Physics-Informed Self-Supervised Generative Model for 3D Localization Microscopy

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

[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
ofrigoldenberg@campus.technion.ac.il, yoavsh@technion.ac.il

Abstract. Localization microscopy has overcome the diffraction limit, *i.e.* the conventional resolution limit of a microscope, enabling nanoscale biological imaging by precisely determining the positions of individual emitters such as single fluorescent molecules. However, the performance of deep learning methods, commonly applied to these tasks, 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 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 model of the Point Spread Function (PSF; the image of a single point source in the microscope) into the decoder, enabling it to disentangle learned realistic environments from emitters. 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 networks, particularly in challenging scenarios such as complex backgrounds and low signal-to-noise ratios. We demonstrate our approach on a variety of experimentally measured microscopy data, including super-resolution imaging in 2D and 3D and particle tracking in live cells, showing substantial improvements in localization precision and emitter detection. The code will be made publicly available.

Keywords: Simulation-to-experiment gap · Single-molecule localization microscopy · Physics-informed generative model · Self-supervised learning

1 Introduction

Determining the precise positions of single molecules or particles in optical microscopy images is a fundamental capability enabling insights at the nanoscale, with applications ranging from biology to materials science. Single-molecule localization microscopy (SMLM) techniques, including PALM [5], STORM [33], and PAINT [39], use sequential localization of emitters to reconstruct images that overcome the diffraction limit, allowing detailed biological visualization [37]. Single-Particle Tracking (SPT) utilizes localization over time to determine the trajectories of moving molecules or nanoparticles, revealing information about their dynamics, interactions, and microenvironments [43]. 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 [15,31,40,42] or other information, such as spectral properties [41,44] or molecular orientation [2,16,17].

Deep learning methods have become essential tools in microscopy and SMLM, enabling powerful analysis and reconstruction [22]. They have been applied to various challenges, including the localization of dense emitters [26,45], the classification of emitter color [13], the determination of molecular orientation [19,47,48], and the facilitation of end-to-end optimization of imaging systems, *e.g.* designing the PSF together with the reconstruction algorithm [9,13,25,26]. Importantly, the performance of many deep learning approaches, particularly supervised methods such as DeepSTORM3D [26], 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 for 3D localization microscopy is practically impossible.

However, building simulations that perfectly mimic experimental conditions is challenging [34]. Accurately capturing complex, spatially varying background, realistic noise statistics, and system-specific aberrations 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, we aim to bridge the sim2exp gap by introducing a physics-informed self-supervised generative framework that learns directly from experimental data. We propose *PILPEL – Physics-Informed Latent Particles for Emitter Localization*, an object-centric self-supervised generative model built on the Deep Latent Particles (DLP [6,7]) framework. We extend DLP by incorporating an explicit physical PSF model into the decoder that takes encoded spatial coordinates as input. Trained directly on unlabeled experimental images, this architecture allows the model to intrinsically learn realistic background and noise patterns while disentangling the learned environment from the precise placement and properties of the emitters.

The trained PILPEL model thus serves as a data generator, producing realistic, fully-labeled training images with known ground truth emitter locations. This generated data, embodying experimentally learned characteristics, serves as highly realistic training material for supervised localization algorithms. Using these generated datasets significantly improves the performance of supervised localization networks on real-world data, particularly in challenging scenarios such as complex backgrounds and low signal-to-noise ratios, 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 generalizability and applicability advantage of automatically learning data characteristics directly from measurements, bridging the sim2exp gap without manual tuning of simulation parameters, or explicit modeling of experimental attributes, thereby unlocking the application of supervised localization networks to novel biological domains.

2 Related Work

Deep learning for microscopy. Recent large vision foundation models, including adaptations for microscopy [1, 11, 24], show promise for tasks such as segmentation, cellular mapping or classification, but their direct application to generating high-precision localization microscopy data is challenging. These models are often pre-trained 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 signal-to-noise ratio (SNR), or involving engineered PSFs where computationally designed optics alter image characteristics. Concurrently, VAE models have been shown to have the ability to encode emitters' z-position in their latent space [46]. Our work contributes to this area by introducing a physics-informed VAE-based model, designed to generate realistic, label-consistent localization-microscopy images.

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. [35] to synthesize realistic super-resolution microscopy images from experimental data for data augmentation. However, the generated data is typically unlabeled, and thus can not be used for training localization models.

Other generative approaches, such as Generative Adversarial Networks (GANs) [4, 10, 18, 49], also attempt to bridge the sim2exp gap by learning mappings from simulated to real domains or generating images from labels. However, such post-hoc image transformations are fundamentally incompatible with the sub-pixel precision demands of SMLM. Any pixel-level modification of a simulated image, however subtle, can deform the PSF shape or shift its apparent center, breaking the correspondence between the image content and the original ground-truth emitter coordinates. At nanoscale resolution, where localization errors on the order of tens of nanometers are meaningful, there is no tolerance for label corruption introduced by an unconstrained generative process. Without an explicit physical model of image formation embedded within the generator itself, there is no mechanism to guarantee that the ground-truth labels remain valid after the image is transformed. This is precisely the design principle behind PILPEL: by placing the PSF model inside the decoder, emitter coordinates are not labels imposed on a generated image - they are the direct cause of it, ensuring exact label integrity by construction. Since, to the best of our knowledge, no other generative baseline outputs 3D sub-pixel coordinates, we compare our generated data against that of the DeepSTORM3D [26] physics-based simulator.

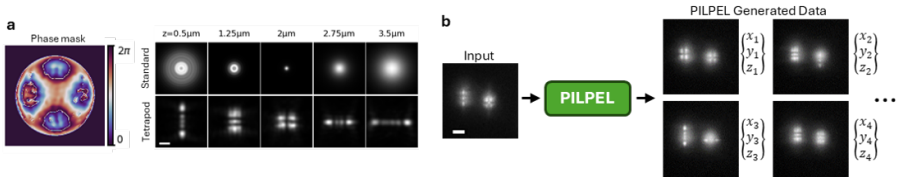

Fig. 1: PILPEL – Physics-Informed Latent Particles for Emitter Localization. (a) A Tetrapod phase mask incorporated in the microscope encodes the emitter's axial position in the shape of the PSF. Scale bar, 1 µm. (b) PILPEL use case. Input – An experimental image of a fluorescently labeled DNA loci in yeast cells, imaged 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). PILPEL learns to generate realistic, fully-labeled images (output examples, right) with known GT coordinates for localization. Scale bar, 2 µm.

3 Background

DeepSTORM3D (DS3D [26]) is a deep learning-based method designed for dense 3D SMLM over large axial ranges. It utilizes a convolutional neural network (CNN) to analyze 2D images containing overlapping engineered PSFs (such as the depth-encoding Tetrapod PSF, Fig. 1a) 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 DS3D analysis pipeline is the generation of training data that resembles the experiment. This involves significant manual tuning and can be highly time-consuming even for an experienced user. Our model, among other contributions, alleviates this manual tuning completely.

In this work, we show results with variations of the Tetrapod mask, but the flexibility of the model makes it applicable to any PSF (*e.g.* astigmatism [15], double-helix [31], etc.), by inserting a different phase mask and accompanied optical parameters to the PSF model. Fig. 1b shows an example application of 3D localization: several yeast cells with fluorescently labeled DNA loci are apparent in the field of view (FOV). The goal here would be to determine the 3D positions of all loci.

Deep Latent Particles (DLP [6, 7]) is a self-supervised, object-centric image representation model that decomposes an input image into a set of latent "particles," each representing a localized foreground object with learned attributes such as position, scale, transparency, and visual features. An additional background particle represents the background visual features. This enables a latent representation that disentangles the objects from the scene. A key component of the original DLP is its use of a spatial softmax (SSM) operator as a spatial prior to particle positions. The model's objective is to reconstruct the original input image, which is done by decoding the latent particles and composing them together with the background to form the output image. Training is performed end-to-end using a variational autoencoder (VAE [21]) framework, by maximizing the evidence lower bound (ELBO), combining image reconstruction loss with a KL-divergence regularization term. For more details, see supplementary Sec. B.

4 Method

Our goal is to design a *controllable* generative model that can generate realistic image data for the task of SMLM (Fig. 1b). 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 (Fig. 1). Accurate localization of these PSFs is essential for achieving nanoscale resolution in SMLM reconstructions. This motivates the use of object-centric generative models, such as DLP, which offer fine-grained

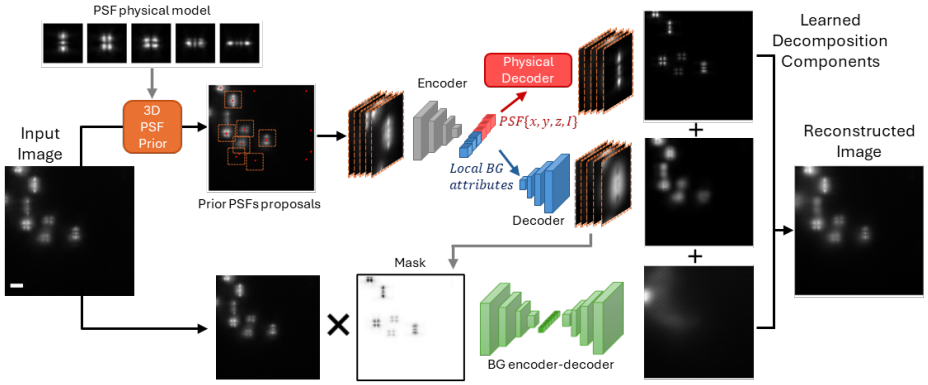

Fig. 2: Physics-informed encoding-decoding with latent representation of PSF-objects. The model divides the input image into overlapping patches and performing 3D template matching using a precomputed PSF dictionary to propose candidate emitter locations. Patches around these locations are then processed by an encoder that generates disentangled latent vectors encoding spatial coordinates, 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 emitter localizations, that can be used as a training sample for supervised localization networks. Scale bar, 2 µm.

control through explicit attributes like spatial coordinates, which are essential for precise localization.

The DLP model decomposes an image into latent particles with explicit 2D positions. Given the challenges of collecting labeled data in real-world microscopy settings, we adapt DLP for localization microscopy by enabling it to recover 3D positions and intensities of point-emitters, observed as individual PSFs.

To this end, we extend the original DLP model to explicitly encode the key emitter properties—3D spatial coordinates ($\{x, y, z\}$) and intensity (I)—such that each latent particle corresponds to a single emitter. Crucially, the model will learn additional image components such as local and global background, SNR, and effective emitter density. We term this extended framework **PILPEL**: **P**hysics-**I**nformed **L**atent **P**articles for **E**mitter **L**ocalization.

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 their surrounding scene, including 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 calibrated, for example, by using phase retrieval methods such as VIPR [8]. Similar to DS3D [26], 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 PSFs of each reconstructed patch.

4.2 Model Architecture

Our model architecture is illustrated in Fig. 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 is used as a confidence metric to filter out patches unlikely to contain any PSFs, resulting in a proposed set of latent particles.

Following this likelihood-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 module. 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. The intensity factor l_I allows the model to capture experiment-specific SNR and brightness distributions, and enables an automatic pruning mechanism — by driving the factor to zero during optimization — which adapts to varying emitter densities by removing unnecessary proposals.

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. These are combined and placed into the final reconstructed image at their corresponding global coordinates.

To learn global background characteristics, the generated object patches are used to mask the 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. An additional encoder learns a latent attribute l_n, used as the standard deviation of random noise added to the general background.

Finally, the generated components (PSFs, local background patches and 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 output image and the original input image, and KL-divergence regularization term for the stochastic latent variables.

An ablation study evaluating the contribution of each module is provided in Supplementary Sec. D.

4.3 Data Generation for Supervised Learning

The decoder of our trained PILPEL model serves as a high-fidelity generator of realistic labeled images. By sampling coordinates and latent variables from the latent space — where the newly provided coordinates to the physical layer are, by definition, the ground truth labels for the generated image — the decoder produces images with known emitter locations and intensities. Most importantly and key to our application: these generated, labeled images inherit the realistic background and noise characteristics learned from the experimental data, as well as emitter densities, global structures and spatial distribution (see Supplementary Fig. S5).

This capability allows for the generation of large and accurately labeled training datasets for supervised localization models, specifically DS3D [26] in this work. This data closely mimic real measurements, thereby narrowing the sim2exp gap while automatically capturing unmodeled experimental nuances.

5 Experiments

Our goal in this section is to show that the generation of realistic data improves the downstream performance of localization tasks. We first establish the method's accuracy on controlled simulations with known ground truth, then demonstrate its impact on real biological data – three datasets of 3D localization and one dataset of 2D localization. Full experimental details are provided in the Supplementary.

5.1 Bridging the sim2exp Gap with Learned Data

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 [28,38] (Fig. 1a), simulated by adding random asymmetrical 2D Gaussians around each PSF to model the typical autofluorescence observed in yeast cells.

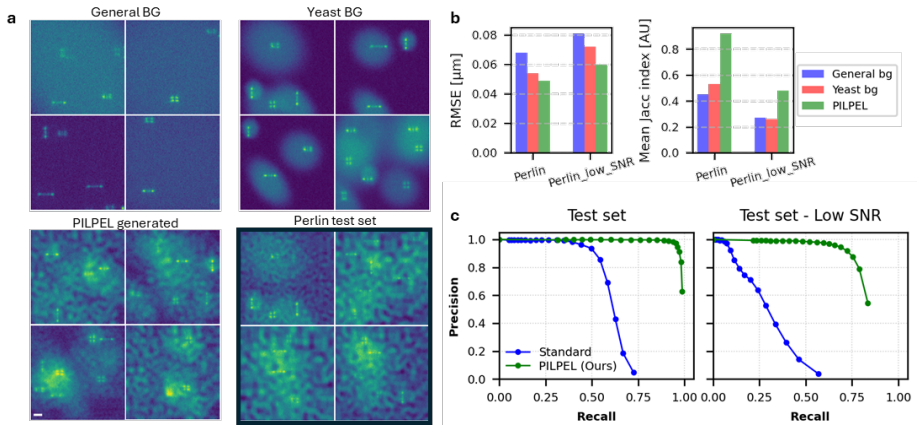

Fig. 3: Robustness experiment on simulations – DS3D localization performance. (a) Comparison of synthetic training datasets used for DS3D localization of the Perlin test set (left column, simulates yeast cells with Perlin noise model). General BG – the standard background in DS3D; Yeast BG – includes local background around emitters to mimic yeast cells; and PILPEL – generated via our model. Scale bar, 2 µm. (b) Improvement of training on generated data. Localization RMSE (lower is better) and Jaccard index (higher is better) for DS3D trained on the generated dataset vs. training on yeast-BG and original-DS3D datasets. (c) Precision-Recall analysis on the PerlinData test set. PILPEL (green) demonstrates superior detection robustness compared to the standard DeepSTORM3D baseline (blue). While the baseline's precision degrades rapidly—particularly in the Low-SNR regime—PILPEL maintains high precision (> 0.9) across a significantly wider recall range.

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, used in microscopy image analysis to model non-uniform illumination and other artifacts [45]. 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, practically introducing prior knowledge manually); and (3) a dataset generated by our model after training on the `PerlinData` test set.

Fig. 3a illustrates these datasets, showing that, as expected, the generated data most closely resembles the test set. PILPEL successfully learned Perlin noise with various levels and complex 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. Consequently, DS3D trained on this generated data achieves superior performance, both in localization accuracy and in the number of detected emitters. Fig. 3b summa-

rizes the performance gains of using the generated dataset under challenging conditions, split into `PerlinData` and `low-SNR-PerlinData` to show the benefit is even greater in low-SNR cases. 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, compared to the baseline. While introducing prior knowledge into synthetic training data improves performance (yeast background, red bar), training on generated data (PILPEL, green bar) is still significantly more effective, and also eliminates the need for implementing additional physical attributes in the DS3D simulator. To further evaluate the detection capability of the proposed model and its robustness against false positives, we performed a Precision-Recall (P-R) analysis on the `PerlinData` test set. Ground truth emitters were matched to localized coordinates using a strict lateral distance threshold of 100 nm, the detection confidence threshold varied from 0 to 800 (the range of output possibilities) to generate the curves. As shown in Fig. 3c, there is a decisive performance gap between our method and the standard baseline. The baseline fails significantly on unmodeled noise, with precision collapsing after approximately 50% recall. In contrast, PILPEL maintains near-perfect precision ($\sim$1.0) up to $\sim$95% recall. Notably, at a fixed high precision of 0.9, PILPEL yields approximately 2$\times$ higher recall than the baseline. This advantage is even more pronounced in the challenging Low-SNR regime. Here, PILPEL maintains > 0.9 precision up to 75% recall, whereas the baseline fails (dropping below 0.9 precision) at only 15% recall. These results confirm that our approach significantly mitigates hallucinated artifacts while maintaining high sensitivity. See supplementary Sec. C for extended details.

5.2 Super-resolution 3D Imaging

Having validated our approach under controlled conditions with known ground truth, we next evaluate its performance on complex, real-world biological data. Specifically, we apply the method to a 3D super-resolution dataset of mitochondria [29] imaged via SMLM (see Supplementary Sec. E for experimental details). As shown in Fig. 4, training on our generated data dramatically improves the final reconstruction. The number of detected localizations increased from 119K to 581K ($\sim 4.9\times$ increase) across 20K frames. This factor, namely, the number of localizations, is crucial in SMLM, and translates to improved image quality, *i.e.* fuller features (as seen in Fig. 4b), or faster acquisition. Fig. 4c shows measured frames and corresponding PILPEL reconstructions.

It is possible that this significant increase in detections also includes a higher rate of false positives; however, this can be addressed using standard density filters commonly applied in SMLM reconstruction [30]. For the visualization in Fig. 4, the reconstruction from our method was post-processed with such a filter to match the original's noise level, while the original reconstruction was left as-is, as filtering would degrade the sparse biological structure. Crucially, the high detection rate achieved by our method makes the application of such filters feasible. While our dense reconstructions survive filtering and preserve

the underlying structure, the baseline's sparse detections vanish under the same processing. The unfiltered version is available in Supplementary Fig. S4.

Additionally, our method learns 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 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. While recent work aims to automate this process [36], such approaches are still limited by the simulator's explicit model and pre-defined parameters, which may not capture unmodeled traits that our generative approach learns directly.

5.3 3D Localization of DNA Loci in Yeast Cells

Beyond super-resolution imaging, we further assess the generalizability of our approach in a distinct biological system characterized by local background features not modeled in the original DS3D simulator. We apply the same pipeline to experimental images of fluorescently labeled DNA loci in fixed yeast cells [38], where each cell exhibits a single fluorescent spot shaped by the Tetrapod phase mask (see Supplementary Sec. F for experimental details). In such a biological system, an example application would be tracking loci over time in 3D under different conditions [3,38,43]; here, we use it to demonstrate our method under a noise regime distinct from STORM.

Dense low-SNR measurements were used as a challenging dataset for DS3D localization, comparing a model trained with the original simulator versus a model trained on generated data derived via our data generation pipeline. Fig. 5a shows the measured dataset and the corresponding PILPEL reconstructions. Fig. 5b displays the measured images alongside localization results from both localization networks. The general detection rate increased by $\sim 1.47\times$, from 12,561 detections with the original net to 18,461 detections using the PILPEL-generated dataset, across three time-lapse measurements (100 time points each) in 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 standard GT estimation in single cells is MLE localization - fitting xyz coordinates of the PSF model together with asymmetric Gaussian background. This is relevant when the model matches the measurement, and MLE is not 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 test set 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

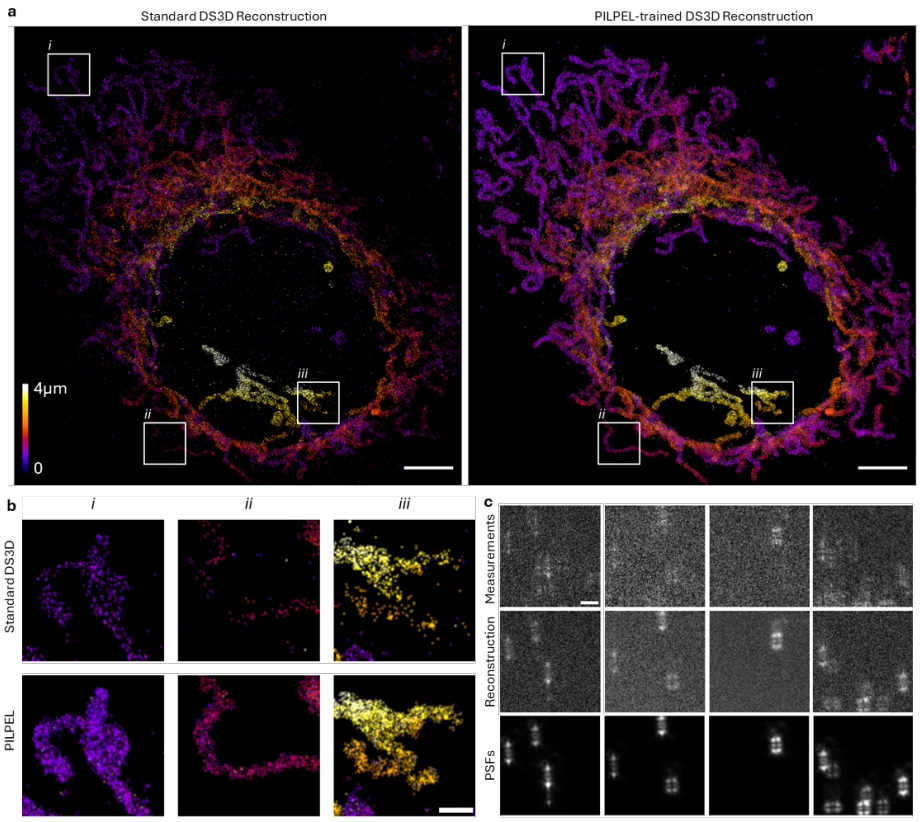

Fig. 4: Comparison of 3D super-resolution reconstruction of experimental data. (a) Super-resolved reconstructions of mitochondria (depth $\approx 4\,\mu\text{m}$), rendered as 2D histograms with color-coded z-positions, using DS3D trained on the original simulated data (left) versus our generated data (right). Scale bar, $5\,\mu\text{m}$. (b) Magnified views of regions i, ii, and iii as indicated in (a). Scale bar, $1\,\mu\text{m}$. (c) Examples of randomly sampled measured images, accompanied with their matching PILPEL reconstruction generated by our model, used to train DS3D. Bottom row shows the PSFs component of the reconstructions, defined by $\{x, y, z\}$ coordinates, that serves as the GT training labels. Scale bar, $3\,\mu\text{m}$.

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.

Comparing to the synthetic yeast-BG dataset from Sec. 5.1, our model's ability to learn unmodeled features is clearly demonstrated. While one can incorporate local background to the DS3D simulator, Fig. 3a shows that it still does not adequately resemble the actual measurements, seen in Fig. 5. 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.

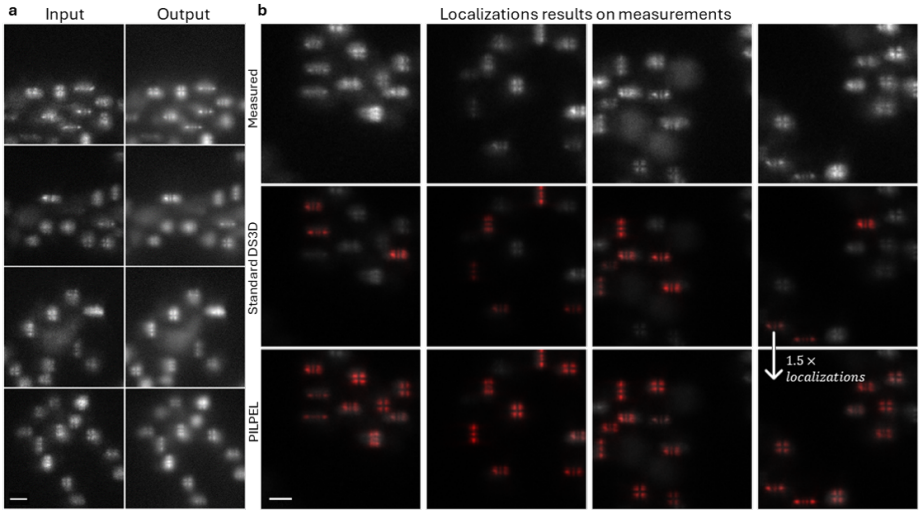

Fig. 5: Fluorescently labeled DNA loci in yeast cells. (a) PILPEL trained on experimental data, showing dense yeast cells with pronounced, spatially varying local background. Input - raw measured images; Output - their reconstructions generated by the model. Reconstructions capture cell-specific background and noise patterns, illustrating that the model learns biological structures directly from the data. (b) Comparison of localization results. Localization overlays on the same measurements (white) with 3D detections rendered (red). Middle row: DS3D trained on standard simulated data; bottom row: DS3D trained on PILPEL-generated data. Evidently, the PILPEL-trained network yields visibly higher red-on-white overlap and recovers more emitters missed by the standard baseline. Scale bars, 3 µm.

5.4 Additional Experiments

To further demonstrate the generalizability of our method, we applied PILPEL to two additional datasets (detailed in Supplementary Sec. G and Sec. H).

First, a **2D super-resolution reconstruction of microtubules**, demonstrating performance enhancement of another localization network in 2D (DeepSTORM [27]). Second, we apply our model to **3D localization data from bacteria**. This dataset utilizes a different PSF (astigmatism) and exhibits a distinct background profile compared to the yeast cell data. These results underscore our model's robustness and applicability to varied data conditions and SMLM modalities.

6 Limitations and Future Work

Despite the advances demonstrated, our method exhibits several limitations. Training the generative model requires additional computational time compared to training on synthetic datasets, although it significantly reduces manual human intervention (see Supplementary Material for computational details). Additionally, accurate reconstruction still necessitates a reliable prior PSF model, obtainable by a numerical model or from experimental calibration followed by phase retrieval. Moreover, the effectiveness of our approach depends 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 includes leveraging this generative capability for complex-to-simulate microscopy scenarios (e.g., particle flows, living organisms like C. elegans), extending to dynamic tracking via DDLP for video [7], and exploring latent space manipulation for tailored dataset generation for downstream tasks (e.g., denoising). Another promising direction is to incorporate learned aberrations [23], enabling also the optimization of the PSF model itself.

7 Conclusions

We introduced a physics-informed generative model that bridges the critical sim2exp gap in localization microscopy. By integrating an explicit PSF model within a self-supervised, object-centric framework, the model can disentangle the emitter properties from complex background and noise characteristics. Trained on real experimental measurements, our approach generates high-fidelity labeled training data that captures the nuances of real measurements. As our benchmarks demonstrate, training supervised models on this generated data significantly enhances localization accuracy and detection rates, especially for challenging or novel experimental conditions. This approach eliminates the need for manual simulation tuning and the explicit modeling of complex image elements, offering a robust, adaptable solution for improving deep learning performance in microscopy.

References

1. Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., Teuber, C., Buckley, G., Von Haaren, S., Gupta, S., Dengel, A., Ahmed, S.,

Pape, C.: Segment Anything for Microscopy. Nature Methods **22**(3), 579–591 (Mar 2025). https://doi.org/10.1038/s41592-024-02580-4, https://www.nature.com/articles/s41592-024-02580-4
2. Backer, A.S., Moerner, W.E.: Single-molecule orientation measurements with a quadrated pupil. Optics Letters **38**(9), 1521–1523 (2013). https://doi.org/10.1364/OL.38.001521
3. Backlund, M.P., Joyner, R., Weis, K., Moerner, W.E.: Correlations of three-dimensional motion of chromosomal loci in yeast revealed by the double-helix point spread function microscope. Molecular Biology of the Cell **25**(22), 3619–3629 (Nov 2014). https://doi.org/10.1091/mbc.e14-06-1127, https://www.molbiolcell.org/doi/10.1091/mbc.e14-06-1127
4. Baniukiewicz, P., Lutton, E.J., Collier, S., Bretschneider, T.: Generative adversarial networks for augmenting training data of microscopic cell images. Frontiers in Computer Science **1**, 10 (2019). https://doi.org/10.3389/fcomp.2019.00010, https://www.frontiersin.org/articles/10.3389/fcomp.2019.00010/full
5. Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., Hess, H.F.: Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. Science **313**(5793), 1642–1645 (Sep 2006). https://doi.org/10.1126/science.1127344, https://www.science.org/doi/10.1126/science.1127344
6. Daniel, T., Tamar, A.: Unsupervised image representation learning with deep latent particles. In: International Conference on Machine Learning. pp. 4644–4665. PMLR (2022)
7. Daniel, T., Tamar, A.: Ddlp: Unsupervised object-centric video prediction with deep dynamic latent particles. Transactions on Machine Learning Research (2024)
8. Ferdman, B., Nehme, E., Weiss, L.E., Orange, R., Alalouf, O., Shechtman, Y.: VIPR: vectorial implementation of phase retrieval for fast and accurate microscopic pixel-wise pupil estimation. Optics Express **28**(7), 10179 (Mar 2020). https://doi.org/10.1364/OE.388248, https://opg.optica.org/abstract.cfm?URI=oe-28-7-10179
9. Goldenberg, O., Ferdman, B., Nehme, E., Ezra, Y.S., Shechtman, Y.: Learning optimal multicolor psf design for 3d pairwise distance estimation. Intelligent Computing **2022**, 0004 (2022). https://doi.org/10.34133/icomputing.0004, https://spj.science.org/doi/abs/10.34133/icomputing.0004
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html
11. Gupta, A., Wefers, Z., Kahnert, K., Hansen, J.N., Leineweber, W., Cesnik, A., Lu, D., Axelsson, U., Ballllosera Navarro, F., Karaletsos, T., Lundberg, E.: SubCell: Vision foundation models for microscopy capture single-cell biology (Dec 2024). https://doi.org/10.1101/2024.12.06.627299, http://biorxiv.org/lookup/doi/10.1101/2024.12.06.627299
12. Haramati, D., Daniel, T., Tamar, A.: Entity-centric reinforcement learning for object manipulation from pixels. In: The Twelfth International Conference on Learning Representations (2024)
13. Hershko, E., Weiss, L.E., Michaeli, T., Shechtman, Y.: Multicolor localization microscopy and point-spread-function engineering by deep learning. Optics Express **27**(5), 6158 (Mar 2019). https://doi.org/10.1364/OE.27.006158, https://opg.optica.org/abstract.cfm?URI=oe-27-5-6158

14. Hoshen, Y., Li, K., Malik, J.: Non-adversarial image synthesis with generative latent nearest neighbors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5811–5819 (2019)
15. Huang, B., Wang, W., Bates, M., Zhuang, X.: Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy. Science **319**(5864), 810–813 (Feb 2008). https://doi.org/10.1126/science.1153529, https://www.science.org/doi/10.1126/science.1153529
16. Hulleman, C.N., Thorsen, R.O., Riegel, B., Stallinga, S.: Simultaneous orientation and 3d localization microscopy with a vortex point spread function. Nature Communications **12**(1), 6228 (2021). https://doi.org/10.1038/s41467-021-26228-5, https://doi.org/10.1038/s41467-021-26228-5
17. Hulleman, C.N., Thorsen, R., Kim, E., Dekker, C., Stallinga, S., Rieger, B.: Simultaneous orientation and 3D localization microscopy with a Vortex point spread function. Nature Communications **12**(1), 5934 (Oct 2021). https://doi.org/10.1038/s41467-021-26228-5, https://www.nature.com/articles/s41467-021-26228-5
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-To-Image Translation With Conditional Adversarial Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017), https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html
19. Jouchet, P., Roy, A.R., Moerner, W.E.: Combining deep learning approaches and point spread function engineering for simultaneous 3d position and 3d orientation measurements of fluorescent single molecules. Opt Commun **542**, 129589 (2023). https://doi.org/10.1016/j.optcom.2023.129589
20. Karempudi, P., Gras, K., Amselem, E., Zikrin, S., Schirman, D., Elf, J.: Three-dimensional localization and tracking of chromosomal loci throughout the Escherichia coli cell cycle. Communications Biology **7**(1), 1443 (Nov 2024). https://doi.org/10.1038/s42003-024-07155-9, https://doi.org/10.1038/s42003-024-07155-9
21. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes (Dec 2022). https://doi.org/10.48550/arXiv.1312.6114, http://arxiv.org/abs/1312.6114, arXiv:1312.6114 [stat]
22. Lelek, M., Gyparaki, M.T., Beliu, G., Schueder, F., Griffié, J., Manley, S., Jungmann, R., Sauer, M., Lakadamyali, M., Zimmer, C.: Single-molecule localization microscopy. Nature Reviews Methods Primers **1**(1), 39 (Jun 2021). https://doi.org/10.1038/s43586-021-00038-x, https://www.nature.com/articles/s43586-021-00038-x
23. Liu, S., Chen, J., Hellgoth, J., Müller, L.R., Ferdman, B., Karras, C., Xiao, D., Lidke, K.A., Heintzmann, R., Shechtman, Y., Li, Y., Ries, J.: Universal inverse modeling of point spread functions for SMLM localization and microscope characterization. Nature Methods **21**(6), 1082–1093 (Jun 2024). https://doi.org/10.1038/s41592-024-02282-x, https://www.nature.com/articles/s41592-024-02282-x
24. Ma, C., Tan, W., He, R., Yan, B.: Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. Nature Methods **21**(8), 1558–1567 (Aug 2024). https://doi.org/10.1038/s41592-024-02244-3, https://www.nature.com/articles/s41592-024-02244-3
25. Nehme, E., Ferdman, B., Weiss, L.E., Naor, T., Freedman, D., Michaeli, T., Shechtman, Y.: Learning optimal wavefront shaping for multi-channel imaging.

IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(7), 2179–2192 (2021). https://doi.org/10.1109/TPAMI.2021.3076873

26. Nehme, E., Freedman, D., Gordon, R., Ferdman, B., Weiss, L.E., Alalouf, O., Naor, T., Orange, R., Michaeli, T., Shechtman, Y.: DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning. Nature Methods **17**(7), 734–740 (Jul 2020). https://doi.org/10.1038/s41592-020-0853-5, https://www.nature.com/articles/s41592-020-0853-5
27. Nehme, E., Weiss, L.E., Michaeli, T., Shechtman, Y.: Deep-STORM: super-resolution single-molecule microscopy by deep learning. Optica **5**(4), 458 (Apr 2018). https://doi.org/10.1364/OPTICA.5.000458, https://opg.optica.org/abstract.cfm?URI=optica-5-4-458
28. Opatovski, N., Shalev Ezra, Y., Weiss, L.E., Ferdman, B., Orange-Kedem, R., Shechtman, Y.: Multiplexed psf engineering for three-dimensional multicolor particle tracking. Nano Letters **21**(13), 5888–5895 (2021)
29. Orange kedem, R., Opatovski, N., Xiao, D., Ferdman, B., Alalouf, O., Kumar Pal, S., Wang, Z., von der Emde, H., Weber, M., Sahl, S.J., Ponjavic, A., Arie, A., Hell, S.W., Shechtman, Y.: Near index matching enables solid diffractive optical element fabrication via additive manufacturing. Light: Science & Applications **12**(1), 222 (Sep 2023). https://doi.org/10.1038/s41377-023-01277-1, https://doi.org/10.1038/s41377-023-01277-1
30. Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z., Hagen, G.M.: Thunderstorm: a comprehensive imagej plug-in for palm and storm data analysis and super-resolution imaging. Bioinformatics **30**(16), 2389–2390 (05 2014). https://doi.org/10.1093/bioinformatics/btu202, https://doi.org/10.1093/bioinformatics/btu202
31. Pavani, S.R.P., Thompson, M.A., Biteen, J.S., Lord, S.J., Liu, N., Twieg, R.J., Piestun, R., Moerner, W.E.: 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 **106**(9), 2995–2999 (Mar 2009). https://doi.org/10.1073/pnas.0900245106, https://pnas.org/doi/full/10.1073/pnas.0900245106
32. Qi, C., Haramati, D., Daniel, T., Tamar, A., Zhang, A.: Ec-diffuser: Multi-object manipulation via entity-centric behavior generation. arXiv preprint arXiv:2412.18907 (2024)
33. Rust, M.J., Bates, M., Zhuang, X.: Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nature Methods **3**(10), 793–796 (Oct 2006). https://doi.org/10.1038/nmeth929, https://www.nature.com/articles/nmeth929
34. Sage, D., Pham, T.A., Babcock, H., Lukes, T., Pengo, T., Chao, J., Velmurugan, R., Herbert, A., Agrawal, A., Colabrese, S., Wheeler, A., Archetti, A., Rieger, B., Ober, R., Hagen, G.M., Sibarita, J.B., Ries, J., Henriques, R., Unser, M., Holden, S.: Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software. Nature Methods **16**(5), 387–395 (May 2019). https://doi.org/10.1038/s41592-019-0364-4, https://www.nature.com/articles/s41592-019-0364-4
35. Saguy, A., Nahimov, T., Lehrman, M., Gómez-de-Mariscal, E., Hidalgo-Cenalmor, I., Alalouf, O., Balakrishnan, A., Heilemann, M., Henriques, R., Shechtman, Y.: This Microtubule Does Not Exist: Super-Resolution Microscopy Image Generation by a Diffusion Model. Small Methods **9**(3), 2400672 (Mar 2025). https:

//doi.org/10.1002/smtd.202400672, https://onlinelibrary.wiley.com/doi/10.1002/smtd.202400672

36. Saguy, A., Xiao, D., Narayanasamy, K.K., Nakatani, Y., Gustavsson, A.K., Heilemann, M., Shechtman, Y.: One-click image reconstruction in single-molecule localization microscopy via deep learning. bioRxiv (2025). https://doi.org/10.1101/2025.04.13.648574, https://www.biorxiv.org/content/early/2025/04/18/2025.04.13.648574
37. Sahl, S.J., Hell, S.W., Jakobs, S.: Fluorescence nanoscopy in cell biology. Nature Reviews Molecular Cell Biology **18**(11), 685–701 (Nov 2017). https://doi.org/10.1038/nrm.2017.71, https://www.nature.com/articles/nrm.2017.71
38. Shalev Ezra, Y., Saguy, A., Levin, G., Weiss, L.E., Alalouf, O., Shechtman, Y.: High-throughput dna repair monitoring in saccharomyces cerevisiae suggests ssb- and dsb-induced chromatin reconfiguration. Scientific Reports **15**(1), 32302 (2025)
39. Sharonov, A., Hochstrasser, R.M.: Wide-field subdiffraction imaging by accumulated binding of diffusing probes. Proceedings of the National Academy of Sciences **103**(50), 18911–18916 (Dec 2006). https://doi.org/10.1073/pnas.0609643104, https://pnas.org/doi/full/10.1073/pnas.0609643104
40. Shechtman, Y., Sahl, S.J., Backer, A.S., Moerner, W.: Optimal Point Spread Function Design for 3D Imaging. Physical Review Letters **113**(13), 133902 (Sep 2014). https://doi.org/10.1103/PhysRevLett.113.133902, https://link.aps.org/doi/10.1103/PhysRevLett.113.133902
41. Shechtman, Y., Weiss, L.E., Backer, A.S., Lee, M.Y., Moerner, W.E.: Multicolour localization microscopy by point-spread-function engineering. Nature Photonics **10**(9), 590–594 (Sep 2016). https://doi.org/10.1038/nphoton.2016.137, https://www.nature.com/articles/nphoton.2016.137
42. Shechtman, Y., Weiss, L.E., Backer, A.S., Sahl, S.J., Moerner, W.E.: Precise Three-Dimensional Scan-Free Multiple-Particle Tracking over Large Axial Ranges with Tetrapod Point Spread Functions. Nano Letters **15**(6), 4194–4199 (Jun 2015). https://doi.org/10.1021/acs.nanolett.5b01396, https://pubs.acs.org/doi/10.1021/acs.nanolett.5b01396
43. Shen, H., Tauzin, L.J., Baiyasi, R., Wang, W., Moringo, N., Shuang, B., Landes, C.F.: Single Particle Tracking: From Theory to Biophysical Applications. Chemical Reviews **117**(11), 7331–7376 (Jun 2017). https://doi.org/10.1021/acs.chemrev.6b00815, https://pubs.acs.org/doi/10.1021/acs.chemrev.6b00815
44. Smith, C., Huisman, M., Siemons, M., Grünwald, D., Stallinga, S.: Simultaneous measurement of emission color and 3D position of single molecules. Optics Express **24**(5), 4996 (Mar 2016). https://doi.org/10.1364/OE.24.004996, https://opg.optica.org/abstract.cfm?URI=oe-24-5-4996
45. Speiser, A., Müller, L.R., Hoess, P., Matti, U., Obara, C.J., Legant, W.R., Kreshuk, A., Macke, J.H., Ries, J., Turaga, S.C.: Deep learning enables fast and dense single-molecule localization with high accuracy. Nature Methods **18**(9), 1082–1090 (Sep 2021). https://doi.org/10.1038/s41592-021-01236-x, https://www.nature.com/articles/s41592-021-01236-x
46. Truedson, A., Gras, K., Elf, J.: The 3d point spread function can be autoencoded in latent space. bioRxiv (2024). https://doi.org/10.1101/2024.10.14.618231, https://www.biorxiv.org/content/early/2024/10/15/2024.10.14.618231
47. Wu, T., Lu, P., Rahman, M.A., Li, X., Lew, M.D.: Deep-smolm: deep learning resolves the 3d orientations and 2d positions of overlapping single molecules with optimal nanoscale resolution. Opt. Express **30**(20), 36761–36780 (2022). https://doi.org/10.1364/OE.480214, https://opg.optica.org/oe/abstract.cfm?uri=oe-30-20-36761

48. Zhang, P., Liu, S., Chaurasia, A., Ma, D., Mlodzianoski, M.J., Culurciello, E., Huang, F.: Analyzing complex single-molecule emission patterns with deep learning. Nature Methods **15**(11), 913–916 (Nov 2018). https://doi.org/10.1038/s41592-018-0153-5, https://www.nature.com/articles/s41592-018-0153-5
49. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)