quad p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) = \mathcal{N}\left(\mu_\theta(\mathbf{y}_t), \beta_t I\right) \tag{6}$$

where we reuse the variance schedule of the forward process (Ho 2020). We seek to learn a denoising model μ_θ which computes the mean of the Gaussian transition density at each time step t. For all $t > 0$, the mean of the transition density is computed as

$$\mu_\theta(\mathbf{y}_t, \hat{\mathbf{y}}, \gamma_t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{y}_t - \frac{(1-\alpha_t)}{\sqrt{1-\gamma_t}} f_\theta(\mathbf{y}, \hat{\mathbf{y}}, \gamma_t) \right) \tag{7}$$

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

4.2 Optimization of the Denoising Model

To reverse the diffusion process, we optimize a neural denoising model f_θ that takes as input $\hat{\mathbf{y}}$ and a noisy target image $\mathbf{y}_t \sim q(\mathbf{y}_t|\mathbf{y}_0)$. That is, this noisy target image $\mathbf{y}_t$ is drawn from the marginal distribution of noisy images at a time step t of the forward diffusion process.

$$\mathbf{y}_t = \sqrt{\gamma}\mathbf{y}_0 + \sqrt{1-\gamma}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{8}$$

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

$$\mathop{\mathbb{E}}_{(\hat{\mathbf{y}}, \mathbf{y}_0)} \mathop{\mathbb{E}}_{(\epsilon, \gamma)} \left[f_\theta \left(x, \sqrt{\gamma}\mathbf{y}_0 + \sqrt{1-\gamma}\epsilon \,\middle|\, \mathbf{y}_t, \gamma \right) - \epsilon \right], \tag{9}$$

where $(\hat{\mathbf{y}}, \mathbf{y}_0)$ 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 $t \sim \{0, ..., T\}$ followed by sampling $\gamma \sim U(\gamma_{t-1}, \gamma_t)$. We set $T = 100$ in all our experiments.

4.3 Optimization of the DeepSTORM architecture

A first pass at localization treats localization as a binary classification problem, such that 0 denotes a vacant pixel and 1 denotes an occupied pixel containing an emitter. Direct learning of pixel-wise classification with cross-entropy loss leads to an imbalance of occupied and unoccupied pixels in dense localization problems (Nehme 2020). CE loss is usually either weighted [51], replaced with a Focal loss [52], or applied to a "blobbed" version of the desired boolean volume e.g. by placing a disk around each GT position [53–55]. Alternative methods take a soft version of the binary classification

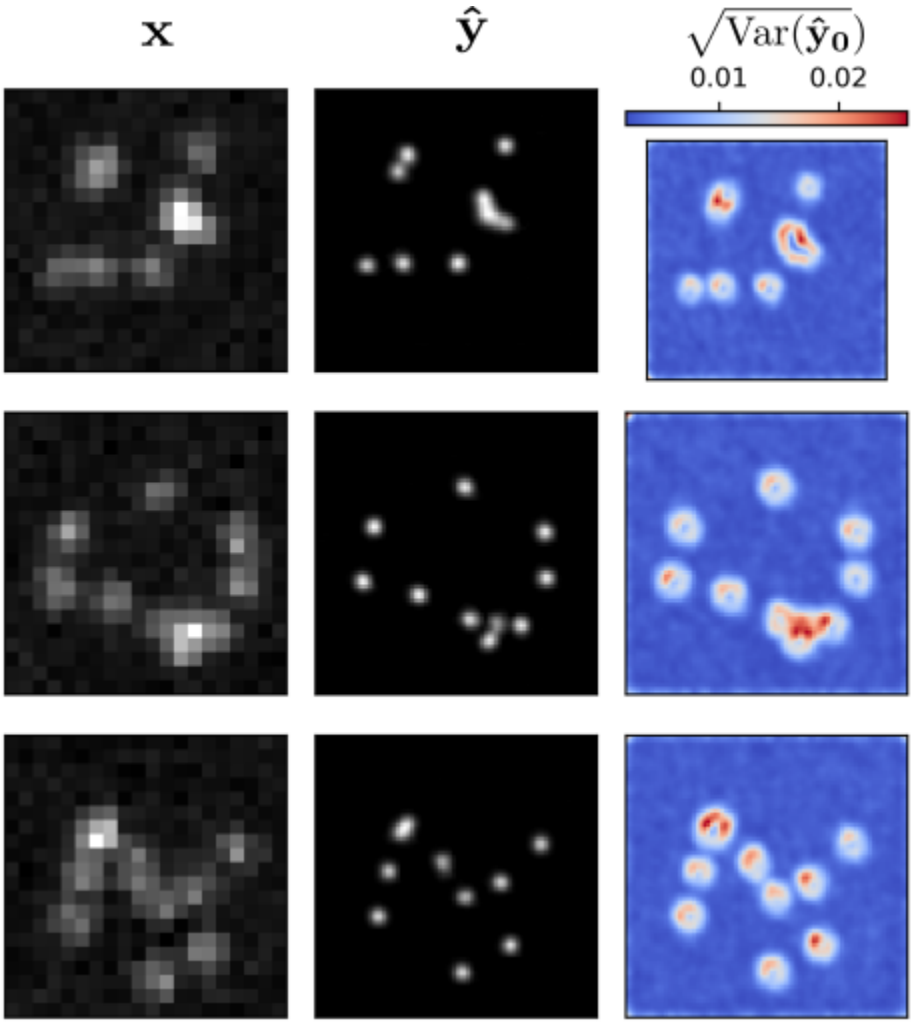

Figure 3: Kernel density estimates for various signal to noise ratios (SNR)

problem. That is, by placing a small Gaussian around each GT position (e.g. with std of 1 pixel), and matching continuous heatmaps, backpropagation yields more meaningful gradients and eases the learning process convergence.

Localization heatmaps thus form a natural encoding for SMLM images, which can be input to our conditional diffusion model. Therefore, to encode raw data $\mathbf{x}$ into a more tractable representation, we train the DeepSTORM architecture (Nehme 2020). Raw coordinates θ are binned into an upsampled image $\mathbf{z}$.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = ||\mathbf{y} - \hat{\mathbf{y}}||^2$$

5 Experiments

Training data was simulated under the image degradation model, drawing coordinates uniformly over a disc. We set $T = 100$ for all experiments and treat forward process variances β_t as hyperparameters, with a linear schedule from $\beta_0 = 10^{-4}$ to $\beta_T = 10^{-2}$. These constants were chosen to be small relative to data scaled to $[-1, 1]$, ensuring that reverse and forward processes have approximately the same functional form while keeping the signal-to-noise ratio at x_T as small as possible ($L_T = D_{KL}(q(x_T|x_0)\|\mathcal{N}(0, I)) \approx 10^{-5}$ bits per dimension in our experiments).

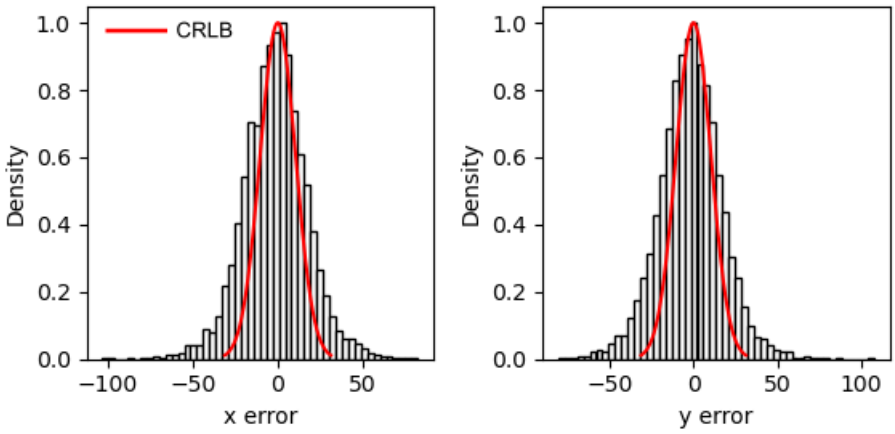

Figure 4: Localization errors of the trained model

To represent the reverse process, we used the DDPM architecture based on a U-Net backbone (Ho 2020). Parameters are shared across time, which is specified to the network using the Transformer sinusoidal position embedding **?**. We use self-attention at the 16×16 feature map resolution **??**. Details are in Appendix A.

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.

6 Related Work

6.1 Diffusion Models

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.

7 Conclusion

References

[1] Nehme, E., et al. *DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning*. Nature Methods 17, 734–740 (2020).

[2] Ouyang, W., et al. *Deep learning massively accelerates super-resolution localization microscopy*. Nature Biotechnology 36, 460–468 (2018).

[3] Speiser, A., et al. *Deep learning enables fast and dense single-molecule localization with high accuracy*. Nature Methods 18, 1082–1090 (2021).

[4] Sohl-Dickstein J., et al. *Deep unsupervised learning using nonequilibrium thermodynamics*. ICLR (2015).

[5] Ho J., et al. *Denoising Diffusion Probabilistic Models*. Advances in Neural Information Processing Systems (2015).

[6] Nanxin C., et al. *WaveGrad: Estimating Gradients for Waveform Generation* . ICLR (2021).

[4] Chao, J., et al. *Fisher information theory for parameter estimation in single molecule microscopy: tutorial*. Journal of the Optical Society of America A 33, B36 (2016).

[5] Schermelleh, L. et al. *Super-resolution microscopy demystified*. Nature Cell Biology vol. 21 72–84 (2019).

[6] Zhang, B., et al. *Gaussian approximations of fluorescence microscope point-spread function models*. (2007).

[7] Smith, C.S., *Fast, single-molecule localization that achieves theoretically minimum uncertainty*. Nature Methods 7, 373–375 (2010).

[8] Nieuwenhuizen, R., et al. *Measuring image resolution in optical nanoscopy*. Nature Methods 10. 557-562 (2013).

[9] Huang, F., et al. *Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms*. Nat Methods 10, 653–658 (2013).

[10] Rust, M., et al. *Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)*. Nat Methods 3, 793–796 (2006).

[11] Betzig, E., et al. *Imaging intracellular fluorescent proteins at nanometer resolution. Science* 313, 1642–1645 (2006).

[12] Weigert, M., et al. *Content-aware image restoration: pushing the limits of fluorescence microscopy*. Nat. Methods 15, 1090 (2018).

[13] Falk, T., et al. *U-net: deep learning for cell counting, detection, and morphometry*. Nat. Methods 16, 67–70 (2019).

[14] Boyd, N., et al. *DeepLoco: fast 3D localization microscopy using neural networks*. Preprint at bioRxiv https://doi.org/10.1101/267096 (2018)

[15] Zelger, P., et al. *Three-dimensional localization microscopy using deep learning*. Opt. Express 26, 33166–33179 (2018)

[16] Zhang, P., et al. *Analyzing complex single-molecule emission patterns with deep learning*. Nat. methods 15, 913 (2018)

[17] Saharia, C., et al. *Image Super-Resolution via Iterative Refinement*. Preprint at arXiv https://doi.org/10.48550/arXiv.2104.07636 (2021)

A Appendix

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. Recently, deep models have generalized SMLM to densely labeled structures by predicting high-resolution kernel density estimates (KDEs) from low resolution images with convolutional networks. However, estimated KDEs may contain irregularities due to finite sample sizes and limited model capacity. Denoising diffusion probabilstic models (DDPMs) are well suited conditional super resolution tasks, demonstrating promising results in detail reconstruction, while directly providing uncertainties in model predictions.

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 molecular photoswitching from dark to fluorescent states, permitting resolution of flurophores beyond the diffraction limit. Estimation of molecular coordinates is typically carried out by modeling the optical impulse response of the imaging system and fitting model functions to the data. However, such models are only well-suited to isolated molecules, reducing the number of molecules in the field of view and limiting temporal resolution in super resolution microscopy. This issue has incited a series of efforts to increase the density of fluorescent molecules imaged in a single frame while developing appropriate models for dense localization.

In fluorescence microscopy, each pixel is treated as a Poisson random variable (Smith 2010; Nehme 2020; Chao 2016), with expected value

$$\omega = i_0 \int O(u)du \int O(v)dv \tag{10}$$

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

$$\int O(u)du = \frac{1}{2}\left(\operatorname{erf}\left(\frac{u_k + \frac{1}{2} - u_0}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{u_k - \frac{1}{2} - u_0}{\sqrt{2}\sigma}\right)\right) \tag{11}$$

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 incorporates a normally distributed white noise per pixel ζ with offset o and variance σ^2. Ultimately, we have a Poisson component of the signal, which scales with N_0 and a Gaussian component, which does not.

Consider,

$$\zeta_k - o_k + \sigma_k^2 \sim \mathcal{N}(\sigma_k^2, \sigma_k^2) \approx \operatorname{Poisson}(\sigma_k^2) \tag{12}$$

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