
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 of
3 living cells at diffraction-unlimited precision. Classical SMLM approaches lever-
4 age the deactivation of fluorescent tags, followed by spontaneous or photoinduced
5 reactivation, which can be used to estimate of the density of a tagged biomolecule
6 in cellular compartments. Standard SMLM localization algorithms based on max-
7 imum likelihood estimators or least squares optimization require tight control
8 of activation and reactivation to maintain sparse emitters, presenting a tradeoff
9 between imaging speed and labeling density. Recently, deep models have gener-
10 alized SMLM to densely labeled structures by predicting high-resolution kernel
11 density estimates (KDEs) from low resolution images with convolutional networks.
12 However, estimated KDEs may contain irregularities due to finite sample sizes
13 and limited model capacity. Denoising diffusion probabilistic models (DDPMs) are
14 well suited conditional super resolution tasks, demonstrating promising results in
15 detail reconstruction, while directly providing uncertainties in model predictions.
16 Here, we adapt DDPM to the task of single molecule localization, and demonstrate
17 that combining traditional CNNs with a DDPM permits uncertainty quantification
18 of KDEs and improves localization precision over a wide range of experimental
19 conditions.

20 1 Introduction

21 Single molecule localization microscopy (SMLM) relies on the temporal resolution of fluorophores
22 whose spatially overlapping point spread functions would otherwise render them unresolvable
23 at the detector. Common strategies for the temporal separation of molecules involve molecular
24 photoswitching from dark to fluorescent. Estimation of molecular coordinates is then carried out via
25 modeling the optical impulse response of the imaging system and fitting model functions to the data.
26 However, such models are only well-suited to isolated molecules, reducing the number of molecules
27 in the field of view and limiting temporal resolution in super resolution microscopy. This issue has
28 incited a series of efforts to increase the density of fluorescent molecules imaged in a single frame
29 while developing appropriate models for dense localization.

30 Previous approaches to this issue has been to predict super-resolution images from a sparse set of
31 localizations with conditional generative adversarial networks (Ouyang 2018) or direct prediction
32 of molecular coordinates using neural networks (Nehme 2020; Speiser 2021). However, diffusion
33 models are an appealing alternative because they model a distribution of high-resolution images that
34 are compatible with a measurement. Although conditional VAEs and conditional GANs can provide
35 a distribution of images with enhanced resolution, both are known to suffer from mode collapse and
36 produce insufficient diversity in their outputs. Diffusion models are a recently developed alternative

37 to VAEs and GANs that excel at producing diverse samples and have been successfully applied to
38 solve inverse problems.

39 Here we focus on the class of models which perform single molecule localization using neural
40 networks. In this approach, one estimates molecular coordinates by predicting kernel density
41 estimates (KDEs) y , which are latent in the raw data x , using a convolutional neural network. Such
42 methods currently perform at the state of the art for localization microscopy, but few models report
43 uncertainty in predictions. Low resolution images provide information regarding fluorescent emitter
44 locations, yet information transfer is limited by fundamental physics and noise in the detection
45 process, resulting in ambiguity in true coordinates. Aleatoric uncertainties in SMLM are well
46 understood, and arise due to the stochastic nature of fluorescent emission from molecules and sensor
47 noise. Previous approaches to uncertainty quantification have thus far relied on additional CNNs to
48 predict uncertainty in coordinates from the data directly (Speiser 2021). However, this approach fails
49 to capture both aleatoric uncertainties in SMLM as well as epistemic uncertainties which can arise
50 due to the model itself. Therefore, we propose an alternative approach, which combines a deep KDE
51 estimator with a denoising diffusion probabilistic model (DDPM). The DDPM models a distribution
52 on KDEs, providing a novel mechanism for uncertainty quantification.

53 2 Background

54 2.1 Image Formation and Origins of Uncertainty

55 The central objective of single molecule localization microscopy is to infer a set of molecular
56 coordinates from measured low resolution images x . We begin by defining the likelihood on measured
57 low-resolution images $p(x|\theta)$. In fluorescence microscopy, each pixel is treated as a Poisson random
58 variable (Smith 2010; Nehme 2020; Chao 2016), with expected value

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

59 where $i_0 = \eta N_0 \Delta$. The scalar parameters η, Δ are the photon detection probability of the sensor and
60 the exposure time, respectively. Without loss of generality, we assume $\eta = \Delta = 1$. Most importantly,
61 N_0 represents the signal amplitude, which we assume maintains a fixed value. The optical impulse
62 response $O(u, v)$ is often approximated as a 2D isotropic Gaussian with standard deviation σ (Zhang
63 2007). This approximation has the convenient property, that the effects of pixelation can be expressed
64 in terms of error functions. For example, given a fluorescent emitter located at $\theta = (u_0, v_0)$, we have
65 that

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

66 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
67 incorporates a normally distributed white noise per pixel ζ with offset o and variance σ^2 . Ultimately,
68 we have a Poisson component of the signal, which scales with N_0 and a Gaussian component, which
69 does not. Therefore, in a single exposure, we measure:

$$x = s + \zeta \quad (3)$$

70 The distribution of x is the convolution of the distributions of s and ζ ,

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

71 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,
72 (4) is difficult to work with, so we look for an approximation. We will use a Poisson-Normal
73 approximation for simplification. Consider,

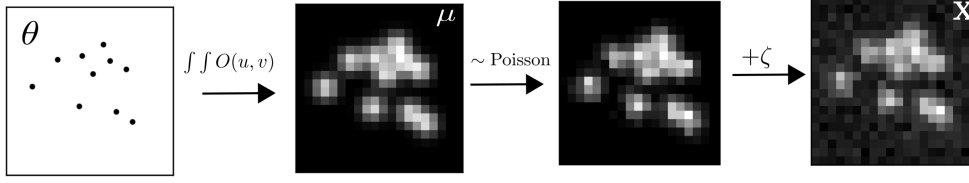


Figure 1: Generative model of single molecule localization microscopy images

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

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 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 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 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 $p(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y})$.

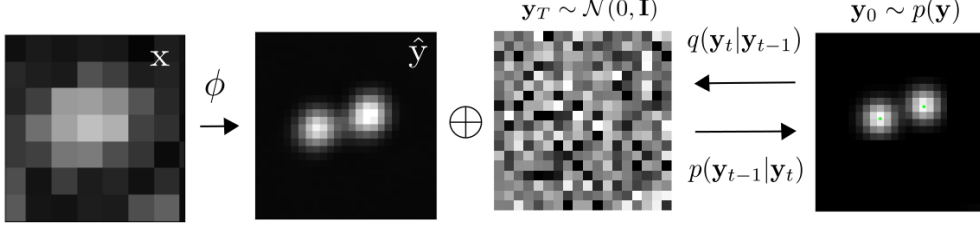


Figure 2: Conditional diffusion model for sampling kernel density estimates

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 $\beta_{0:T}$

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

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

$$q(\mathbf{y}_t|\mathbf{y}_0) = \mathcal{N}(\sqrt{\gamma_t}\mathbf{y}_0, (1 - \gamma_t)\mathbf{I}) \quad (9)$$

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 $p_{\theta}(\mathbf{y}_T) = \mathcal{N}(0, \mathbf{I})$.

$$p_{\theta}(\mathbf{y}_{0:T}) = p_{\theta}(\mathbf{y}_T) \prod_{t=1}^T p_{\theta}(\mathbf{y}_{t-1}|\mathbf{y}_t) \quad p_{\theta}(\mathbf{y}_{t-1}|\mathbf{y}_t) = \mathcal{N}(\mu_{\theta}(\mathbf{y}_t), \beta_t \mathbf{I}) \quad (10)$$

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

$$\mu_{\theta}(\mathbf{y}_t, \hat{\mathbf{y}}, \gamma_t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{y}_t - \frac{(1 - \alpha_t)}{\sqrt{1 - \gamma_t}} f_{\theta}(\mathbf{y}, \hat{\mathbf{y}}, \gamma_t) \right) \quad (11)$$

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

127 4.2 Optimization of the Denoising Model

128 To reverse the diffusion process, we optimize a neural denoising model f_θ that takes as input $\hat{\mathbf{y}}$ and a
 129 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
 130 distribution of noisy images at a time step t of the forward diffusion process.

$$\mathbf{y}_t = \sqrt{\gamma}\mathbf{y}_0 + \sqrt{1-\gamma}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (12)$$

131 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
 132 the sufficient statistics for the variance of the noise γ , and is trained to predict the noise vector ϵ .
 133 We make the denoising model aware of the level of noise through conditioning on a scalar γ . The
 134 proposed objective function for training f_θ is

$$\mathbb{E}_{(\hat{\mathbf{y}}, \mathbf{y}_0)} \mathbb{E}_{(\epsilon, \gamma)} \left[f_\theta \left(x, \sqrt{\gamma}\mathbf{y}_0 + \sqrt{1-\gamma}\epsilon \mid \mathbf{y}_t, \gamma \right) - \epsilon \right], \quad (13)$$

135 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
 136 impact on the quality of the model and the generated outputs. For our training noise schedule, we
 137 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
 138 training, we first uniformly sample a time step $t \sim \{0, \dots, T\}$ followed by sampling $\gamma \sim U(\gamma_{t-1}, \gamma_t)$.
 139 We set $T = 100$ in all our experiments.

140 4.3 Optimization of the DeepSTORM architecture

141 A first pass at localization treats localization as a binary classification problem, such that 0 denotes
 142 a vacant pixel and 1 denotes an occupied pixel containing an emitter. Direct learning of pixel-wise
 143 classification with cross-entropy loss leads to an imbalance of occupied and unoccupied pixels in
 144 dense localization problems (Nehme 2020). CE loss is usually either weighted [51], replaced with a
 145 Focal loss [52], or applied to a "blobbed" version of the desired boolean volume e.g. by placing a disk
 146 around each GT position [53–55]. Alternative methods take a soft version of the binary classification
 147 problem. That is, by placing a small Gaussian around each GT position (e.g. with std of 1 pixel),
 148 and matching continuous heatmaps, backpropagation yields more meaningful gradients and eases the
 149 learning process convergence.

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

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

154 5 Experiments

155 Training data was simulated under the image degradation model, drawing coordinates uniformly over
 156 a disc. We set $T = 100$ for all experiments and treat forward process variances β_t as hyperparameters,
 157 with a linear schedule from $\beta_0 = 10^{-4}$ to $\beta_T = 10^{-2}$. These constants were chosen to be small
 158 relative to data scaled to $[-1, 1]$, ensuring that reverse and forward processes have approximately
 159 the same functional form while keeping the signal-to-noise ratio at x_T as small as possible ($L_T =$
 160 $D_{KL}(q(x_T|x_0)\|\mathcal{N}(0, I)) \approx 10^{-5}$ bits per dimension in our experiments).

161 To represent the reverse process, we used the DDPM architecture based on a U-Net backbone (Ho
 162 2020). Parameters are shared across time, which is specified to the network using the Transformer
 163 sinusoidal position embedding ?. We use self-attention at the 16×16 feature map resolution ??.
 164 Details are in Appendix A.

165 and the channel multipliers at different resolutions (see Appendix A for details). To condition the
 166 model on the input x , we up-sample the low-resolution image to the target resolution using bicubic
 167 interpolation. The result is concatenated with y_t along the channel dimension. We experimented with

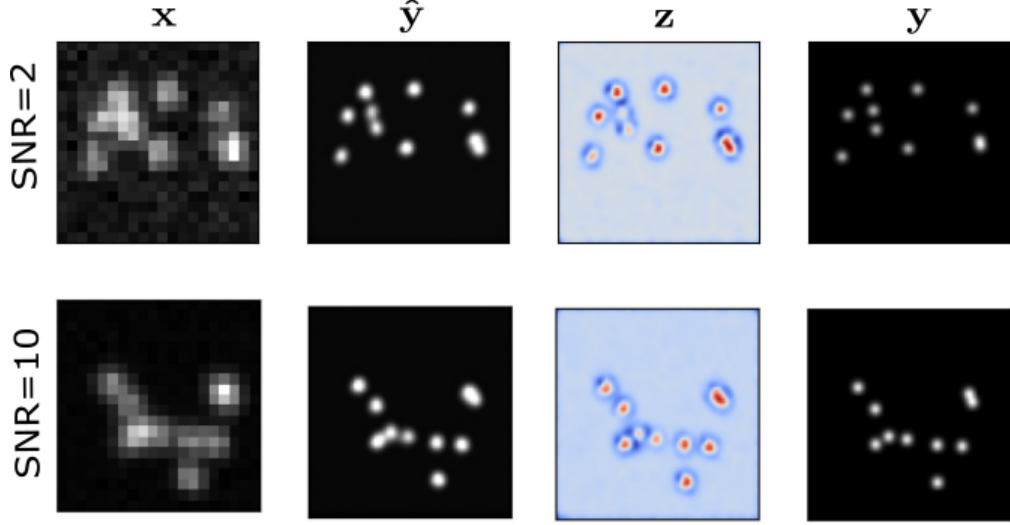


Figure 3: Kernel density estimates for various signal to noise ratios (SNR)

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

7.1 Fisher Information Metric

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)

$$\mathcal{I}_{ij}(\theta) = \mathbb{E}_{\theta} \left(\frac{\partial \ell}{\partial \theta_i} \frac{\partial \ell}{\partial \theta_j} \right) = \sum_k \frac{1}{\omega'_k} \frac{\partial \omega'_k}{\partial \theta_i} \frac{\partial \omega'_k}{\partial \theta_j} \quad (14)$$

References

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

192 [2] Ouyang, W., et al. *Deep learning massively accelerates super-resolution localization microscopy*. Nature
193 Biotechnology 36, 460–468 (2018).

194 [3] Speiser, A., et al. *Deep learning enables fast and dense single-molecule localization with high accuracy*.
195 Nature Methods 18, 1082–1090 (2021).

196 [4] Sohl-Dickstein J., et al. *Deep unsupervised learning using nonequilibrium thermodynamics*. ICLR (2015).

197 [5] Ho J., et al. *Denoising Diffusion Probabilistic Models*. Advances in Neural Information Processing Systems
198 (2015).

199 [6] Nanxin C., et al. *WaveGrad: Estimating Gradients for Waveform Generation*. ICLR (2021).

200 [4] Chao, J., et al. *Fisher information theory for parameter estimation in single molecule microscopy: tutorial*.
201 Journal of the Optical Society of America A 33, B36 (2016).

202 [5] Schermelleh, L. et al. *Super-resolution microscopy demystified*. Nature Cell Biology vol. 21 72–84 (2019).

203 [6] Zhang, B., et al. *Gaussian approximations of fluorescence microscope point-spread function models*. (2007).

204 [7] Smith, C.S., *Fast, single-molecule localization that achieves theoretically minimum uncertainty*. Nature
205 Methods 7, 373–375 (2010).

206 [8] Nieuwenhuizen, R., et al. *Measuring image resolution in optical nanoscopy*. Nature Methods 10. 557-562
207 (2013).

208 [9] Huang, F., et al. *Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms*.
209 Nat Methods 10, 653–658 (2013).

210 [10] Rust, M., et al. *Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)*.
211 Nat Methods 3, 793–796 (2006).

212 [11] Betzig, E., et al. *Imaging intracellular fluorescent proteins at nanometer resolution*. Science 313, 1642–1645
213 (2006).

214 [12] Weigert, M., et al. *Content-aware image restoration: pushing the limits of fluorescence microscopy*. Nat.
215 Methods 15, 1090 (2018).

216 [13] Falk, T., et al. *U-net: deep learning for cell counting, detection, and morphometry*. Nat. Methods 16, 67–70
217 (2019).

218 [14] Boyd, N., et al. *DeepLoco: fast 3D localization microscopy using neural networks*. Preprint at bioRxiv
219 <https://doi.org/10.1101/267096> (2018)

220 [15] Zelger, P., et al. *Three-dimensional localization microscopy using deep learning*. Opt. Express 26,
221 33166–33179 (2018)

222 [16] Zhang, P., et al. *Analyzing complex single-molecule emission patterns with deep learning*. Nat. methods 15,
223 913 (2018)

224 [17] Saharia, C., et al. *Image Super-Resolution via Iterative Refinement*. Preprint at arXiv
225 <https://doi.org/10.48550/arXiv.2104.07636> (2021)