Physics-Informed Generative Model for 3D Localization Microscopy

Ofri Goldenberg[1], Tal Daniel[2], Dafei Xiao[3], Yael Shalev Ezra[1], Yoav Shechtman[1,4]

Abstract—Localization microscopy techniques have overcome the diffraction limit, enabling nanoscale biological imaging by precisely determining the positions of individual emitters. However, the performance of deep learning methods commonly applied to these tasks often depends significantly on the quality of training data, typically generated through simulation. Creating simulations that perfectly replicate experimental conditions remains challenging, resulting in a persistent simulation-to-experiment (sim2exp) gap. To bridge this gap, we propose a physics-informed generative model leveraging self-supervised learning directly on experimental data. Our model extends the Deep Latent Particles (DLP) framework by incorporating a physical Point Spread Function (PSF) model into the decoder, enabling it to disentangle learned realistic environments from precise emitter properties. Trained directly on unlabeled experimental images, our model intrinsically captures realistic background, noise patterns, and emitter characteristics. The decoder thus acts as a high-fidelity generator, producing fully labeled, realistic training images with known emitter locations. Using these generated datasets significantly improves the performance of supervised localization algorithms, particularly in challenging scenarios such as complex backgrounds and low signal-to-noise ratios. Our results demonstrate substantial improvements in localization accuracy and emitter detection, underscoring the practical benefit of our approach for real-world microscopy applications. We will make our code publicly available.

Index Terms—Computational Microscopy, Single Molecule Localization, Deep Learning, Generative Models, Self-Supervised Learning, Object-centric Representation

1 Introduction

Determining the precise positions of single molecules or particles in optical microscopy images is a fundamental capability enabling critical insights at the nanoscale, with applications ranging from biology to materials science. Single-molecule localization microscopy (SMLM) techniques, including PALM [1], STORM [2], and PAINT [3], use sequential localization of emitters to reconstruct images that overcome the diffraction limit, allowing detailed biological visualization [4]. Single-Particle Tracking (SPT) utilizes localization over time to determine the trajectories of moving molecules or nanoparticles, revealing information about their dynamics, interactions, and microenvironment [5]. Extending localization microscopy to three dimensions (3D) typically requires physical modification of the microscope; a common approach involves Point Spread Function (PSF) engineering, where optical elements modify the PSF to encode depth [6], [7], [8], [9] or other information, such as spectral properties [10], [11] or molecular orientation [12], [13], [14].

Deep learning (DL) methods have become essential tools in microscopy and SMLM, enabling powerful analysis and reconstruction [15]. They have been applied to various challenges, including the localization of dense emitters [16], [17], the classification of emitter color [18], the determination of molecular orientation [19], [20], [21], and the facilitation of end-to-end optimization of imaging systems, e.g. designing the PSF together with the reconstruction algorithm [17], [18], [22], [23]. Importantly, the performance of many DL approaches, particularly supervised methods such as DeepSTORM3D [17], depends on the quality and quantity of the training data; such training data is often generated via simulation, which relies on modeling the PSF and various other experimental factors, because acquiring large, accurately annotated experimental datasets can be prohibitively time consuming and resource intensive.

However, building simulations that perfectly mimic experimental conditions is challenging [24]. Accurately capturing complex, spatially varying background, realistic noise statistics (which may go beyond simple Gaussian or Poisson models), system-specific aberrations, and subtle experimental factors—such as intricate biological structures, complex dynamics in live cells or particle flows, or other factors difficult for humans to discern or model— requires significant efforts and in many cases is impractical. This leads to a persistent simulation-to-experiment (sim2exp) gap, where models trained on synthetic data underperform when applied to real measurements. Furthermore, even when an adequate simulation model exists, modifying it to reflect new experimental conditions can require time-consuming manual tuning or substantial redevelopment efforts.

In this work, our aim is to bridge this sim2exp gap and propose a physics-informed generative approach that takes advantage of self-supervised learning on experimental data. We develop an object-centric self-supervised generative model built on the Deep Latent Particles (DLP [25], [26]) framework. DLP is a variational autoencoder (VAE [27]) that encodes images as a set of latent particles, compared

[1]*Department of Biomedical Engineering, Technion–Israel Institute of Technology, Haifa, Israel.*
[2]*Carnegie Mellon University, Pittsburgh, PA, USA.*
[3]*Russell Berrie Nanotechnology Institute, Technion–Israel Institute of Technology, Haifa, Israel*
[4]*Department of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Haifa, Israel*
Corresponding author: Yoav Shechtman, `yoavsh@technion.ac.il`

to the standard single-vector holistic representation, where each particle is composed of several learned attributes such as position (2D keypoint), scale, transparency, and latent visual features. The model is trained solely to reconstruct the original input image by stitching the decoded latent particles together. We propose PILPEL - physics informed latent particles for emitter localiztion, where we develop a DLP-based decoder to incorporate an explicit physical PSF model that takes the encoded spatial coordinates derived from the latent space as input. This allows the model to separate the realistic environment learned from the precise placement and properties of the emitters, which are also learned as part of the training process. We train the model directly on unlabeled experimental localization microscopy images to intrinsically learn the realistic background and noise patterns, and implicitly capture variations in emitter density and intensity within its learned representations. In this work, we show results with variations of the Tetrapod mask (Figure 1a), but the flexibility of the model makes it applicable to any PSF (e.g. astigmatism [9], double-helix [7], etc.), by inserting a different phase mask and accompanied optical parameters to the PSF model. Figure 1b shows an example application of 3D localization: several yeast cells with fluorescently labeled DNA loci are apparent in the field of view. the goal here would be to determine the 3D positions of all loci.

Trained on measurements, the PILPEL model serves as a realistic data generator, where the decoder functions as high fidelity generator, producing realistic images with learned background features, image structure, and noise characteristics, as well as known ground truth emitter locations and associated intensities, reflecting the learned emitter-density and brightness distributions. This generated data, embodying experimentally learned characteristics, serves as highly realistic training material for supervised localization algorithms, enhancing their performance on real-world data, enabling their use across diverse experimental conditions, and improving their ease of application.

Our contributions are summarized as follows: (1) we propose a physics-informed Deep Latent Particle model, a novel self-supervised object-centric generative image representation model that builds on DLP and is used to generate realistic, labeled training data embodying experimentally learned characteristics that can be used to train supervised localization models such as DeepSTORM3D; (2) we demonstrate that the use of the generated data leads to superior performance in downstream localization tasks, showing significant improvements in localization accuracy and detection rates, particularly on challenging test data with complex backgrounds and low signal-to-noise ratios, and (3) we highlight the advantage of automatically learning data characteristics directly from measurements, bridging the simulation-to-experiment gap without manual tuning of simulation parameters.

2 Related work

Deep learning for microscopy. Recent large vision foundation models, including adaptations for microscopy [28], [29], [30], show promise for tasks such as segmentation or mapping protein organization or classification, but their direct application to generating high-precision localization microscopy data is challenging. These models are often pretrained on different data domains (e.g., natural images or other microscopy modalities) and are typically designed for tasks like segmentation rather than sub-pixel localization. Additionally, fine-tuning or training such large models specifically for localization microscopy is hindered by the lack of extensive labeled datasets, particularly for diverse experimental conditions and SNR, or involving engineered PSFs where computationally designed optics alter image characteristics. VAE models have been shown to have the ability to encode emitters' z-position in their latent space [31]. Our work contributes to this area by introducing a physics-informed generative VAE-based object-centric latent model trained on experimental data, designed to generate realistic, label-consistent localization-microscopy images.

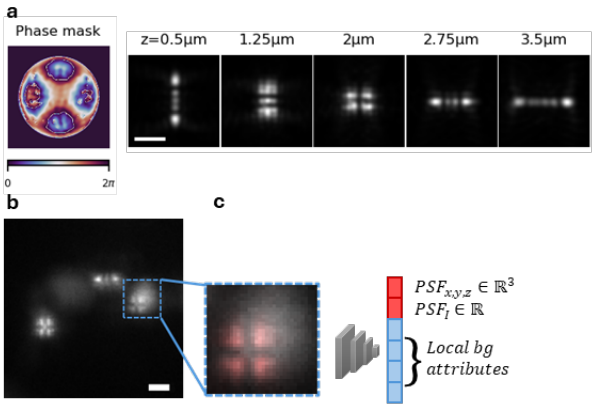


Fig. 1. Deep latent representation of PSF-objects. a. A Tetrapod phase mask incorporated in the microscope encodes the emitter's axial position in the shape of the PSF. b. Experimentally measured yeast cells with fluorescently labeled DNA loci, imaged in a 4f system using the Tetrapod phase mask. The DNA loci are located within the 3D volume of the nucleus; hence the different shapes of the PSF encoding 3D locations of emitters (one in each cell). The blow-up shows a single locus (in red) as well as background fluorescence from the yeast cell. c. The model learns to represent an image as a set of objects, each encoding a PSF component, characterized by 3D coordinates and intensity, as well as local background attributes. Scale bar, 2 μm.

Generative models for SMLM. In the domain of data synthesis, the challenge of generating realistic training data for deep learning in localization microscopy remains critical due to the difficulty in accurately simulating complex experimental conditions. Several generative modeling strategies have been explored to tackle this. Diffusion models, for instance, were adapted by Saguy et al. [32] to synthesize realistic super-resolution microscopy images from experimental data, demonstrating utility for data augmentation in downstream tasks like single-image super-resolution. However, the generated data is typically unlabeled, and thus can not be used for training localization models. Other generative approaches, such as image-to-image translation with Generative Adversarial Networks (GANs) [33], have also been applied in microscopy, often aiming to bridge the sim2exp gap. VAE models Methods like conditional GANs (e.g., pix2pix [34]) attempt to learn direct mappings between simulated and real domains or generate images from labels.

However, effectively applying these generative methods to produce useful training data for localization microscopy faces significant difficulties. The need for sub-pixel accuracy for localization is a demanding requirement for standard generative frameworks, and ensuring the preservation of precise ground-truth localization labels during image translation or generation can be challenging and potentially introduce artifacts.

3 Background

DeepSTORM3D DeepSTORM3D (DS3D [17]) is a deep learning-based method designed for dense 3D single-molecule localization microscopy (SMLM) over large axial ranges. It utilizes a convolutional neural network (CNN) to analyze 2D images containing overlapping, engineered point-spread functions (PSFs), such as the Tetrapod PSF, and outputs a 3D localization map. The supervised network is trained entirely on simulated data, where large datasets are generated using a physics-based model of the microscope and PSF. These simulations incorporate realistic experimental variability, including different emitter densities, signal-to-noise ratios, background levels, and noise models. A significant step in the DeepSTORM3D analysis pipeline is the generation of training data that resembles the experiment. This involves significant manual tuning and can be highly time demanding even for an experienced user. In this work, among other contributions, we alleviate this manual tuning completely.

Deep Latent Particles (DLP) Deep Latent Particles (DLP [25], [26])[1] is a self-supervised, object-centric image representation model that decomposes an input image $x \in \mathbb{R}^{H \times W \times C}$ into a set of K structured foreground latent particles $L = \{l_p, l_s, l_d, l_t, l_f\}_{k=0}^{K-1}$. Each particle encodes the following attributes: $l_p \in \mathbb{R}^2$ denotes the 2D position (keypoint), $l_s \in \mathbb{R}^2$ the spatial scale (bounding box size), $l_d \in \mathbb{R}$ the appearance order, $l_t \in \mathbb{R}$ the transparency, and $l_f \in \mathbb{R}^m$ the latent visual features. A key component of DLP is the spatial prior over keypoints, where particle positions are determined using a spatial softmax (SSM) operator applied to activation maps from a patch-based convolutional network. In addition to the K foreground particles, the model includes a single background particle $l_{\text{bg}} \in \mathbb{R}^{m_{\text{bg}}}$ to represent the background visual features. The values of K, m, and m_{bg} are model hyperparameters. To learn these attributes, the model decodes the latent particles into glimpses and composes them together with the background to reconstruct the input image. Training is performed end-to-end using a variational autoencoder (VAE [27]) framework, with the objective of maximizing the evidence lower bound (ELBO), combining image reconstruction loss with a KL-divergence regularization term. The disentangled latent representations learned by DLP have been successfully applied to a variety of downstream tasks, including video prediction [26], reinforcement learning [35], and imitation learning for robotics [36]. For a detailed overview of the model, we refer the reader to the original work [26].

1. In our work, we use the improved DLPv2 [26] as our base model and refer to it simply as DLP.

4 Method

Our goal is to design a *controllable* generative model that can generate realistic image data for the task of SMLM. In the following, we describe our proposed model that extends the DLP framework to address specific challenges in localization microscopy.

4.1 Physics-Informed Deep Latent Particle Model for Localization Microscopy

In localization microscopy, the objects of interest are individual emitters, each represented by a PSF (Figure 1c). Accurate localization of these PSFs is essential for achieving nanoscale resolution in single-molecule localization microscopy (SMLM) reconstructions. This motivates the use of object-centric generative models, such as DLP, which offer fine-grained control through explicit attributes like spatial coordinates—attributes that are essential for precise localization.

The Deep Latent Particles (DLP) model provides a powerful object-centric representation by decomposing an image into a set of latent particles, each with explicit attributes such as 2D position, as described in Section 3. Given the challenges of collecting labeled data in real-world microscopy settings, we aim to adapt DLP for localization microscopy by enabling it to recover 3D positions and intensities of point-emitters, observed as individual point spread functions (PSFs). To this end, we extend the original DLP model to explicitly encode the 3D coordinates and intensity of each emitter, such that particles correspond to individual emitters, and reconstruct not only the PSFs but also, crucially, additional image components such as local and global background, signal-to-noise ratio (SNR), and effective emitter density.

This is achieved by incorporating a physical model of the PSF into the decoder, where the encoded $\{x, y, z\}$ coordinates are used as input to simulate PSFs as part of the image formation process. This PSF model serves as a strong inductive prior, embedding physical constraints and domain knowledge into the generative process. The resulting physics-informed architecture allows the model to disentangle the PSFs from other image structures, as well as from noise and measurement artifacts. It also encourages the latent space to explicitly capture the key parameters of interest–namely, emitter location and intensity—enabling fine-grained control and accurate reconstruction.

In our modified model, the physics-informed prior replaces the spatial softmax (SSM)-based prior used in the original DLP to propose keypoint locations. Specifically, we employ a physical model that captures depth-dependent PSF variations and accounts for the optical parameters used in the experiment. This PSF model can be theoretically derived or experimentally estimated, for example, by using phase retrieval methods such as VIPR [37]. Similarly to DeepSTORM3D [17], our model includes a differentiable optical layer that simulates image formation from emitter coordinates to the final sensor image. The same PSF model is used both in the prior module to propose emitter locations and in the decoder to generate the PSF component of each reconstructed patch, ensuring consistency between proposal and generation.

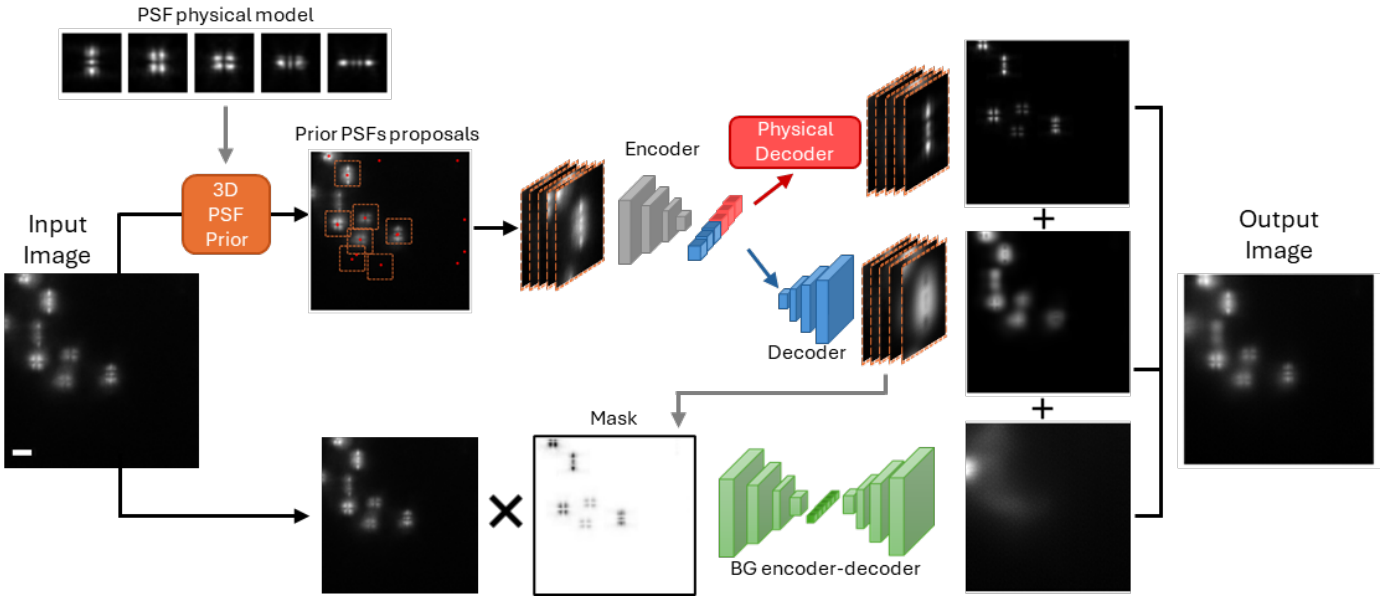


Fig. 2. Physics-informed encoding-decoding. The model begins by dividing the input image into overlapping patches and performing 3D template matching using a precomputed PSF dictionary to propose candidate emitter locations. Patches around these locations are then processed by an encoder that generates disentangled latent vectors encoding spatial offsets, intensity, and appearance features. A decoder reconstructs the image using a physics-based PSF generator and a CNN-based local background generator. These patches, along with a separately modeled global background, are combined to form the final image. The entire system is trained end-to-end within a VAE framework, balancing reconstruction accuracy and latent regularization. The output is a new generated image, with known GT emitters localizations, that can be used as a training sample for supervised localization networks. Scale bar, 2 μm.

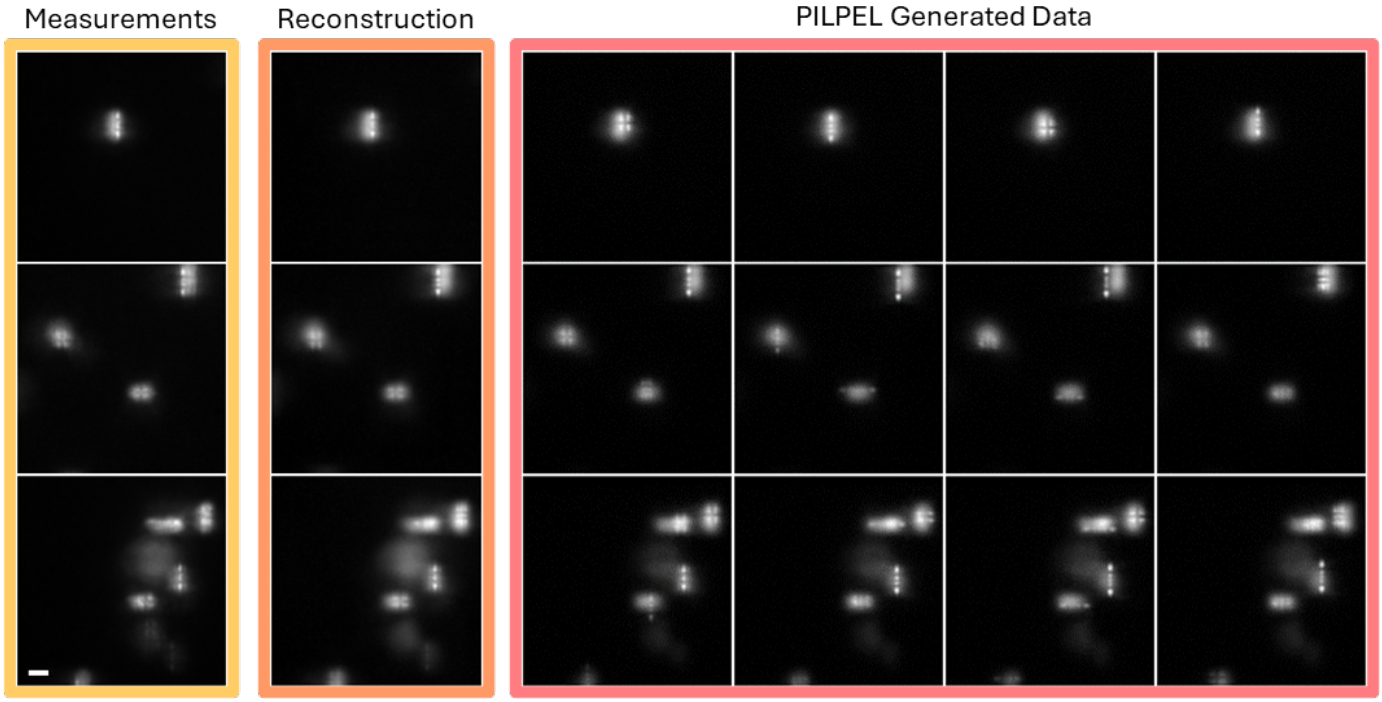


Fig. 3. Realistic labeled training data generation. Measured images are passed through the encoder-decoder model to generate the reconstructions. To create a labeled dataset, the encoded PSF attributes are randomly perturbed, and the stochastic latent variables are resampled. This process produces realistic image variants that resemble the original measurements. Since the PSF coordinates fed into the generator are known, the resulting dataset is fully labeled and can be used to train a 3D localization network. Scale bar, 2 μm.

4.1.1 Model Architecture

Our model architecture is illustrated in Figure 2. The input image is initially divided into overlapping patches. Each patch undergoes a 3D template matching process against a dictionary of pre-calculated PSFs spanning the expected axial z range. This yields an initial estimation of potential 3D PSF positions (l_x, l_y, l_z coordinates) within each patch. The maximum cross-correlation score from this matching

can be subsequently used as a confidence metric to filter out patches unlikely to contain any PSFs, resulting in a proposed set of latent particles.

Following this prior-driven proposal stage, we extract image patches centered around each proposed PSF location. These patches, intended to capture both the PSF and its local surroundings, serve as input to a particle encoder architecture. The encoder outputs the latent attributes for each proposed object. These are disentangled by construction, with specific latent attributes representing offsets (l_{dx}, l_{dy}, l_{dz}) to refine the initial position estimate, the PSF's intensity l_i, and additional latent attributes encoding the appearance features of the patch $(l_s, l_d, l_t, l_f$ like before). Learning a distinct intensity factor for each proposed PSF allows the model to capture experiment-specific SNR. Furthermore, this intensity factor can be driven towards zero during optimization, effectively removing initially proposed PSF locations that are not needed to reconstruct the image. This mechanism enables the model to adapt to varying emitter densities by starting with numerous proposals and automatically pruning the unnecessary ones.

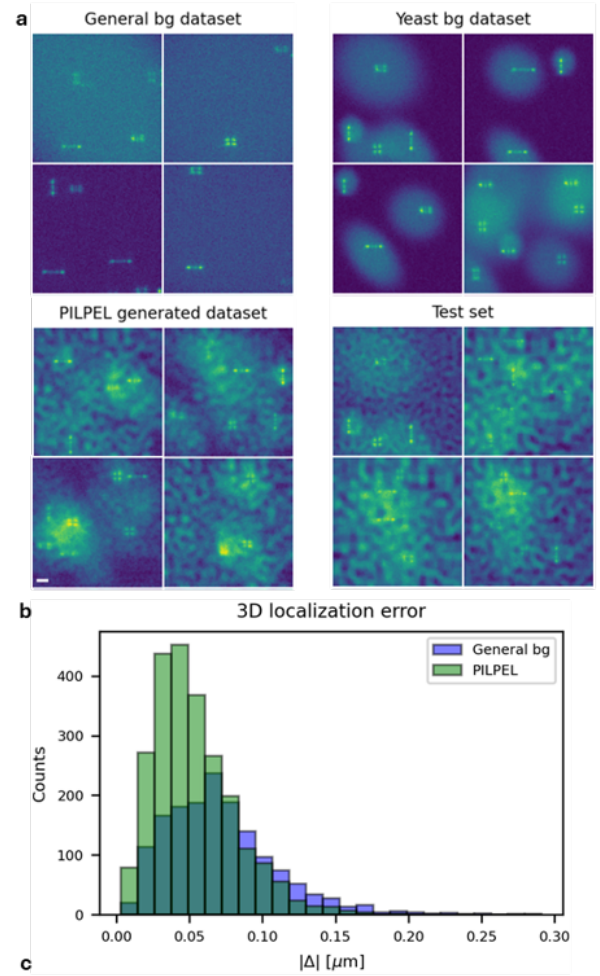
The decoder reconstructs the image from these latent representations and consists of two distinct components. The first is the differentiable physical layer that takes the refined spatial coordinates (initial estimate and the learned offset, $l_x + l_{dx}$, etc.) and the intensity factor from the latent vector l_i to generate a PSF image. The second component is a CNN-based decoder which takes the remaining latent appearance attributes and generates a corresponding image patch representing the local background surrounding the PSF. The output of the physical layer (the PSF) is combined with the output of the CNN decoder (local background) to form a complete patch containing both the emitter and its immediate environment. These reconstructed patches are then placed into the final reconstructed image at their corresponding global coordinates.

To learn the global background characteristics, the generated object patches act as masks on the original input image. This masked input is processed by a separate encoder-decoder network dedicated to modeling and reconstructing the general background component of the scene.

Finally, the generated components—the PSFs, the local background patches, and the general background—are added together to produce the final reconstructed image. The entire model is optimized end-to-end within the VAE framework by minimizing the ELBO - the reconstruction loss (typically MSE or a perceptual loss) between the final output image and the original input image, and KL-divergence regularization term for the stochastic latent variables.

4.1.2 Data Generation for Supervised Learning

The decoder of our trained DLP model serves as a high-fidelity generator of realistic labeled images. By sampling coordinates and latent variables from the latent space, the decoder can generate images with known ground-truth emitter locations and intensities, which are the input to the physical layer. Most importantly and key to our application: these generated, labeled images (in the sense that precise emitter positions are known) inherit the realistic



	PILPEL vs. General bg			PILPEL vs. Yeast bg		
Test set	ΔRMSE		Δ Jacc	ΔRMSE		Δ Jacc
Perlin	-19 [nm]	**-28 [%]**	**+0.47**	-5 [nm]	-9 [%]	+0.40
Perlin Low SNR	-21 [nm]	**-26 [%]**	**+0.21**	-12 [nm]	-16 [%]	+0.22

Fig. 4. Robustness experiment on simulations– DS3D localization performance. a. Comparison of synthetic training datasets used for DS3D: General BG – the standard background in DS3D; Yeast BG – includes local background around emitters to mimic yeast cells; Test set – simulates yeast cells with Perlin noise model; and Generated – produced via our model. Scale bar, 2 μm b. 3D localization error comparison: DS3D trained on the generated dataset vs. DS3D trained on the standard simulated dataset. Mean absolute error for the generated model $= 0.053\mu$m, compared to 0.071μm for the original. c+d. Improvement of training on generated data. Localization RMSE and Jaccard index for DS3D trained on the generated dataset vs. training on yeast-BG and original-DS3D datasets.

background and noise characteristics learned from the experimental data (Figure 3).

This capability allows for the generation of large and accurately labeled training datasets for supervised localization models, specifically DeepSTORM3D (DS3D [17]) in this work. Training on this generated data, which closely mirrors real measurements, significantly improves the performance and generalization of supervised localization algorithms, by reducing the sim2exp gap. Furthermore, this approach completely eliminates the need to manually tune simulation parameters for new experimental conditions or add new components to the synthetic data simulator, which can be hard, time-consuming, or possibly unknown based on observation.

Notably, the reconstruction of the encoding-decoding process results in images that appear similar to the input images, but the encoded PSF coordinates do not necessarily achieve the sub-pixel accuracy required for precise localization directly (the error between the encoded PSF coordinates of the input image and the GT coordinates in simulations is 300nm 3D distance RMSE, which is unacceptable for single molecule localization). Thus, we cannot use these encodings as the GT labels. However, the latent coordinates, which serve as input to the physical model in the decoder, do represent the ground-truth positions in the output image, by definition. In other words, we can now generate images that highly resemble the measurement, and contain emitters in precisely known positions, even if those positions are not exactly those of the experimentally measured emitters. We generate ground-truth data by sampling perturbations of the same input image, as shown in Fig 3. The global and local background are learned from the image, but the PSF coordinates are new and serve as the training labels. We use these images, with their precisely correct labels, namely emitter positions, to train a supervised model.

5 Experiments

In this section, we detail our experimental benchmark followed by a comparison with baselines, both on real and simulated data. Our aim is to provide evidence that the generation of realistic data can improve the downstream performance of localiaztion tasks. Our code for reproducing the results will be made publicly available.

5.1 Closing the Simulation-to-Reality Gap with Learned Data Generation

To demonstrate the advantage of using data generated from experimental measurements and to assess performance improvements, we evaluated DS3D for 3D localization across different training datasets. We created a test set resembling yeast cell measurements expressing point sources shaped by the Tetrapod phase mask (shown in Figure 1a), simulated by adding random asymmetrical 2D Gaussians around each PSF to model the typical autofluorescence observed in yeast cells. To mimic a realistic "measured" test set with different unknown noise patterns not included in the synthetic training data, we introduced noise patterns that were not incorporated into the original DS3D simulator and cannot be manually tuned. We used Perlin noise to simulate inhomogeneous backgrounds. This technique has been applied in microscopy image analysis, as demonstrated in [16], where it was used to model non-uniform illumination and other artifacts. Our implementation uses scale parameters ranging from 0.5 to 1.7 and intensity values from 5 to 14, which control the spatial frequency and amplitude of the noise, respectively, to produce diverse patterns. We term this test set `PerlinData`.

Datasets. We trained DS3D on three different datasets: (1) the original simulator with a wide range of SNR and background values; (2) the original simulator with an added local background (the same asymmetrical Gaussians used in the test set, manually introducing prior knowledge); and (3) a dataset generated by our model after training on the `PerlinData` test set.

Figure 4a illustrates these datasets, showing that, as expected, the generated data most closely resembles the measured test set. Consequently, the performance of DS3D when trained on this data is improved; Figure 4b shows the improved localization accuracy of DS3D trained on generated data compared to the original training set, as well as, importantly, a higher number of detected emitters. Figure 4c and 4d summarize the performance gains of using the generated dataset under challenging conditions with Perlin noise: localization RMSE improves by 19nm (a 28% improvement), and the Jaccard index (measuring the overlap between detected and ground truth localizations) doubles from 0.45 to 0.92. While introducing prior knowledge (e.g., yeast background) into synthetic training data improves performance, training on generated data is still significantly more effective, and eliminates the need for implementing additional physical attributes in the DS3D simulator.

5.2 Super-resolution 3D Imaging

To demonstrate the capability of our data generation approach on complex biological structures imaged via SMLM, we applied it to a super-resolution dataset of mitochondria. This dataset, originally presented in Nehme et al. (2020) [17], involves 3D STORM imaging of mitochondria in fixed COS7 cells. The mitochondria were fluorescently labeled with Alexa Fluor 647 and imaged using a 4f system equipped with a Tetrapod phase mask. As shown in Figure 5, the 3D super-resolved reconstruction appears comparable to the original, with improvements in several regions. Since the DS3D training data had been manually refined and carefully optimized to resemble the experimental data (Figure 5c), it already performs well on the measurements. However, our method further improves the reconstruction results - the localizations increased from 636K detections to 886K (39% increase), on 20K frames. This factor, namely, the number of localizations, is crucial in SMLM, and is translatable to improved image quality, i.e. fuller features (as seen in Figure 5b), or faster acquisition. Additionally, it works in an automatic manner, by learning directly from the measurements without the need for manual tuning of the DS3D simulation parameters, which requires hours if not days, depending on user expertise. Human capability to verify and check vast amounts of images is limited, whereas PILPEL learns to generate based on the entire dataset, providing more comprehensive optimization. Furthermore, manual

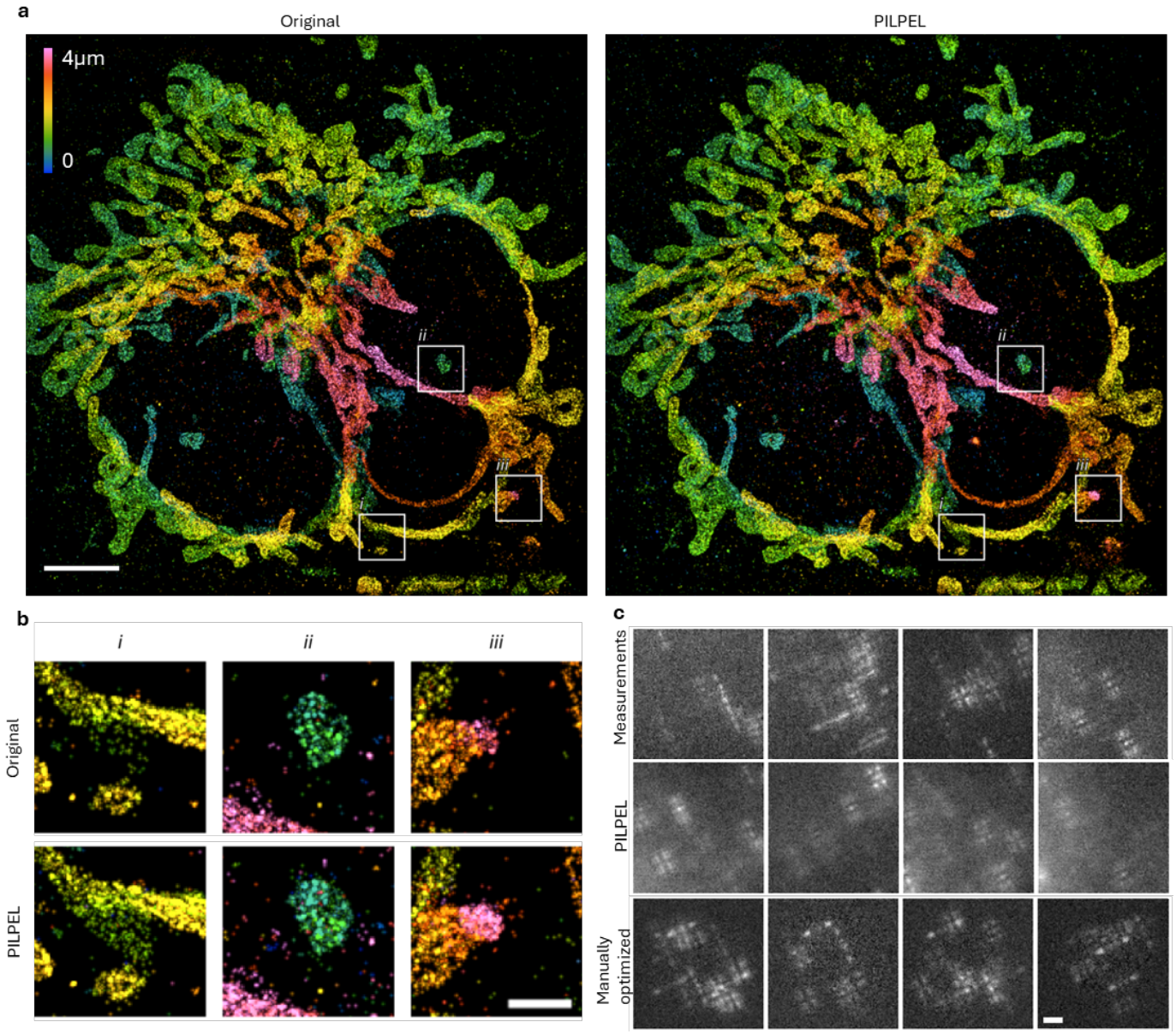


Fig. 5. Comparison of 3D super-resolution reconstruction. (a) Super-resolved reconstructions of mitochondria (depth ≈ 4 μm), rendered as 2D histograms with color-coded z-positions, using DS3D trained on the original simulated data (left) versus our generated data (right). Scale bar, 5 μm (b) Magnified views of regions i, ii, and iii as indicated in (a). Scale bar, 1 μm (c) Examples of randomly sampled training data images, generated by our model, and the original manually-optimized DS3D simulator. Scale bar, 2 μm

tuning relies on subjective visual assessment ("does it look similar enough?"), while PILPEL employs computationally optimized image similarity loss between measurements and generated data, which is more efficient and robust.

5.3 3D Localization of DNA Loci in Yeast Cells

We tested our model on experimental images of fluorescently labeled DNA loci in fixed *Saccharomyces cerevisiae* yeast cells. The cells were tagged with mCherry, generating red fluorescent spots at the POA1 DNA loci (chromosome II); each cell exhibits a single fluorescent spot [38]. In such a biological system, an example application would be to track loci over time in 3D under different conditions [5], [38], [39]; here, we use it to demonstrate our method under a different noise model than STORM. The imaging was performed using a 4f extension to the optical system (equipped with an oil immersion objective of 1.49 NA and a magnification of 100x), incorporating a Tetrapod phase mask designed for $\lambda = 605$nm with an axial range of 3μm. From the measurements, we cropped 193×193 pixels ROIs, which were input into our model along with a phase-retrieved PSF model generated using VIPR [37] from a calibration bead.

Datasets. Two datasets were collected from different fields of view and stages of the experiment: one with high SNR and relatively sparse cells, and another with low SNR and higher density, including naturally denser FOVs resulting from cell aggregation, acquired later in the experiment after some photobleaching. Figure 3 and Figure 6 show the measured data sets and their corresponding reconstructions, respectively. The dense, low-SNR dataset was used as a

challenging test set for DS3D localization, comparing a model trained with the original simulator versus a model trained on generated data derived via our data generation pipeline.

Results. Figure 6b displays the measured images alongside localization results from both networks. We report that the general detection rate increased by 1.5-fold, from $12,561$ detections with the original net, and $18,461$ detections using the PILPEL generated dataset, on 3 timelapse measurements of 100 timepoints each, on a 1200×1200 FOV. In addition to that, mean detection confidence values increased by 40% using PILPEL. Ground-truth positions are unavailable in such experiments, but state-of-the-art GT estimation in single cells is MLE localizations - fitting xyz coordinates of the PSF model together with asymmetric gaussian background. This is relevant when the model matches the measurement, and MLE isn't always correct in such difficult cases. Hence, we conducted a small experimental test set of $1,500$ single-cell measurements with relatively high SNR where we assume the MLE fitting will work, and compared the results to the two networks' localizations. Localization RMSE improves from 107 nm to 96 nm, a 10% improvement, using PILPEL. Importantly - this dataset comprised high SNR data, whereas a more significant improvement will be on low SNR cases, where the model deviates substantially from the measurement. The advantage of our method is that it generalizes the model and can identify more emitters, in a large FOV with multiple emitters. This is reflected in greater detection rate, seen also on the provided images, including low-SNR emitters missed by the original DS3D.

6 Limitations and Future Work

Despite the advances demonstrated, our method exhibits several limitations. Specifically, training the generative model requires additional computational time compared to training on synthetic datasets, although it significantly reduces manual human intervention. Additionally, accurate reconstruction still necessitates a reliable prior PSF model, which can be obtained from a numerical model or from experimental calibration followed by phase retrieval. Moreover, the effectiveness of our approach depends heavily on the reconstruction quality and its resemblance to actual measurements. If reconstruction quality is insufficient, adjusting parameters such as patch size or the number of latent particles may be required.

Future work may involve extending the method to dynamic scenarios through Deep Dynamic Latent Particles (DDLP [26]), for applications such as video-based molecule tracking and flow visualization in live cells. Another promising avenue is leveraging this generative capability to produce training data for complex-to-simulate microscopy scenarios, including particle flows, living organisms like C. elegans, and large-field-of-view imaging. Moreover, controlled manipulation of data properties within the latent space could facilitate the generation of datasets specifically tailored for downstream microscopy tasks, including denoising and super-resolution, which would be especially beneficial when high-quality training data are scarce or difficult to obtain. Additionally, we plan to explore further flexibility by incorporating model aberrations encoded and learned from the experiments [40], enabling also the optimization of the PSF model itself.

7 Conclusion

We introduced a physics-informed generative model that bridges a major simulation-to-experiment gap in localization microscopy. By incorporating an explicit PSF model in the self-supervised object-centric DLP model and training it on real experimental measurements, our approach generates high-fidelity, labeled training data. The main advantage in our method is experiment-learned-features for training data; the generalization ability of the method comes in effect on the noise patterns, general and local background, emitters density and intensity – basically everything other than the PSF model. This learning of unmodeled features is clearly demonstrated in the yeast cells experiments. While one can incorporate local background to the DS3D simulator, Fig. 4a shows it still doesn't adequately resemble actual measurements. PILPEL's generated data improves this correspondence and removes the need for manual implementation, allowing its use for future datasets with different noise patterns and shapes. PILPEL also successfully learned Perlin noise with various levels and patterns, autofluorescence without PSFs, and random background patterns. These elements are not modeled in the original DS3D simulator and contribute to the sim2exp gap that can significantly degrade localization network performance. As our experimental benchmark demonstrates, training supervised algorithms on this data significantly enhances localization accuracy and detection, especially for challenging or novel experimental conditions, outperforming methods trained on standard simulations. This eliminates the need for manual simulation tuning and offers a robust, adaptable solution for improving deep learning performance in localization microscopy.

Acknowledgments

We thank Prof. Johan Elf and Axel Truedson for helpful discussion. We also thank Dr Elisa Dultz and Prof. Karsten Weis for providing us with the yeast strains used in the yeast cells experiments. This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 450/18). OG was supported in part by VATAT Scholarship Program For PhD Students in Data Science.

References

[1] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging Intracellular Fluorescent Proteins at Nanometer Resolution," *Science*, vol. 313, no. 5793, pp. 1642–1645, Sep. 2006. [Online]. Available: https://www.science.org/doi/10.1126/science.1127344

[2] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nature Methods*, vol. 3, no. 10, pp. 793–796, Oct. 2006. [Online]. Available: https://www.nature.com/articles/nmeth929

[3] A. Sharonov and R. M. Hochstrasser, "Wide-field subdiffraction imaging by accumulated binding of diffusing probes," *Proceedings of the National Academy of Sciences*, vol. 103, no. 50, pp. 18 911–18 916, Dec. 2006. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.0609643104

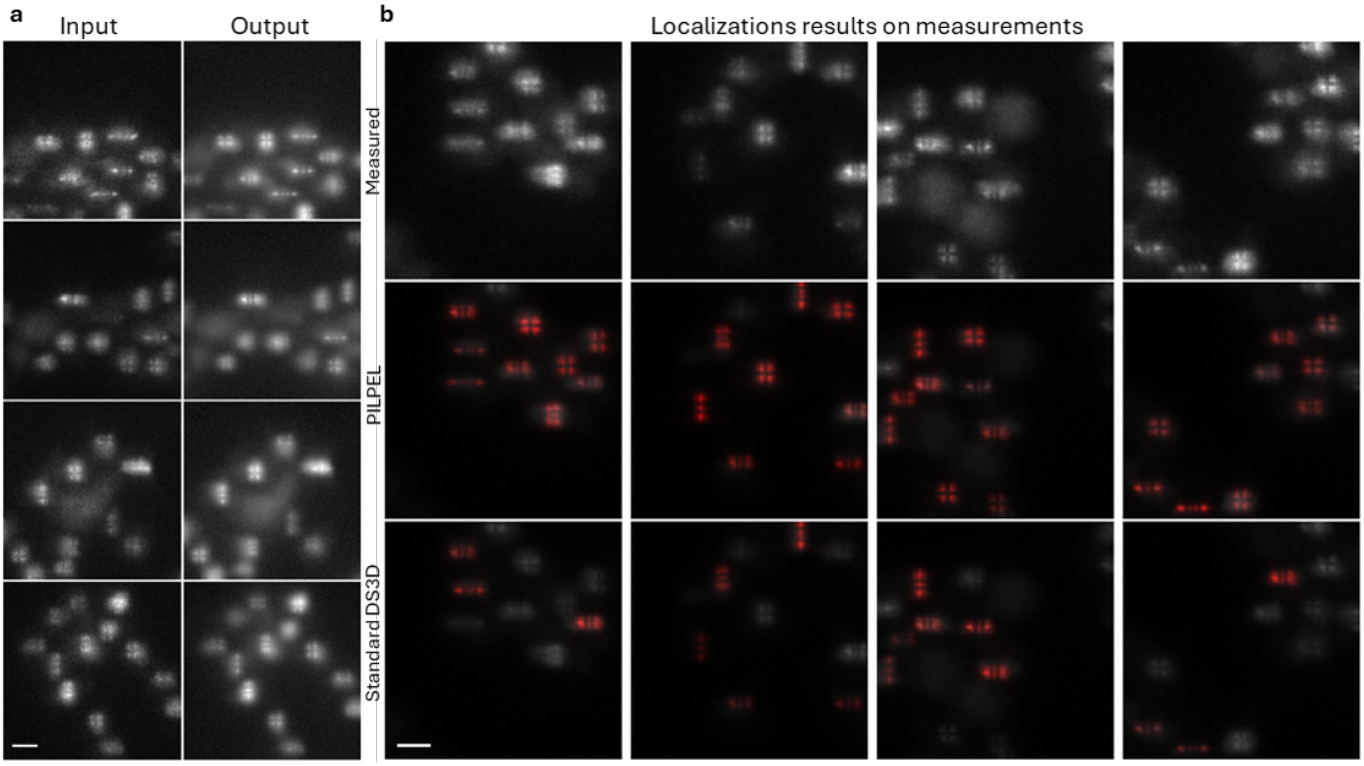


Fig. 6. Yeast Cell Experiment – Dense and Low SNR Dataset. a. PILPEL trained on real experimental data. Input - raw measured images, Output - their reconstructions generated (i.e. decoded) by the model. b. Comparison of localization results. 3D localizations rendering (red) are overlaying the measurement (white): First row shows four example experimental images. DS3D trained on the dataset generated using our method (second row) is compared versus DS3D trained on standard simulated data (third row). DS3D trained using our method detects a significantly higher number of emitters, apparent by the higher overlap (red on white) in the second row.

[4] S. J. Sahl, S. W. Hell, and S. Jakobs, "Fluorescence nanoscopy in cell biology," *Nature Reviews Molecular Cell Biology*, vol. 18, no. 11, pp. 685–701, Nov. 2017. [Online]. Available: https://www.nature.com/articles/nrm.2017.71

[5] H. Shen, L. J. Tauzin, R. Baiyasi, W. Wang, N. Moringo, B. Shuang, and C. F. Landes, "Single Particle Tracking: From Theory to Biophysical Applications," *Chemical Reviews*, vol. 117, no. 11, pp. 7331–7376, Jun. 2017. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.chemrev.6b00815

[6] Y. Shechtman, S. J. Sahl, A. S. Backer, and W. Moerner, "Optimal Point Spread Function Design for 3D Imaging," *Physical Review Letters*, vol. 113, no. 13, p. 133902, Sep. 2014. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.113.133902

[7] S. R. P. Pavani, M. A. Thompson, J. S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. E. Moerner, "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function," *Proceedings of the National Academy of Sciences*, vol. 106, no. 9, pp. 2995–2999, Mar. 2009. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.0900245106

[8] Y. Shechtman, L. E. Weiss, A. S. Backer, S. J. Sahl, and W. E. Moerner, "Precise Three-Dimensional Scan-Free Multiple-Particle Tracking over Large Axial Ranges with Tetrapod Point Spread Functions," *Nano Letters*, vol. 15, no. 6, pp. 4194–4199, Jun. 2015. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.nanolett.5b01396

[9] B. Huang, W. Wang, M. Bates, and X. Zhuang, "Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy," *Science*, vol. 319, no. 5864, pp. 810–813, Feb. 2008. [Online]. Available: https://www.science.org/doi/10.1126/science.1153529

[10] Y. Shechtman, L. E. Weiss, A. S. Backer, M. Y. Lee, and W. E. Moerner, "Multicolour localization microscopy by point-spread-function engineering," *Nature Photonics*, vol. 10, no. 9, pp. 590–594, Sep. 2016. [Online]. Available: https://www.nature.com/articles/nphoton.2016.137

[11] C. Smith, M. Huisman, M. Siemons, D. Grünwald, and S. Stallinga, "Simultaneous measurement of emission color and 3D position of single molecules," *Optics Express*, vol. 24, no. 5, p. 4996, Mar. 2016. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=oe-24-5-4996

[12] C. N. Hulleman, R. Thorsen, E. Kim, C. Dekker, S. Stallinga, and B. Rieger, "Simultaneous orientation and 3D localization microscopy with a Vortex point spread function," *Nature Communications*, vol. 12, no. 1, p. 5934, Oct. 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-26228-5

[13] A. S. Backer and W. E. Moerner, "Single-molecule orientation measurements with a quadrated pupil," *Optics Letters*, vol. 38, no. 9, pp. 1521–1523, 2013.

[14] C. N. Hulleman, R. O. Thorsen, B. Riegel, and S. Stallinga, "Simultaneous orientation and 3d localization microscopy with a vortex point spread function," *Nature Communications*, vol. 12, no. 1, p. 6228, 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-26228-5

[15] M. Lelek, M. T. Gyparaki, G. Beliu, F. Schueder, J. Griffié, S. Manley, R. Jungmann, M. Sauer, M. Lakadamyali, and C. Zimmer, "Single-molecule localization microscopy," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 39, Jun. 2021. [Online]. Available: https://www.nature.com/articles/s43586-021-00038-x

[16] A. Speiser, L.-R. Müller, P. Hoess, U. Matti, C. J. Obara, W. R. Legant, A. Kreshuk, J. H. Macke, J. Ries, and S. C. Turaga, "Deep learning enables fast and dense single-molecule localization with high accuracy," *Nature Methods*, vol. 18, no. 9, pp. 1082–1090, Sep. 2021. [Online]. Available: https://www.nature.com/articles/s41592-021-01236-x

[17] E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, T. Naor, R. Orange, T. Michaeli, and Y. Shechtman, "DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning," *Nature Methods*, vol. 17, no. 7, pp. 734–740, Jul. 2020. [Online]. Available: https://www.nature.com/articles/s41592-020-0853-5

[18] E. Hershko, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Multicolor localization microscopy and point-spread-function engineering by deep learning," *Optics Express*, vol. 27, no. 5,

p. 6158, Mar. 2019. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=oe-27-5-6158

[19] P. Zhang, S. Liu, A. Chaurasia, D. Ma, M. J. Mlodzianoski, E. Culurciello, and F. Huang, "Analyzing complex single-molecule emission patterns with deep learning," *Nature Methods*, vol. 15, no. 11, pp. 913–916, Nov. 2018. [Online]. Available: https://www.nature.com/articles/s41592-018-0153-5

[20] T. Wu, P. Lu, M. A. Rahman, X. Li, and M. D. Lew, "Deep-smolm: deep learning resolves the 3d orientations and 2d positions of overlapping single molecules with optimal nanoscale resolution," *Opt. Express*, vol. 30, no. 20, pp. 36 761–36 780, 2022. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?uri=oe-30-20-36761

[21] P. Jouchet, A. R. Roy, and W. E. Moerner, "Combining deep learning approaches and point spread function engineering for simultaneous 3d position and 3d orientation measurements of fluorescent single molecules," *Opt Commun*, vol. 542, p. 129589, 2023.

[22] O. Goldenberg, B. Ferdman, E. Nehme, Y. S. Ezra, and Y. Shechtman, "Learning optimal multicolor psf design for 3d pairwise distance estimation," *Intelligent Computing*, vol. 2022, p. 0004, 2022. [Online]. Available: https://spj.science.org/doi/abs/10.34133/icomputing.0004

[23] E. Nehme, B. Ferdman, L. E. Weiss, T. Naor, D. Freedman, T. Michaeli, and Y. Shechtman, "Learning optimal wavefront shaping for multi-channel imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2179–2192, 2021.

[24] D. Sage, T.-A. Pham, H. Babcock, T. Lukes, T. Pengo, J. Chao, R. Velmurugan, A. Herbert, A. Agrawal, S. Colabrese, A. Wheeler, A. Archetti, B. Rieger, R. Ober, G. M. Hagen, J.-B. Sibarita, J. Ries, R. Henriques, M. Unser, and S. Holden, "Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software," *Nature Methods*, vol. 16, no. 5, pp. 387–395, May 2019. [Online]. Available: https://www.nature.com/articles/s41592-019-0364-4

[25] T. Daniel and A. Tamar, "Unsupervised image representation learning with deep latent particles," in *International Conference on Machine Learning*. PMLR, 2022, pp. 4644–4665.

[26] ——, "Ddlp: Unsupervised object-centric video prediction with deep dynamic latent particles," *Transactions on Machine Learning Research*, 2024.

[27] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," Dec. 2022, arXiv:1312.6114 [stat]. [Online]. Available: http://arxiv.org/abs/1312.6114

[28] A. Gupta, Z. Wefers, K. Kahnert, J. N. Hansen, W. Leineweber, A. Cesnik, D. Lu, U. Axelsson, F. Ballllosera Navarro, T. Karaletsos, and E. Lundberg, "SubCell: Vision foundation models for microscopy capture single-cell biology," Dec. 2024. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2024.12.06.627299

[29] A. Archit, L. Freckmann, S. Nair, N. Khalid, P. Hilt, V. Rajashekar, M. Freitag, C. Teuber, G. Buckley, S. Von Haaren, S. Gupta, A. Dengel, S. Ahmed, and C. Pape, "Segment Anything for Microscopy," *Nature Methods*, vol. 22, no. 3, pp. 579–591, Mar. 2025. [Online]. Available: https://www.nature.com/articles/s41592-024-02580-4

[30] C. Ma, W. Tan, R. He, and B. Yan, "Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration," *Nature Methods*, vol. 21, no. 8, pp. 1558–1567, Aug. 2024. [Online]. Available: https://www.nature.com/articles/s41592-024-02244-3

[31] A. Truedson, K. Gras, and J. Elf, "The 3d point spread function can be autoencoded in latent space," *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2024/10/15/2024.10.14.618231

[32] A. Saguy, T. Nahimov, M. Lehrman, E. Gómez-de-Mariscal, I. Hidalgo-Cenalmor, O. Alalouf, A. Balakrishnan, M. Heilemann, R. Henriques, and Y. Shechtman, "This Microtubule Does Not Exist: Super-Resolution Microscopy Image Generation by a Diffusion Model," *Small Methods*, vol. 9, no. 3, p. 2400672, Mar. 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/smtd.202400672

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-To-Image Translation With Conditional Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html

[35] D. Haramati, T. Daniel, and A. Tamar, "Entity-centric reinforcement learning for object manipulation from pixels," in *The Twelfth International Conference on Learning Representations*, 2024.

[36] C. Qi, D. Haramati, T. Daniel, A. Tamar, and A. Zhang, "Ec-diffuser: Multi-object manipulation via entity-centric behavior generation," *arXiv preprint arXiv:2412.18907*, 2024.

[37] B. Ferdman, E. Nehme, L. E. Weiss, R. Orange, O. Alalouf, and Y. Shechtman, "VIPR: vectorial implementation of phase retrieval for fast and accurate microscopic pixel-wise pupil estimation," *Optics Express*, vol. 28, no. 7, p. 10179, Mar. 2020. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=oe-28-7-10179

[38] Y. S. Ezra, A. Saguy, G. Levin, L. E. Weiss, O. Alalouf, and Y. Shechtman, "Monitoring dna double strand break repair in saccharomyces cerevisiae at high throughput and high resolution," *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2024/08/25/2024.08.23.609327

[39] M. P. Backlund, R. Joyner, K. Weis, and W. E. Moerner, "Correlations of three-dimensional motion of chromosomal loci in yeast revealed by the double-helix point spread function microscope," *Molecular Biology of the Cell*, vol. 25, no. 22, pp. 3619–3629, Nov. 2014. [Online]. Available: https://www.molbiolcell.org/doi/10.1091/mbc.e14-06-1127

[40] S. Liu, J. Chen, J. Hellgoth, L.-R. Müller, B. Ferdman, C. Karras, D. Xiao, K. A. Lidke, R. Heintzmann, Y. Shechtman, Y. Li, and J. Ries, "Universal inverse modeling of point spread functions for SMLM localization and microscope characterization," *Nature Methods*, vol. 21, no. 6, pp. 1082–1093, Jun. 2024. [Online]. Available: https://www.nature.com/articles/s41592-024-02282-x