OPTIMAL TRANSPORT UNLOCKS END-TO-END LEARNING FOR SINGLE-MOLECULE LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-molecule localization microscopy (SMLM) allows reconstructing biology-relevant structures beyond the diffraction limit by detecting and localizing individual fluorophores — fluorescent molecules stained onto the observed specimen — over time to reconstruct super-resolved images. Currently, efficient SMLM requires non-overlapping emitting fluorophores, leading to long acquisition times that hinders live-cell imaging. Recent deep-learning approaches can handle denser emissions, but they rely on variants of non-maximum suppression (NMS) layers, which are unfortunately non-differentiable and may discard true positives with their local fusion strategy. In this presentation, we reformulate the SMLM training objective as a set-matching problem, deriving an optimal-transport loss that eliminates the need for NMS during inference and enables end-to-end training. Additionally, we propose an iterative neural network that integrates knowledge of the microscope's optical system inside our model. Experiments on synthetic benchmarks and real biological data show that both our new loss function and architecture surpass the state of the art at moderate and high emitter densities. Code is available at `anonymized_url`.

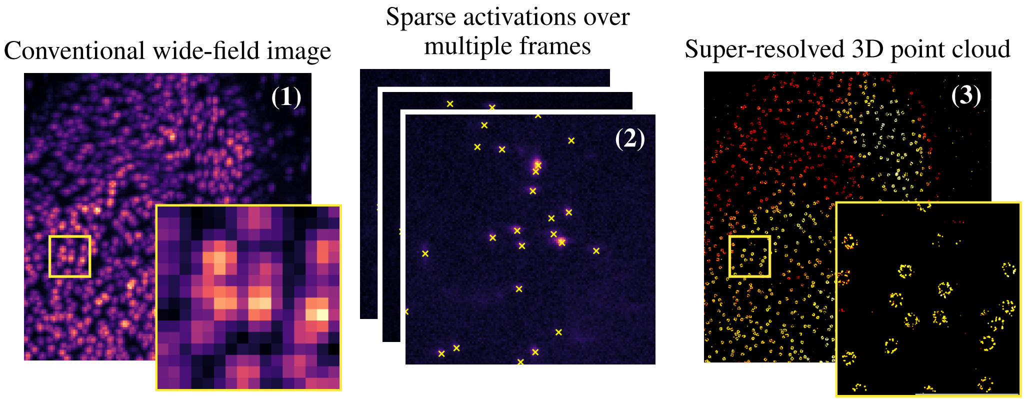


Figure 1: Illustration of the SMLM principle using our method. Data (Fei et al., 2025) show Nup96 in human bone cancer (U2-OS) cells. (1) A conventional wide-field microscope would record an image with limit resolution of $\sim 200\,\mathrm{nm}$. (2) Instead, SMLM captures many frames where only a sparse subset of fluorophores actively emit in each one. These can be detected and localized with sub-pixel precision. (3) The union of all detections is rendered as a 3D point cloud (color encodes depth), producing a super-resolved representation of the specimen.

1 INTRODUCTION

Fluorescence microscopy remains a cornerstone tool of biological research, recording photon emissions from fluorophores (fluorescent molecules) stained onto a specimen to characterized its structure. However, light diffraction restricts the final image resolution to approximately half the wavelength of light, preventing analysis of structures or organelles feature smaller than $\sim 200\,\mathrm{nm}$ in practice (Mccutchen, 1967; Schermelleh et al., 2019).

Multiple experimental techniques have been developed to surpass the diffraction limit (Hell & Wichmann, 1994; Gustafsson, 2000; Dertinger et al., 2009; Laine et al., 2023), collectively described as *super-resolution microscopy* methods. Among them, *single molecule localization microscopy* (SMLM) takes advantage of the stochastic flickering of fluorophores over a long sequence of images (Betzig et al., 1991). Compared to conventional fluorescence microscopy, the laser power is tuned to achieve a low density of simultaneously active fluorophores such that, with high probability, no two emitters occupy the same diffraction-limited area at the same time (Lelek et al., 2021). As modelling light propagation for point emitters in the microscope can be approximated, each emitter pixel pattern can be deconvolved into point localisation with positionning error massively inferior to the diffraction limit. Accumulating detections across all frames yields a point cloud representation of the underlying specimen (Rust et al., 2006), effectively achieving super-resolution. Figure 1 illustrates this method. For additional details, on both experimental methods and deconvolution approaches see the review by Lelek et al. (2021).

However, the low-density constraint inherent to this approach limits the number of active fluorophores that can be captured in a single frame, requiring thousands of frames to reconstruct a complete specimen, which hinders live-cell imaging and the observation of dynamic processes (Heilemann et al., 2008). Consequently, high-density setups are desirable, but overlapping fluorophores within the same diffraction-limited area usually led to uncertainties in the number of fluorophores and reduced spatial resolution, yielding deteriorated reconstruction.

Deep learning methods have shown success at handling higher densities. Top methods (Speiser et al., 2021; Fei et al., 2025) predict a detection map trained with pixel-wise objectives, and decide at inference whether a candidate exists or not by binarizing their map using a variant of non-maximum suppression (NMS) (Girshick et al., 2014). This NMS-variant uses two thresholds to (i) suppress spurious local maxima while (ii) not merging nearby emitters. We see three main issues with this framework. (1) These pixel-wise loss functions do not account for multiple emitters within the same pixel. (2) Objectives (i) and (ii) are inherently in conflict, and this issue only worsens as density increases, where the probability of multiple emitters activating simultaneously at sub-pixel distance rises. (3) The precision–recall tradeoff is difficult to tune due to the two required hand-set thresholds. Figure 5 in the Appendix illustrates problems (1) and (2).

In this paper, we frame the SMLM training objective based on one-to-one matching between predicted and true emitters, using a new loss function constructed from optimal transport (Peyré et al., 2019), and solve the decision problem at inference with a simple individual one-threshold filtering. These changes solve problem (1) by removing pixel-wise assignments in the training objective, problem (2) by removing decision pipelines based on spatial proximity like NMS, and problem (3) by using a single threshold during filtering, which directly controls the precision-recall tradeoff. Furthermore, NMS non-differentiability prevents the model from optimizing for it: discarding it allows us to benefit from the flexibility of deep neural networks at the final model layer, unlocking end-to-end learning. Additionally, inspired by the success of iterative refinement networks for optical flow estimation (Teed & Deng, 2020; Hur & Roth, 2019), we propose a novel iterative neural network architecture that leverages a reconstruction of the expected frame given the current estimated set of fluorophores, introducing knowledge of the microscope's optic system into the model. We demonstrate that both our loss function and architecture choices improve the state of the art at both low- and high-density regimes on synthetic benchmarks and real data.

2 RELATED WORK

Single-molecule localization microscopy. SMLM has been enabled by the development of photoactivable and photoswitchable fluorophores, which allows individual molecules to emit efficiently and in a controllable manner sufficient amount of photons to be individually located (Betzig et al., 2006; Hess et al., 2006). Early tools perform detection by locating local maxima and localizing with a Gaussian estimation of the point-spread function (PSF) (Patterson et al., 2010; Rust et al., 2006), assuming simplified light propagation in the microscope. Shortly after, the introduction of asymmetry along the z-axis of the PSF enabled 3D localisation (Huang et al., 2008); the most common setup is to introduce astigmatism in the microscope optics, which we employ in this work. To improve localisation accuracy, PSF models has transitioned from being only theory-derived to experimentally augmented, in order to incorporate effects of real light propagation in the microscope and unmodelled effect of light propagation in the cell (Babcock et al., 2012). This usually requires a pre-calibration step using specially designed fluorescent beads, which are imaged to capture how a single point of light appears at different locations. Note that while this calibration phase can be resource- and time-consuming, recent works propose live estimation of the PSF (Liu et al., 2024). 3D-DAOSTORM (Babcock et al., 2012) is a widely used classical method that uses experimentally derived PSFs, and we use it as a baseline in our comparisons.

SMLM can deliver high resolution ($10-20\,\mathrm{nm}$) in optimised experimental conditions and low phototoxicity at the cost of slow acquisition speed (Lelek et al., 2021). Methods such as SIM (Gustafsson, 2000), SOFI (Dertinger et al., 2009), and eSRRF (Laine et al., 2023) trade speed for reduced resolution, while STED (Hell & Wichmann, 1994) offers faster imaging at the cost of higher phototoxicity. MINFLUX (Balzarotti et al., 2017) is a promising novel technique with comparable resolution to SMLM, but suffers from a small field of view and requires ultra-stable microscopes (Scheiderer et al., 2025). Therefore, SMLM offers an attractive middle ground for biologists among all super-resolution methods, which explains why enhancing performance for high-density setups is of major interest (Lelek et al., 2021).

Deep-learning methods have been widely applied to fluorescence microscopy (Nehme et al., 2018; Ouyang et al., 2018; Boyd et al., 2018; Cachia et al., 2023; Li et al., 2023; Mentagui et al., 2024; Fei et al., 2025). Among these, DeepLoco (Boyd et al., 2018) introduces a set formulation like ours with a loss function based on maximum mean discrepancy (Gretton et al., 2012). DECODE (Speiser et al., 2021) combines pixel-wise detection and Gaussian-mixture localization losses, and at the time of writing is ranked first on the EPFL SMLM challenge (Sage et al., 2019), a popular benchmark for SMLM tools. More recently, LiteLoc (Fei et al., 2025) has slightly improved on DECODE's architecture with additional technical refinements. We use those last two methods for comparison in our benchmarks.

Optimal transport for set matching. Optimal transport (Peyré et al., 2019; Villani, 2021) has become a popular tool for set matching by deep learning. Recent works in object detection (Carion et al., 2020; Zhu et al., 2021; Zhang et al., 2023; Li et al., 2022) have demonstrated success in predicting sets of variable and unknown size using bipartite matching loss functions, while other modern works have employed entropic regularization (Cuturi, 2013) to achieve fully differentiable pipelines (Zareapoor et al., 2024). By framing SMLM as a set matching problem, we draw a direct connection to this line of work — substituting objects for fluorophores — enabling the design of an end-to-end training procedure.

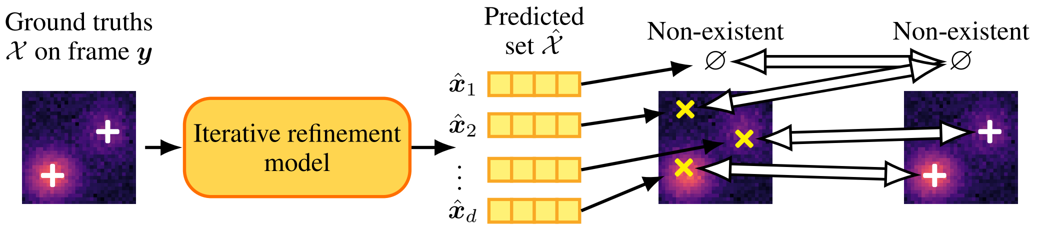
Iterative refinement network. Iterative refinement within neural networks has proven effective for tasks that benefit from sequential solution improvement (Carreira et al., 2016; Yu et al., 2023). In computer vision, Putzky & Welling (2017) have applied this approach to inverse problems such as image denoising, super-resolution, and inpainting, while Hur & Roth (2019) proposed iterative optical-flow refinement using a feedback loop with a rewarping operator. As the physics of SMLM is well understood (Etheridge et al., 2022), we show that an accurate simulator of the microscope's physics can provide similar visual feedback, enabling progressive refinement of the solution.

3 METHOD

3.1 PROBLEM FORMULATION

In this section, we first introduce the image formation model for SMLM and formulate the corresponding inverse problem as a set matching task. We then present a differentiable loss function and an iterative refinement architecture that explicitly leverages the image formation process.

Image formation model An *activation* is defined as an emission event from a fluorophore within a given frame (a single fluorophore may produce several activations accross multiple frames). Throughout this work, an activation is represented by a 4D vector $\boldsymbol{x} = (x, y, z, n)$, where (x, y) denote the 2D coordinates in the camera frame (with the origin at the top-left corner), z represents the axial coordinate relative to the focal plane, and n is the photon count. Given N activations within a frame, we denote the complete set as $\mathcal{X} = \{\boldsymbol{x}_i\}_{1\leq i\leq N}$.



(1) Model maps input frame $\boldsymbol{y}$ to candidate activations set $\hat{\mathcal{X}}$.

(2) Loss function is the optimal transport cost between $\hat{\mathcal{X}}$ and $\mathcal{X}$.

Figure 2: Illustration of our loss function for end-to-end training. (1) Given a simulated image and its ground truth activations (see Section 3.1), our model (see Section 3.3) predicts d candidate activations, each with a detection score quantifying the plausibility of its existence. (2) We solve a regularized optimal transport problem — conceptually similar to a bi-matching between ground truths and predictions — over a cost involving both localisation and detection tasks. Our loss function is the optimal cost yielded by this solution.

Diffraction within the microscope's optical system is modeled by a convolution where the kernel is called the *point spread function* (PSF) (Rossmann, 1969). It can be thought of as the image of a single point source. We represent the PSF as a function $\mathbf{P} : \mathbb{R}^3 \longmapsto \mathbb{R}^{H \times W}$, that outputs the normalized $H \times W$ image resulting from the diffraction of a single point source given its 3D coordinates. To ensure photon count independence, the output image is normalized to sum to unity in the focal plane, i.e. for $z = 0$. Given the set of activations $\mathcal{X}$ in a frame, the observed $H \times W$ image, denoted by $\mathbf{H}(\mathcal{X})$, is formed as a weighted sum of PSFs, where weights are the photon count n for each activation:

$$\mathbf{H}(\mathcal{X}) = \sum_{(x,y,z,n) \in \mathcal{X}} n\mathbf{P}(x, y, z). \tag{1}$$

The dependence of the PSF on depth z enables 3D localization of activations from the observed image, see (Ovesnỳ et al., 2014). Following Babcock & Zhuang (2017), we assume that the PSF is pre-calibrated on synthetic fluorescent beads and implemented as a collection of 3D splines. This approach is a standard tool in SMLM used in many works (Ries, 2020; Li et al., 2020; Speiser et al., 2021; Etheridge et al., 2022); see Babcock & Zhuang (2017) for further details.

Noise model. We adopt the noise model of Sage et al. (2019), which combines shot noise (stemming from the photon detector, and modeled by a Poisson distribution), amplification noise (which allow to increase the number of generated electron per photon, and modeled by a Gamma distribution only for EM-CCD camera) and readout noise (modeled by a normal distribution). In-depth description of all camera parameters is available in Appendix A.1. The noise for each camera sensor being independent and identically distributed (Fazel & Wester, 2022), it is applied independently to all pixels of $\mathbf{H}(\mathcal{X})$. Then, we denote by $\mathbb{P}$ the distribution of images $\boldsymbol{y}$ produced by a set $\mathcal{X}$ of fluorophores under the noise model such that

$$\boldsymbol{y} \sim \mathbb{P}\left(\mathcal{X}\right). \tag{2}$$

Risk minimization formulation for set matching. As no ground truth is available in most scientific imaging applications, supervised-learning models for SMLM have to be trained with a simulator, which is able to generate realistic $\boldsymbol{y}$ from $\mathbb{P}(\mathcal{X})$ given sets $\mathcal{X}$ of activations from a distribution $\mathcal{D}$. Our approach consists of training a neural network f_θ which directly predicts a set of activations given an observation $\boldsymbol{y}$, by minimizing the risk

$$\theta^* = \operatorname*{arg\,min}_{\theta \in \Theta} \mathbb{E}_{\mathcal{X} \sim \mathcal{D}, \boldsymbol{y} \sim \mathbb{P}(\mathcal{X})} \left[\mathcal{L}(f_\theta(\boldsymbol{y}), \mathcal{X})\right]. \tag{3}$$

Such a formulation raises two major challenges: we need to design a differentiable loss function $\mathcal{L}$ and an architecture f_θ that are appropriate to the context of SMLM.

3.2 OPTIMAL TRANSPORT LOSS FUNCTION

We argue that framing SMLM as a supervised-learning problem leads to a set matching formulation, for which optimal transport theory is a natural fit. To the best of our knowledge, however, this framework has not yet been applied to SMLM. Figure 2 provides an overview of our method.

Let $\mathcal{X} = \{\boldsymbol{x}_i\}_{1\leq i\leq N}$ be the ground truth set of activations. The size of this set, N, is unknown and varies between frames, but it can be bounded by the physics of the fluorophore and the experimental protocol. We simulate an acquisition $\boldsymbol{y} \sim \mathbb{P}(\mathcal{X})$ and aim to retrieve $\mathcal{X}$ from $\boldsymbol{y}$.

Given $\boldsymbol{y}$, our neural network f_θ outputs a set of candidate activations $\hat{\mathcal{X}} = \{\hat{\boldsymbol{x}}_i\}_{1\leq i\leq d}$ of fixed size d, each associated with a detection score $\hat{s}_i$ in $(0,1)$ gathered in a set $\hat{\mathcal{S}} = \{\hat{s}_i\}_{1\leq i\leq d}$. The network architecture is detailed in Section 3.3. The number of candidates d is fixed by the architecture, and defines the maximum possible number of detectable activations, see Appendix A.2 for further analyze on the impact of this parameter.

We first define $\boldsymbol{L}$, a squared cost matrix of size $d \times d$, whose components are:

$$\forall 1 \leq i,j \leq d, L_{i,j} = \begin{cases} (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_j) + \log\det(\boldsymbol{\Sigma}) & \text{if } j \leq N, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_x^2, \sigma_y^2, \sigma_z^2, \sigma_n^2)$ is a diagonal weighting matrix. Quadratic costs are a natural and principled choice for regression tasks. Extending this formulation to the negative log-likelihood of a multivariate normal distribution allows to learn $\boldsymbol{\Sigma}$ end-to-end, which can be viewed as an automatic weighting strategy that balances the difficulty of predicting each dimension, similar to the homoscedastic uncertainty weighting method proposed by Kendall et al. (2018). Experimentally, we have found σ_z^2 to be $\sim 2\times$ larger than σ_x^2 and σ_y^2 after training, which is consistent with the optical theory of confocal microscopy (Pawley, 2006).

Similarly, we define $\boldsymbol{D}$, another $d \times d$ cost matrix whose components are:

$$\forall 1 \leq i,j \leq d, D_{i,j} = \begin{cases} -\log(s_i) & \text{if } j \leq N, \\ -\log(1-s_i) & \text{otherwise.} \end{cases} \tag{5}$$

The binary cross-entropy cost is a natural choice for detection tasks. It favors a high score s_i when $\hat{\boldsymbol{x}}_i$ is paired with an element of $\mathcal{X}$ and low score otherwise, hence promoting good detection. Finally, we define the total cost matrix $\boldsymbol{C} = \boldsymbol{L} + \boldsymbol{D}$, which integrates both localization and detection tasks.

Considering the initial set matching problem, the optimal solution $(\hat{\mathcal{X}}^*, \hat{\mathcal{S}}^*)$ given a target $\mathcal{X}$ would consist of N elements, each identical to one element of $\mathcal{X}$ and with detection scores close to 1. The remaining $d - N$ elements have detection scores close to 0. Naturally, one would like to compare each candidate in $\hat{\mathcal{X}}^*$ to its nearest counterpart in $\mathcal{X}$ and minimize a loss function over these pairs. This can be achieved by solving an optimal-transport problem over $\boldsymbol{C}$ — conceptually creating a bipartite matching between the predictions and the ground truths — where the minimal cost accounts for all pairwise contributions. Therefore, we would ideally like our loss function to be the optimal-transport cost with respect to $\boldsymbol{C}$, i.e. solve:

$$\min_{\boldsymbol{\Gamma}\in\mathcal{B}} \langle \boldsymbol{\Gamma} \mid \boldsymbol{C} \rangle_{\mathcal{F}} \quad \text{where } \mathcal{B} = \left\{ \boldsymbol{\Gamma} \in \mathbb{R}_+^{d\times d} \mid \boldsymbol{\Gamma}\mathbf{1}_d = \boldsymbol{\Gamma}^\top \mathbf{1}_d = \mathbf{1}_d \right\}, \tag{6}$$

and $\langle . \mid . \rangle_{\mathcal{F}}$ is the Frobenius inner product. However, while the Hungarian algorithm can exactly solve this problem in $\mathcal{O}(d^3)$ (Kuhn, 1955), its algorithmic step is non-differentiable, which prevents end-to-end learning. We circumvent this issue by finding $\boldsymbol{\Gamma}$ through the entropy-regularized optimal transport problem, see (Cuturi, 2013), and therefore define our loss function as follows:

$$\mathcal{L}(\hat{\mathcal{X}}, \hat{\mathcal{S}}, \mathcal{X}) = \langle \boldsymbol{\Gamma}^* \mid \boldsymbol{C} \rangle_{\mathcal{F}}, \quad \text{where } \boldsymbol{\Gamma}^* = \underset{\boldsymbol{\Gamma}\in\mathcal{B}}{\arg\min} \langle \boldsymbol{\Gamma} \mid \boldsymbol{C} \rangle_{\mathcal{F}} - \epsilon H(\boldsymbol{\Gamma}), \tag{7}$$

H is the Shannon entropy and ϵ the entropic regularization parameter. A good approximation of $\boldsymbol{\Gamma}^*$ in Eq. (7) can be found efficiently with a few iterations of the Sinkhorn algorithm, whose steps are differentiable with respect to the elements of $\boldsymbol{C}$, enabling its use within a deep learning framework, see (Genevay et al., 2018; Mialon et al., 2021). A more detailed analysis of the impact of ϵ and a comparison with the Hungarian algorithm are available in Appendix A.4.

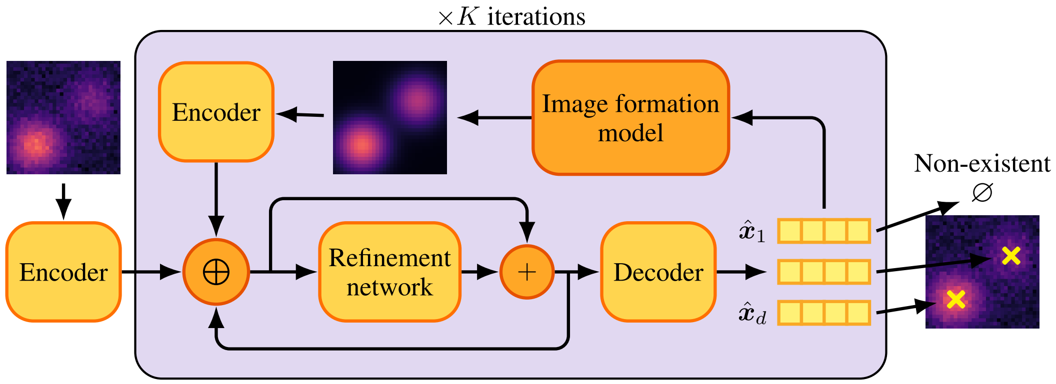


Figure 3: Illustration of our iterative refinement model. Within a classic encoder-decoder architecture, we leverage prior knowledge about the known image formation model (not learned) to simulate the expected frame given the current latent representation. This feedback is used to iteratively refine the model's inner latent representation for K steps. The encoders are identical. $+$ and $\oplus$ respectively denotes element-wise addition and concatenation.

Algorithm 1 Iterative refinement architecture

```
Require: input frame $\boldsymbol{y} \in \mathbb{R}^{H \times W}$, encoder $\boldsymbol{E}$, decoder $\boldsymbol{D}$, refinement module $\boldsymbol{R}$, camera model $\mathbb{P}(\cdot)$, number of iterations $K \in \mathbb{N}$
Ensure: final proposals $(\hat{\mathcal{X}}^{(K)}, \hat{\mathcal{S}}^{(K)})$
$\boldsymbol{z}^{(0)} \leftarrow \boldsymbol{E}(\boldsymbol{y})$    ▷ encode original frame
$(\hat{\mathcal{X}}^{(0)}, \hat{\mathcal{S}}^{(0)}) \leftarrow \boldsymbol{D}(\boldsymbol{z}^{(0)})$    ▷ decode initial proposals
for $k = 0$ to $K - 1$ do
    $\hat{\boldsymbol{y}}^{(k)} \leftarrow \mathbb{E}[\mathbb{P}(\hat{\mathcal{X}}^{(k)}, \hat{\mathcal{S}}^{(k)})]$    ▷ simulate reconstruction from current proposals
    $\hat{\boldsymbol{z}}^{(k)} \leftarrow \boldsymbol{E}(\hat{\boldsymbol{y}}^{(k)})$    ▷ encode reconstruction
    $\boldsymbol{z}^{(k+1)} \leftarrow \boldsymbol{z}^{(k)} + \boldsymbol{R}\big(\boldsymbol{z}^{(k)},\ \hat{\boldsymbol{z}}^{(k)},\ \boldsymbol{z}^{(0)}\big)$    ▷ refine latent iteratively
    $(\hat{\mathcal{X}}^{(k+1)}, \hat{\mathcal{S}}^{(k+1)}) \leftarrow \boldsymbol{D}(\boldsymbol{z}^{(k+1)})$    ▷ decode refined proposals
end for
return $(\hat{\mathcal{X}}^{(K)}, \hat{\mathcal{S}}^{(K)})$
```

3.3 ITERATIVE REFINEMENT SCHEME

To solve Eq. (3), we investigate architectures that explicitly leverage the image formation process. To this end, we adopt an iterative architecture, an idea that has proven successful for optical flow estimation (Hur & Roth, 2019). At each iteration the network produces a set of candidate activations, turns those proposals into a simulated image, which is then used as feedback to refine the next proposals. This iterative method is illustrated in Figure 3.

Concretely, let $\boldsymbol{y}$ be the input frame of size $H \times W$. An encoder $\boldsymbol{E} : \mathbb{R}^{H \times W} \longmapsto \mathbb{R}^{C \times H \times W}$ maps $\boldsymbol{y}$ to a latent representation $\boldsymbol{z}^{(0)}$, where C is a hyperparameter controlling the dimension of the latent space. A decoder $\boldsymbol{D}$ then maps latent variables to a set of candidate activations $\hat{\mathcal{X}} = \{\hat{\boldsymbol{x}}_i\}_{1 \leq i \leq d}$ and a corresponding set of detection scores $\hat{\mathcal{S}} = \{\hat{s}_i\}_{1 \leq i \leq d}$.

Given $(\hat{\mathcal{X}}, \hat{\mathcal{S}})$, we compute a reconstructed frame $\hat{\boldsymbol{y}} = \mathbb{E}[\hat{\mathbf{y}} | \hat{\mathcal{X}}, \hat{\mathcal{S}}]$. $\hat{\boldsymbol{y}}$ is the expected image produced by the current proposal set, and thus provides a visual summary of what the model's output currently explain in the SMLM frame. Comparing the reconstructed image $\hat{\boldsymbol{y}}$ to the original frame $\boldsymbol{y}$ supplies informative feedback, which helps the model correct errors and refine the candidates set over iterations.

Concretely, we define an iterative refinement operator $\boldsymbol{R} : \mathbb{R}^{3 \times C \times H \times W} \longmapsto \mathbb{R}^{C \times H \times W}$ which produces a residual update of the latent representation given it's current estimate, the representation of the simulated frame, and the encoded original frame. Algorithm 1 shows how this proposal is updated successively over K steps.

During training, the final decoded output $(\hat{\mathcal{X}}^{(K)}, \hat{\mathcal{S}}^{(K)})$ is used as input for our loss function, see Section 3.2. Details about the decoder architecture and the computation of $\hat{\boldsymbol{y}}$ given $\hat{\mathcal{X}}$ and $\hat{\mathcal{S}}$ can be found in Appendix A.2.

4 EXPERIMENTS

Implementation and training details. We construct a synthetic target activations set $\mathcal{X}$ from $\mathcal{D}$ by uniformly sampling between 10 and 30 activations per frame, assigning each activation independent coordinates that are uniformly distributed across all dimensions. This guarantees that the network cannot learn any specific prior about the activation distribution.

Following (Speiser et al., 2021), we augment $\boldsymbol{y}$ by the previous and the next frame into a tensor $\bar{\boldsymbol{y}}$ of size $3 \times H \times W$. Including these provides additional context, without introducing a too complex prior about the physics of the fluorophore, about the frame of interest, and yield improved performance. We also randomly scaled each camera parameters by a coefficient e^{ρ}, where $\rho \sim \mathcal{N}(0, 0.03)$: this data-augmentation ”trick” increases the model robustness to experimental complexities in fully controlling and characterising experimental parameters.

The encoder $\boldsymbol{E}$ is a two-layer U-Net (Ronneberger et al., 2015) with SiLU activation functions (Hendrycks & Gimpel, 2016), LayerNorm normalization layers (Ba et al., 2016), and an internal channel width of 48. It maps the input image to a latent image with $C = 96$ channels. The iterative refinement stage is implemented with a similar two-layer U-Net. For the decoder $\boldsymbol{D}$, instead of adopting a vision transformer (Dosovitskiy et al., 2021) as done in DETR-like object detectors (Carion et al., 2020), we found that a light CNN yields better performance, see Appendix A.2 for additional details and benchmarks. The resulting network predicts $d = HW/4$ candidates and contains ~ 3 millions learnable parameters. Empirically, performance improvement stops after three or more refinement iterations; we thus use $K = 2$ in our experiments.

For training, we use AdamW (Loshchilov & Hutter, 2019) for 100,000 steps with a batch size of 128 on a NVIDIA-H100 gpu, taking approximately 20h. The iterative architecture incurs a higher computational burden than single-pass models like DECODE or LiteLoc; further details about our model computational footprint for training and inference is available in Appendix A.3.

Inference and detection-localisation trade-off. During inference, we only retain candidate activations from $\hat{\mathcal{X}}^{(K)}$ whose associated detection scores in $\hat{\mathcal{S}}^{(K)}$ exceed a user-defined threshold τ in $[0, 1]$. This simple filtering strategy makes τ an easy lever to control the precision-recall trade-off: $\tau = 0$ keeps every candidate while $\tau = 1$ discards all. By contrast, DECODE and LiteLoc use a two-threshold variant of a NMS strategy (Speiser et al., 2021; Fei et al., 2025) that may be harder to tune and harder to adapt to changing dynamics during the recording.

As the default value for τ, we propose using the one maximizing the $\mathrm{E}_{3\mathrm{D}}$ metric (defined in Section 4) on a separate synthetic dataset generated by our simulator. This procedure yields a threshold that achieves the same detection–localisation trade-off as the one proposed by the EPFL challenge (Sage et al., 2019). To ensure that this choice does not bias the results of Table 1 in our favor, we performed the same optimization for DECODE's and LiteLoc's NMS parameters. When performing experimental data analysis hyperparameters can be tuned to match experimental settings and ensure consistent precision for long time recordings.

Synthetic data. Because no ground-truth annotations exist for real SMLM acquisitions, we performe the initial evaluation on the open synthetic datasets provided by Sage et al. (2019) on the 2016 EPFL challenge, and adopte their set of metrics.

To evaluate candidate activations in a frame, we first solve a Hungarian assignment between ground-truths and predicted activations. A prediction is considered a *true positive* (TP) if it lies within $\pm 250\,\mathrm{nm}$ in both x and y directions, and $\pm 500\,\mathrm{nm}$ in z relative to its matched ground-truth (both thresholds come from the EPFL challenge). Otherwise, predictions (resp. ground truths) are labeled as *false positives* (resp. *false negatives*). Detection performance is quantified by computing *precision*, *recall* and *Jaccard Index* (*area under the curve* is not commonly employed in this field). Localisation performance is evaluated by computing the *root-mean-square error* (RMSE) for TPs, for the lateral plan ($\mathrm{RMSE}_{\mathrm{lat}}$), the axial dimension ($\mathrm{RMSE}_{\mathrm{ax}}$), and all three dimensions together

Density	SNR	Method	Precision ↑	Recall ↑	Jaccard ↑	RMSE$_{\text{lat}}$ ↓	RMSE$_{\text{ax}}$ ↓	E$_{\text{3D}}$ ↑
0.2	High	3D-DAOSTORM	0.964	0.919	0.914	11.9	16.9	0.821
		DECODE	0.961 ± 0.003	**0.998 ± 0.001**	0.959 ± 0.003	8.8 ± 0.1	10.7 ± 0.1	0.895 ± 0.003
		LiteLoc	0.996 ± 0.002	0.987 ± 0.001	**0.983 ± 0.001**	9.0 ± 0.1	11.7 ± 0.1	0.912 ± 0.001
		Ours	**0.998 ± 0.002**	0.978 ± 0.016	0.980 ± 0.010	**7.5 ± 0.4**	**10.0 ± 3.5**	**0.920 ± 0.007**
	Low	3D-DAOSTORM	0.978	0.835	0.833	19.3	29.8	0.685
		DECODE	0.918 ± 0.002	**0.978 ± 0.001**	0.903 ± 0.002	20.5 ± 0.1	26.2 ± 0.1	0.757 ± 0.001
		LiteLoc	**0.995 ± 0.001**	0.939 ± 0.001	0.934 ± 0.001	**17.0 ± 0.1**	25.0 ± 0.4	0.798 ± 0.001
		Ours	0.985 ± 0.001	0.961 ± 0.001	**0.947 ± 0.001**	18.8 ± 0.2	**24.5 ± 0.1**	**0.802 ± 0.002**
2.0	High	3D-DAOSTORM	0.914	0.678	0.643	56.8	76.6	0.373
		DECODE	0.923 ± 0.003	**0.946 ± 0.002**	0.876 ± 0.004	32.2 ± 0.3	33.0 ± 0.4	0.706 ± 0.004
		LiteLoc	**0.993 ± 0.001**	0.863 ± 0.002	0.858 ± 0.001	30.7 ± 0.2	36.0 ± 0.3	0.699 ± 0.001
		Ours	0.992 ± 0.002	0.895 ± 0.011	**0.883 ± 0.007**	**24.8 ± 0.6**	**28.4 ± 0.5**	**0.750 ± 0.004**
	Low	3D-DAOSTORM	0.914	0.496	0.475	74.4	120.0	0.116
		DECODE	0.859 ± 0.034	**0.874 ± 0.006**	0.756 ± 0.027	56.4 ± 0.3	65.3 ± 0.4	0.468 ± 0.008
		LiteLoc	**0.992 ± 0.001**	0.729 ± 0.002	0.725 ± 0.001	**46.2 ± 0.4**	63.8 ± 0.2	0.500 ± 0.002
		Ours	0.973 ± 0.003	0.812 ± 0.007	**0.794 ± 0.005**	48.4 ± 0.6	**59.5 ± 0.4**	**0.536 ± 0.003**
8.0*	High	3D-DAOSTORM	0.910	0.392	0.379	83.3	133.9	0.009
		DECODE	0.973 ± 0.001	**0.627 ± 0.002**	**0.617 ± 0.002**	59.58 ± 0.02	71.5 ± 0.5	0.371 ± 0.006
		LiteLoc	0.988 ± 0.001	0.557 ± 0.003	0.553 ± 0.003	60.4 ± 0.2	80.5 ± 0.5	0.319 ± 0.004
		Ours	**0.989 ± 0.001**	0.578 ± 0.008	0.574 ± 0.008	**52.5 ± 0.4**	**64.3 ± 0.7**	**0.384 ± 0.003**
	Low	3D-DAOSTORM	0.908	0.211	0.217	92.6	175.8	−0.216
		DECODE	0.93 ± 0.03	**0.415 ± 0.005**	**0.402 ± 0.009**	80.25 ± 0.04	105.2 ± 1.0	0.090 ± 0.007
		LiteLoc	0.986 ± 0.001	0.339 ± 0.002	0.338 ± 0.002	76.1 ± 0.1	110.1 ± 0.8	0.055 ± 0.002
		Ours	**0.983 ± 0.002**	0.376 ± 0.008	0.374 ± 0.008	**74.3 ± 0.5**	**99.4 ± 1.1**	**0.103 ± 0.002**

Table 1: Comparative evaluation of SMLM algorithms on the EPFL 2016 challenge datasets and metrics. Densities are expressed in activations/µm/frame. For each method, means and standard deviations are estimated over four independent training seeds (3D-DAOSTORM is deterministic). *The EPFL 2016 challenge does not include a dataset with a density of 8.0; see the main text for details about its creation process.

(RMSE$_{\text{vol}}$). A global performance metric called *3D efficiency* (E$_{\text{3D}}$) is then defined as:

$$\mathrm{E}_{\text{3D}} = \frac{\mathrm{E}_{\text{ax}} + \mathrm{E}_{\text{lat}}}{2} \quad \text{where} \quad \begin{cases} \mathrm{E}_{\text{lat}} = & 1 - \sqrt{(1 - \text{Jaccard})^2 + \alpha_{\text{lat}}^2 \mathrm{RMSE}_{\text{lat}}^2}, \\ \mathrm{E}_{\text{ax}} = & 1 - \sqrt{(1 - \text{Jaccard})^2 + \alpha_{\text{ax}}^2 \mathrm{RMSE}_{\text{ax}}^2}, \end{cases} \tag{8}$$

$\alpha_{\text{lat}} = 1.0\,\text{nm}^{-1}$ and $\alpha_{\text{ax}} = 0.5\,\text{nm}^{-1}$, following definitions of the EPFL challenge. All metrics are computed frame by frame and averaged.

Our benchmark includes 3D-DAOSTORM (Babcock et al., 2012), DECODE (Speiser et al., 2021) and LiteLoc (Fei et al., 2025). All algorithms are evaluated on the open-access EPFL 2016 challenge datasets (Sage et al., 2019), all with astigmatism PSFs. To assess performance in a very high-density regime, we have synthesized a density-8.0 benchmark by temporally binning groups of 4 frames in the original density-2.0 sequences. For each newly binned frame, we have re-sampled camera noise using the known camera parameters. This extra step prevents an artificial SNR improvement caused by the frame-averaging process.

Results are reported in Table 1. We observe that while our approach yields lower recall than the other methods, it preserves excellent precision and almost always achieves the lowest RMSE in all spatial dimensions. Most notably, it also outperforms all competitors on the E$_{\text{3D}}$ metric for all densities and SNRs, establishing itself as the most balanced method with respect to this criterion.

Real data. We have evaluated our method on three publicly available datasets, all of which provide beads for calibrating their astigmatic PSFs. The Tubulin and NPC-Nup107 datasets from Li et al. (2018) depict, respectively, the microtubule network and nuclear pore complexes in U2OS cells. The NPC-Nup96 dataset from Fei et al. (2025) also features nuclear pore complexes in the same cell line. All datasets were acquired with conventional SMLM activation densities; therefore, to test our method's robustness at higher densities, we applied 16-frame temporal binning to Tubulin and NPC-Nup107 and 32-frame binning to NPC-Nup96. We refer to the temporally-binned versions as T16-Tubulin, T16-NPC-Nup107, and T32-NPC-Nup96. Note that this approach is an imperfect proxy for truly high-density imaging, as it improves the signal-to-noise ratio via noise averaging.

Figure 4 compares 3D SMLM reconstructions, rendered with SMAP (Ries, 2020), for 3D-DAOSTORM (Babcock et al., 2012), LiteLoc (Fei et al., 2025) and our method. We chose to include

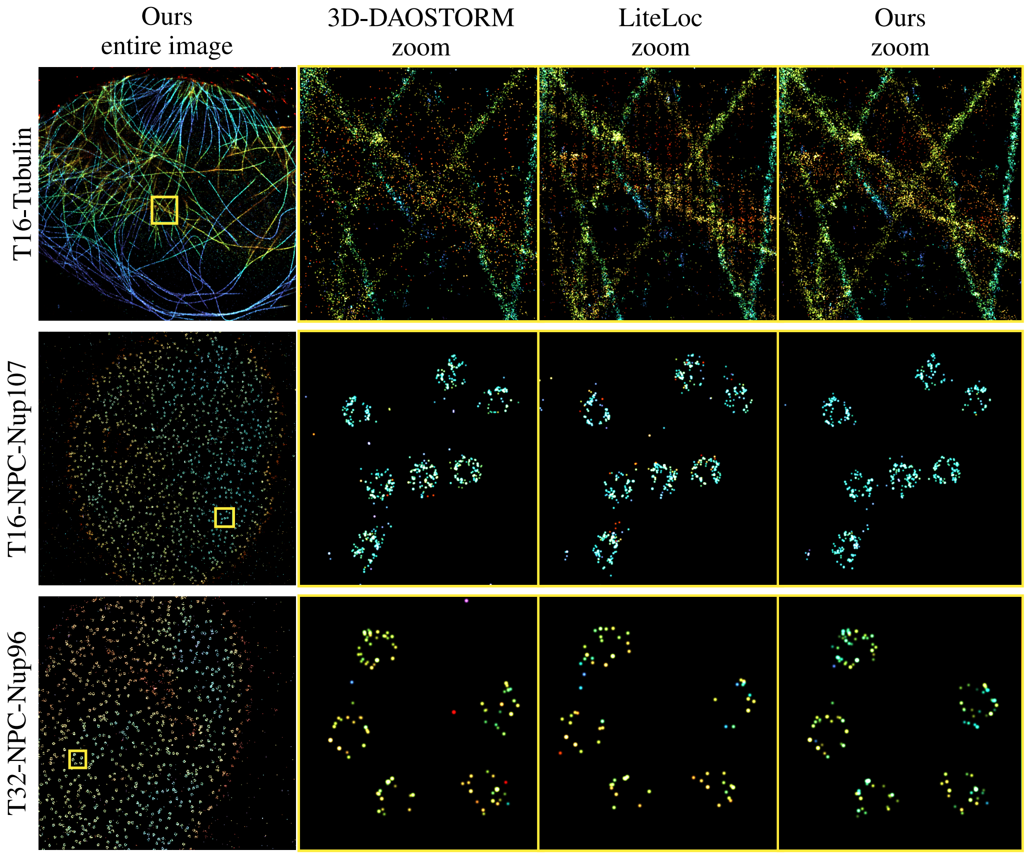


Figure 4: Qualitative comparison of SMLM methods on real data. Although ground truths are unavailable, results show that our approach yields fewer grid-reconstruction artifacts (line 1), improved depth estimation consistency (line 2), and more accurate nuclear pore complex reconstruction (line 3). Refer to the main text for a more thorough discussion.

LiteLoc over DECODE because the former delivers comparable or slightly better performance. In addition, the authors of DECODE note that their method may benefit from an extra filtering step applied to the predicted uncertainties associated with each activation. However, this post-processing step requires selecting additional arbitrary thresholds that are difficult to tune, making it challenging to perform a fair and objective comparison with other methods.

On the T16-Tubulin dataset, our algorithm yields a higher-fidelity reconstruction than 3D-DAOSTORM and eliminates the artefacts that appear with LiteLoc. On the T16-NPC-Nup107 dataset, all methods recover comparable structures; however, our method delivers more consistent depth estimates (as indicated by colors), whereas 3D-DAOSTORM and LiteLoc exhibit spatially varying detections. On the T32-NPC-Nup96 dataset, our approach reconstructed NPC's structures with clear greater fidelity than LiteLoc and 3D-DAOSTORM.

Quantitatively, the absence of ground-truth data prevents the use of the metrics introduced in Section 4. To evaluate the resolution and fidelity of a reconstructed super-resolution image, we adopted two widely used metrics: *Fourier ring correlation* (FRC) (Banterle et al., 2013) and the *resolution-scaled Pearson's coefficient* (RSP) Culley et al. (2018). FRC reconstructs two super-resolution images by splitting localisations into two subsets, computing their Fourier transforms, and then measuring the correlation of their spatial frequency signals against each other. The resulting curve provides an estimate of the spatial frequency at which signal can no longer be distinguished from noise (Banterle et al., 2013). RSP is defined as the Pearson correlation coefficient between the reconstructed super-resolution image and a reference image, typically the mean of all raw wide-field frames. Values close to one indicate strong agreement between the reconstruction and the reference.

Dataset	Bin size	Method	FRC (nm) ↓	RSP ↑
Tubulin (Li et al., 2018)	×1	LiteLoc	**29.7 ± 0.3**	**0.708**
		Ours	31.9 ± 0.2	0.692
	×16	LiteLoc	63.0 ± 0.8	0.649
		Ours	**58.1 ± 1.1**	**0.672**
NPC-Nup107 (Li et al., 2018)	×1	LiteLoc	19.3 ± 0.3	**0.696**
		Ours	**18.8 ± 0.4**	0.686
	×16	LiteLoc	25.9 ± 0.3	0.682
		Ours	**22.1 ± 0.1**	**0.684**
NPC-Nup96 (Fei et al., 2025)	×1	LiteLoc	**29.8 ± 0.1**	**0.713**
		Ours	31.7 ± 0.1	0.693
	×32	LiteLoc	71.5 ± 0.5	0.671
		Ours	**44.2 ± 0.4**	**0.689**

Table 2: Quantitative results on real datasets. We temporally binned them to simulate very high-density setups. Our method consistently scored first in those denser regimes.

Iterative arch.	OT loss func.	Jaccard ↑	RMSE_{vol} ↓	E_{3D} ↑
✗	✗	0.876 ± 0.004	47.9 ± 0.5	0.705 ± 0.004
✗	✓	0.867 ± 0.004	39.6 ± 0.3	0.740 ± 0.002
✓	✗	0.854 ± 0.005	45.4 ± 0.7	0.703 ± 0.005
✓	✓	**0.883 ± 0.007**	**39.2 ± 0.5**	**0.750 ± 0.004**

Table 3: Ablation study of our different modules over EPFL synthetic with high SNR and a density of 2.0. ✗ means we used DECODE's original loss function or model architecture.

Results with these metrics on real datasets are reported in Table 2. In dense-activations regimes, our approach consistently yields lower FRC and higher RSP values than other methods, confirming the visual improvements illustrated in Figure 4.

Ablation study. We have conducted an ablation study on synthetic data to validate the effectiveness of our loss function and our iterative architecture. Results are reported in Table 3. It can be seen that the loss function drives most of the improvement, with our iterative architecture providing a modest boost. Given the additional memory and compute overhead of our architecture, a lightweight variant that retains only the optimal loss function can be considered for deployment scenarios with constrained resources.

5 DISCUSSION AND CONCLUDING REMARKS

We have presented a novel deep-learning SMLM method that surpasses existing methods in medium and high-density regimes, all without the need for handcrafted layers. By enabling faster data acquisition, our approach extend SMLM's temporal resolution, allowing more accurate observation of rapid biological processes but also stable inference precision to degrading conditions induced by evolution of recording parameters during time. Furthermore, the integration of optimal transport theory to SMLM could open a path to new localization algorithms.

The main limitation of our method is the longer training and inference times that result from its iterative design. However, training is a one-time cost per experimental setup, and inference remains fast enough ($\sim$ 200 fps on a modern GPU) to let biologists run multiple experiments sequentially with minimal delay. Another systemic limitation shared by most top-performing methods is the dependence for precise PSF calibration (Lelek et al., 2021). Future work could focus on robust methods invariant to PSF variations, pursue blind SMLM super-resolution without sacrificing precision or include PSF optimisation to the microscopy setup design during training and to adapt to celullar based optical anomalies affecting the PSF at inference time.

ETHICS STATEMENT

In this work, we explore new model architecture and loss function to improve single-molecule localization microscopy (SMLM) reconstruction. We do not anticipate ethical or societal harms: the work is computational only and all biological data used are public and were used according to their licenses. We believe that by improving SMLM reconstruction and releasing our code openly, this work can broadly benefit biological research and make advanced tools accessible to communities worldwide.

REPRODUCIBILITY STATEMENT

The project repository includes all requirements to reproduce our results. We provide the full source code and model implementation, all datasets are publicly available, training and evaluation scripts are provided with all hyperparameters set (including random seeds), and python environment specification are supplied, making all experiments from section 4 reproducible.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. 1607.06450, 2016.

Hazen Babcock, Yaron M Sigal, and Xiaowei Zhuang. A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Optical Nanoscopy*, 1:1–10, 2012.

Hazen P Babcock and Xiaowei Zhuang. Analyzing single molecule localization microscopy data using cubic splines. *Scientific Reports*, 7(552), 2017.

Francisco Balzarotti, Yvan Eilers, Klaus C Gwosch, Arvid H Gynnå, Volker Westphal, Fernando D Stefani, Johan Elf, and Stefan W Hell. Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science*, 335:606–612, 2017.

Niccolo Banterle, Khanh Huy Bui, Edward A Lemke, and Martin Beck. Fourier ring correlation as a resolution criterion for super-resolution microscopy. *Journal of Structural Biology*, 183:363–367, 2013.

Eric Betzig, JK Trautman, TD Harris, JS Weiner, and RL Kostelak. Breaking the Diffraction Barrier: Optical Microscopy on a Nanometric Scale. *Science*, 251:1468–1470, 1991.

Eric Betzig, George H Patterson, Rachid Sougrat, O Wolf Lindwasser, Scott Olenych, Juan S Bonifacino, Michael W Davidson, Jennifer Lippincott-Schwartz, and Harald F Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313:1642–1645, 2006.

Nicholas Boyd, Eric Jonas, Hazen Babcock, and Benjamin Recht. DeepLoco: Fast 3D Localization Microscopy Using Neural Networks. *BioRxiv preprint*, pp. 267096, 2018.

Mayeul Cachia, Vasiliki Stergiopoulou, Luca Calatroni, Sébastien Schaub, and Laure Blanc-Féraud. Fluorescence image deconvolution microscopy via generative adversarial learning (FluoGAN). *Inverse Problems*, 39(5), 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4733–4742, 2016.

Siân Culley, David Albrecht, Caron Jacobs, Pedro Matos Pereira, Christophe Leterrier, Jason Mercer, and Ricardo Henriques. Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods*, 15:263–266, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 26:2292–2300, 2013.

Thomas Dertinger, Ryan Colyer, Gopal Iyer, Shimon Weiss, and Jörg Enderlein. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proceedings of the National Academy of Sciences (PNAS)*, 106(52):22287–22292, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*, 2021.

Thomas J Etheridge, Antony M Carr, and Alex D Herbert. GDSC SMLM: Single-molecule localisation microscopy software for ImageJ. *Wellcome open research*, 7(241), 2022.

Mohamadreza Fazel and Michael J Wester. Analysis of super-resolution single molecule localization microscopy data: A tutorial. *AIP Advances*, 12(010701), 2022.

Yue Fei, Shuang Fu, Wei Shi, Ke Fang, Ruixiong Wang, Tianlun Zhang, and Yiming Li. Scalable and lightweight deep learning for efficient high accuracy single-molecule localization microscopy. *Nature Communications*, 16(7217), 2025.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pp. 1608–1617, 2018.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Mats GL Gustafsson. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *Journal of Microscopy*, 198:82–87, 2000.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Mike Heilemann, Sebastian Van De Linde, Mark Schüttpelz, Robert Kasper, Britta Seefeldt, Anindita Mukherjee, Philip Tinnefeld, and Markus Sauer. Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angewandte Chemie-International Edition*, 47(33), 2008.

Stefan W Hell and Jan Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics Letters*, 19(11):780–782, 1994.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint*, 1606.08415, 2016.

Samuel T Hess, Thanu PK Girirajan, and Michael D Mason. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical Journal*, 91(11):4258–4272, 2006.

Bo Huang, Sara A Jones, Boerries Brandenburg, and Xiaowei Zhuang. Whole-cell 3D STORM reveals interactions between cellular structures with nanometer-scale resolution. *Nature Methods*, 5:1047–1052, 2008.

Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5754–5763, 2019.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2018.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2:83–97, 1955.

Romain F Laine, Hannah S Heil, Simao Coelho, Jonathon Nixon-Abell, Angélique Jimenez, Theresa Wiesner, Damián Martínez, Tommaso Galgani, Louise Régnier, Aki Stubb, et al. High-fidelity 3D live-cell nanoscopy through data-driven enhanced super-resolution radial fluctuation. *Nature Methods*, 20:1949–1956, 2023.

Mickaël Lelek, Melina T Gyparaki, Gerti Beliu, Florian Schueder, Juliette Griffié, Suliana Manley, Ralf Jungmann, Markus Sauer, Melike Lakadamyali, and Christophe Zimmer. Single-molecule localization microscopy. *Nature Reviews Methods Primers*, 1(39), 2021.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13619–13627, 2022.

Xinyang Li, Xiaowan Hu, Xingye Chen, Jiaqi Fan, Zhifeng Zhao, Jiamin Wu, Haoqian Wang, and Qionghai Dai. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nature Computational Science*, 3:1067–1080, 2023.

Yiming Li, Markus Mund, Philipp Hoess, Joran Deschamps, Ulf Matti, Bianca Nijmeijer, Vilma Jimenez Sabinina, Jan Ellenberg, Ingmar Schoen, and Jonas Ries. Real-time 3D single-molecule localization using experimental point spread functions. *Nature Methods*, 15:367–369, 2018.

Yiming Li, Elena Buglakova, Yongdeng Zhang, Jervis Vermal Thevathasan, Joerg Bewersdorf, and Jonas Ries. Accurate 4Pi single-molecule localization using an experimental PSF model. *Optics Letters*, 45(13):3765–3768, 2020.

Sheng Liu, Jianwei Chen, Jonas Hellgoth, Lucas-Raphael Müller, Boris Ferdman, Christian Karras, Dafei Xiao, Keith A Lidke, Rainer Heintzmann, Yoav Shechtman, et al. Universal inverse modeling of point spread functions for SMLM localization and microscope characterization. *Nature Methods*, 21:1082–1093, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

Charles W Mccutchen. Superresolution in microscopy and the Abbe resolution limit. *Journal of the Optical Society of America*, 57(10):1190–1192, 1967.

Hamza Mentagui, Luca Calatroni, Sébastien Schaub, and Laure Blanc-Féraud. Physics-inspired generative adversarial modelling for fluctuation-based super-resolution microscopy. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2024.

Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations (ICLR)*, 2021.

Elias Nehme, Lucien E Weiss, Tomer Michaeli, and Yoav Shechtman. Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica*, 5(4):458–464, 2018.

Wei Ouyang, Andrey Aristov, Mickaël Lelek, Xian Hao, and Christophe Zimmer. Deep learning massively accelerates super-resolution localization microscopy. *Nature Biotechnology*, 36:460–468, 2018.

Martin Ovesný, Pavel Křížek, Josef Borkovec, Zdeněk Švindrych, and Guy M Hagen. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics*, 30(16):2389–2390, 2014.

George Patterson, Michael Davidson, Suliana Manley, and Jennifer Lippincott-Schwartz. Superresolution imaging using single-molecule localization. *Annual Review of Physical Chemistry*, 61: 345–367, 2010.

James Pawley. *Handbook of Biological Confocal Microscopy*, volume 3. Springer, 2006.

Gabriel Peyré, Marco Cuturi, et al. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Patrick Putzky and Max Welling. Recurrent inference machines for solving inverse problems. *arXiv preprint*, 1706.04008, 2017.

Jonas Ries. SMAP: a modular super-resolution microscopy analysis platform for SMLM data. *Nature Methods*, 17:870–872, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Kurt Rossmann. Point spread-function, line spread-function, and modulation transfer function: tools for the study of imaging systems. *Radiology*, 93(2):257–272, 1969.

Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods*, 3:793–796, 2006.

Duncan P Ryan, Megan K Dunlap, Martin P Gelfand, James H Werner, Alan K Van Orden, and Peter M Goodwin. A gain series method for accurate EMCCD calibration. *Scientific Reports*, 11 (18348), 2021.

Daniel Sage, Thanh-An Pham, Hazen Babcock, Tomas Lukes, Thomas Pengo, Jerry Chao, Ramraj Velmurugan, Alex Herbert, Anurag Agrawal, Silvia Colabrese, et al. Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software. *Nature Methods*, 16: 387–395, 2019.

Lukas Scheiderer, Zach Marin, and Jonas Ries. MINFLUX–molecular resolution with minimal photons. *Nature Protocols*, 19:238–247, 2025.

Lothar Schermelleh, Alexia Ferrand, Thomas Huser, Christian Eggeling, Markus Sauer, Oliver Biehlmaier, and Gregor PC Drummen. Super-resolution microscopy demystified. *Nature cell biology*, 21:72–84, 2019.

Artur Speiser, Lucas-Raphael Müller, Philipp Hoess, Ulf Matti, Christopher J Obara, Wesley R Legant, Anna Kreshuk, Jakob H Macke, Jonas Ries, and Srinivas C Turaga. Deep learning enables fast and dense single-molecule localization with high accuracy. *Nature Methods*, 18:1082–1090, 2021.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision (ECCV)*, pp. 402–419, 2020.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Changdong Yu, Xiaojun Bi, and Yiwei Fan. Deep learning for fluid velocity field estimation: A review. *Ocean Engineering*, 271(113693), 2023.

Masoumeh Zareapoor, Pourya Shamsolmoali, Huiyu Zhou, Yue Lu, and Salvador García. Fractional correspondence framework in detection transformer. In *International Conference on Multimedia*, volume 1, pp. 5498–5506, 2024.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *International Conference on Learning Representations (ICLR)*, 2023.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations (ICLR)*, 2021.

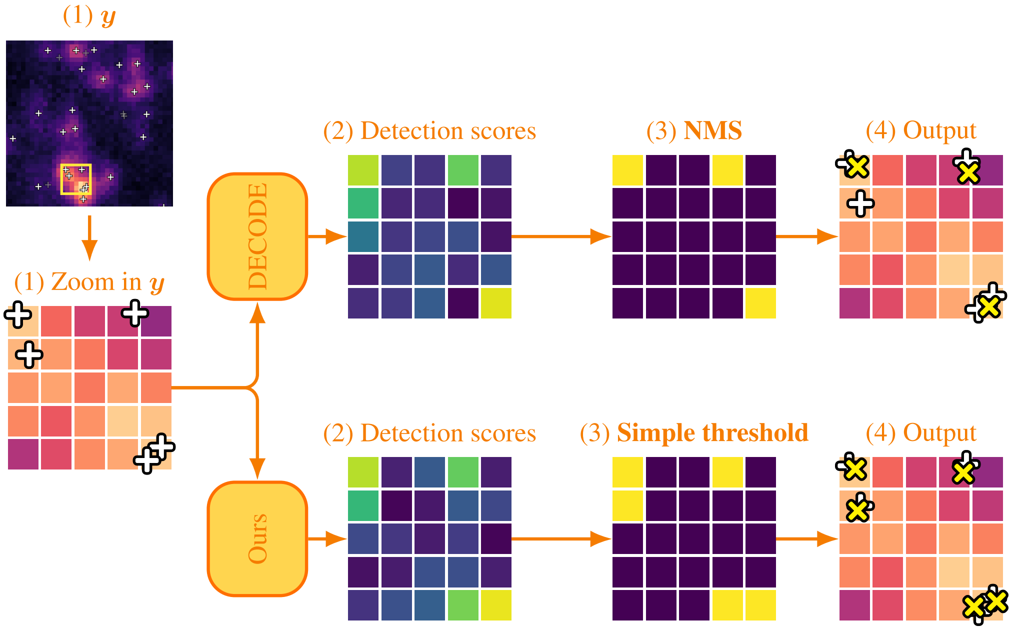


Figure 5: Toy illustration of the negative effects of pixel-wise loss functions and NMS post processing. (1) A synthetic image $\boldsymbol{y}$ with a zoomed version. White crosses represent the ground truth emitters. (2) DECODE's pixel-wise loss function enforces a high score for the bottom-right pixel, but does not consider that it is shared by two targets. By contrast, our optimal transport loss function also assigns a high score to a neighboring pixel, which can contribute to its neighborhood with an extended localization range. (3) DECODE's NMS variant merges the two adjacent predictions in the top-left corner if one of the prediction scores isn't high enough to be automatically retained. By contrast, our simple single-threshold filtering keeps all scores above the threshold, regardless of their spatial distribution. (4) Models' final outputs, showing the predicted emitters for each method with yellow crosses.

A APPENDIX

A.1 IMAGE FORMATION PROCESS

SMLM experimental setups typically employ either Electron-Multiplying CCD (EM-CCD) or scientific CMOS (sCMOS) cameras. Their sensors converts incident photons into a digital intensity value (ADU) through a sequence of physical processes, each of which introduces noise.

Let n be the incident photon count on the camera sensor. Initially, photon detection is modeled as a Poisson process — known as *shot noise* — with a mean proportional to n and the quantum efficiency (QE), and an offset known as the spurious charge (c):

$$n_{e,1} \sim \mathcal{P}\left(\mathrm{QE} \times n + c\right). \tag{9}$$

EM-CCD cameras introduce an additional amplification stage, modeled as a Gamma distribution with parameters $n_{e,1}$ and the electromagnetic gain (EM):

$$n_{e,2} \sim \Gamma(n_{e,1}, \mathrm{EM}) \text{ for EM-CCD camera, or } n_{e,2} = n_{e,1} \text{ for sCMOS camera.} \tag{10}$$

Subsequently, *read noise* is modeled by a normal distribution with mean $n_{e,2}$ and standard deviation σ_R:

$$n_{e,3} \sim \mathcal{N}(n_{e,2}, \sigma_R). \tag{11}$$

Finally, the analog-to-digital conversion process yields the observed ADU, scaled by the electrons per ADU (e_{ADU}) and offset by a baseline (B):

$$y = \min\left(\left\lfloor \frac{n_{e,3}}{e_{\mathrm{ADU}}} \right\rfloor + B \ , \ 65535\right) \tag{12}$$

Parameter	Evolve Delta 512	Dhyana 400BSI V3
Camera type	EMCCD	sCMOS
Quantum efficiency (QE)	0.90	0.95
Spurious charge (c)	0.002	0.002
EM gain (EM)	300	—
Readout noise (σ_R)	74.4	1.535
Electrons per ADU (e_{ADU})	45	0.7471
ADU baseline (B)	100	100

Table 4: Reported parameters of two typical cameras used in SMLM by their manufacturer.

Architecture	d	Parameters↓	Jaccard ↑	RMSE$_{\text{vol}}$ ↓	E$_{\text{3D}}$ ↑
CNN	HW	**2.31M**	0.866 ± 0.003	41.4 ± 0.7	0.732 ± 0.002
	$HW/4$	2.81M	$\mathbf{0.883 \pm 0.007}$	$\mathbf{39.2 \pm 0.5}$	$\mathbf{0.750 \pm 0.004}$
	$HW/16$	3.48M	0.875 ± 0.008	40.5 ± 0.5	0.739 ± 0.002
ViT	$HW/4$	5.95M	0.852 ± 0.006	40.6 ± 0.7	0.722 ± 0.001

Table 5: Evaluation of different decoder architectures with different number of predicted candidates over the MT0N1HDAS dataset (Sage et al., 2019), with standard deviations computed for three different training seeds. CNNs differ by the architecture of their head module, composed of alternating 2×2 max-pooling layers and double-convolution blocks.

No algebraic solution exists for the resulting distribution relating n and y (Ryan et al., 2021). Table 4 presents the parameters for two commonly used cameras in SMLM: the *Evolve Delta 512* camera for the Tubulin and NPC-Nup107 datasets (Li et al., 2018) and the *Dhyana 400BSI V3* camera for the NPC-Nup96 dataset (Fei et al., 2025).

A.2 ARCHITECTURE DETAILS

Decoder architecture. The decoder maps a latent representation $\boldsymbol{z}$ of the image - implemented as a $C \times H \times W$ tensor - to a set of d activations, implemented as a $d \times 5$ matrix (one activation contains five elements: the three spatial coordinates (x, y, z), the number of emitted photons n and the detection score s).

As we aim to predict a set from an image, and given the recent success of object detection by transformer architectures (Carion et al., 2020), we have considered using a vision transformer (ViT)

Range factor	Jaccard ↑	RMSE$_{\text{vol}}$ ↓	E$_{\text{3D}}$ ↑
1.0	0.869	47.0	0.710
1.1	0.874	45.1	0.722
1.2	0.880	42.6	0.737
1.5	**0.889**	**39.8**	**0.750**
2.0	0.884	40.1	0.749
3.0	0.880	41.3	0.738

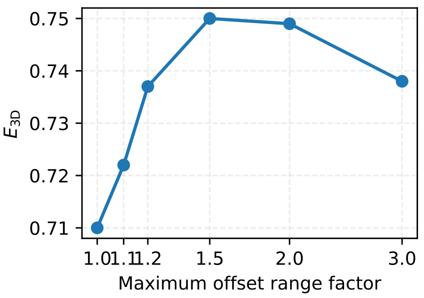


Figure 6: Evaluation of the impact of the maximum offset range of our CNN decoder. The maximum prediction range is controlled by a range factor; the final range equals the range factor multiplied by the output pixel size (which is 2× larger than the original image pixel size). Our results show that a range factor of 1.5 times the output pixel size, i.e. 3× the original pixel size, yields the best performance. Reported results are computed on the MT0N1HDAS dataset.

(Dosovitskiy et al., 2021). However, this architecture produced unsatisfactory results, see Table 5. We attribute this to two factors:

1. SMLM requires sub-pixel precision, but each ViT's token spans the whole image, so a small prediction error can severely affect the output.
2. The individual localisation-detection problem is highly local. Hence, ViT's global attention mechanism offers little benefit.

Therefore, we propose a convolution-based decoder architecture, composed of 2×2 max-pooling layers and residual blocks (He et al., 2016). We experimented with different numbers of max-pooling operations and found that a single max-pooling layer, followed by a residual block and an element-wise output layer yield the best results, see Table 5.

Formally, given the latent image $\boldsymbol{z}$, our decoder is defined as $\boldsymbol{D} : \mathbb{R}^{C\times H\times W} \longmapsto \mathbb{R}^{5\times H/2\times W/2}$, mapping a latent variable to a $H/2 \times W/2$ map with 5 channels, where each pixel is an activation prototype. Consider a single pixel i of $\boldsymbol{D}$'s output, and let $(\tilde{x}_i, \tilde{y}_i)$ be its 2D coordinates in the camera coordinate system. The five elements output for this pixel encode the characteristics of the underlying candidate activation: the detection score $\hat{s}_i$, the relative lateral coordinates $(\Delta\hat{x}_i, \Delta\hat{y}_i)$, the depth $\hat{z}_i$ and the number of emitted photons $\hat{n}_i$. The absolute lateral coordinates $(\hat{x}_i, \hat{y}_i)$ are reconstructed by summing $(\Delta\hat{x}_i, \Delta\hat{y}_i)$ with $(\tilde{x}_i, \tilde{y}_i)$. The magnitude of the relative coordinate offsets $(\Delta\hat{x}, \Delta\hat{y})$ predicted by the decoder is set to 1.5 times the output pixel size (or in other words 3× the pixel size of the original image, given the final max pooling). This extended range permits multiple activations to be mapped within a single pixel area, as neighbouring activations can contribute to their surroundings. Figure 6 shows experiments for various magnitude of the coordinate offsets. With a factor of 1× the output pixel size, each output pixel can predict locations only on the exact surface it covers. In this regime, the optimal transport solution reduces to an identity pixel-wise mapping.

Finally, the output is formatted into a candidate set $\hat{\mathcal{X}} = \{(\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{n}_i)\}_{1\leq i\leq d}$ and a detection scores set $\mathcal{S} = \{\hat{s}_i\}_{1\leq i\leq d}$. We integrate this reconstruction process into the decoder, meaning $\boldsymbol{D}(z) = (\hat{\mathcal{X}}, \hat{\mathcal{S}})$.

Differentiable simulation within our model. During inference, our algorithm selects a subset of candidate detections by thresholding their confidence scores. However, this operation is non-differentiable, preventing direct gradient propagation during training. To mimic this behaviour while retaining differentiability, we replace it by a soft weighting that scales the photon count of each candidate by its detection confidence. For each candidate $\boldsymbol{x}_i$ in $\hat{\mathcal{X}}$, the network outputs the 3D coordinates $(\hat{x}_i, \hat{y}_i, \hat{z}_i)$, the raw photon count $\hat{n}_i$, and a detection confidence $\hat{s}_i \in (0, 1)$. We choose to modulate the photon count by the confidence, producing the weighted activation

$$\tilde{\boldsymbol{x}}_i = \big(\hat{x}_i,\ \hat{y}_i,\ \hat{z}_i,\ \hat{s}_i\,\hat{n}_i\big),$$

and the set of all such activations is denoted $\tilde{\mathcal{X}} = \{\tilde{\boldsymbol{x}}_i\}_{i=1}^{d}$. This causes activations with low detection scores to have a number of emitted photons near zero, making them almost non-existent, while keeping almost untouched activations with a detection score close to one, mimicking the effect of a hard threshold while remaining fully differentiable.

After derivation, the expected image $\hat{\boldsymbol{y}}$ is obtained by:

$$\hat{\boldsymbol{y}} = \mathbb{E}[\hat{\mathbf{y}}|\tilde{\mathcal{X}}] = \frac{\mathrm{QE} \times \mathrm{EM}}{e_{\mathrm{ADU}}}\mathbf{H}(\tilde{\mathcal{X}}) + B, \tag{13}$$

and with $\mathrm{EM} = 1$ for sCMOS camera. $\tilde{\boldsymbol{y}}$ is an end-to-end differentiable approximation of the reconstructed output, and can be used inside our iterative refinement scheme.

A.3 COMPUTATIONAL FOOTPRINT

Training is performed using an NVIDIA H100 GPU using the AdamW optimizer with a learning rate of 4×10^{-4}, a weight decay of 0.01, and a cosine annealing scheduler. We chose a batch size of 128 to maximize GPU usage, filling all 80 GB of VRAM. It can be lowered using smaller batch sizes or gradient accumulation.

Subpart	Multiply-Accumulate operations	Parameters
Encoder	1.03 GMac	1.05 M
Decoder	512.56 MMac	499.4 k
Residual Network	1.87 GMac	1.26 M
Renderer	94.64 MMac	0

Table 6: Multiply-Accumulate operations and number of parameters of our model subparts.

Algorithm		Jaccard ↑	RMSE_{vol} ↓	E_{3D} ↑
Sinkhorn's	$\epsilon = 10^{-2}$	0.531 ± 0.007	67.0 ± 5.2	0.397 ± 0.011
	$\epsilon = 10^{-3}$	0.883 ± 0.006	42.0 ± 1.3	0.739 ± 0.002
	$\epsilon = 10^{-4}$	0.883 ± 0.007	39.2 ± 0.5	$\mathbf{0.750 \pm 0.004}$
	$\epsilon = 10^{-5}$	0.877 ± 0.012	39.1 ± 1.1	0.747 ± 0.002
Hungarian		0.873 ± 0.007	39.7 ± 0.9	0.742 ± 0.001

Table 7: Evaluation of different algorithms for solving the optimal transport problem used during the computation of our loss function. Results are reported for the MT0N1HDAS dataset (Sage et al., 2019), with standard deviations computed for three different training seeds.

We trained for 14 hours 100 epochs of 1024 steps each, totaling approximately 100,000 steps. Excellent results ($\text{E}_{\text{3D}} \geq 0.72$ on EPFL's density=2.0 and high SNR dataset) are achieved after only 20 minutes of training, at around 2000 steps.

During inference, a batch size of 16 produces a peak VRAM usage of $8.7\,\text{GB}$ and processes 2500 64x64 images in $30\,\text{s}$, or $12\,\text{ms/frame}$.

Table 6 shows an overview of the computational resources for each subpart of our model.

A.4 REGULARIZED OPTIMAL TRANSPORT

As explained in the main text, we solve the regularized optimal transport problem of Eq. (7) with Sinkhorn's algorithm. Our motivations are both analytical and computational: compared to the standard bipartite matching, Sinkhorn's algorithm avoids the need for stop-gradient operations, is differentiable, and is computationally efficient on GPUs. In our implementation, we run Sinkhorn's algorithm in log space and compute gradients automatically via PyTorch's autograd module. Our implementation uses 20 iterations, as we have found that additional iterations do not improve performance. Additionally, we have included a masking step to ensure that candidates are only assign to target activations if they are capable of reaching it within their limited prediction range. The observed performance boosts for increased range factors highlight the benefits of using optimal transport rather than pixel-wise assignments.

Table 7 compares results for various regularization constant ϵ in regularized optimal transport problem. We also report results with a bipartite matching performed by the Hungarian algorithm, that yields lower performance. We observe that smaller values for ϵ yield better performance, with no improvement beyond $\epsilon = 10^{-4}$; thus our we set $\epsilon = 10^{-4}$ in our implementation. Note that our implementation of the Sinkhorn's algorithm includes the common practical heuristic of scaling ϵ by the median of the cost matrix (Flamary et al., 2021).

A.5 ROBUSTNESS TO CAMERA PARAMETER MISMATCH

We have conducted a study to analyze the robustness of our model to mismatch in all camera parameters listed in Table 4.

To this end, we have generated a synthetic dataset of 2048 frames with a mean density of 2.0, each rendered with camera parameters independently jittered by noise from a log-normal distribution, i.e. scaled by e^{ρ} where $\rho \sim \mathcal{N}(0, \sigma)$. Note that we apply a similar data augmentation process during training, with $\sigma = 0.03$.

Jitter strength	Jaccard ↑	RMSE_{vol} ↓	E_{3D} ↑
$\sigma = 0$	0.956 ± 0.001	34.31 ± 0.26	0.811 ± 0.001
$\sigma = 0.03$	0.958 ± 0.002	34.52 ± 0.20	0.810 ± 0.001
$\sigma = 0.10$	0.957 ± 0.001	34.50 ± 0.25	0.810 ± 0.001
$\sigma = 0.30$	0.954 ± 0.001	35.21 ± 0.32	0.805 ± 0.002

Table 8: Evaluation of the robustness of our model to domain mismatch. Multiplicative noise of increasing strength, sampled from a zero-mean log-normal distribution, is applied to the simulator's camera parameters. Our model demonstrates strong resilience to this perturbation.

We have evaluated our model performance under increasing noise strength, to assess performance for increasing domain gap between training and test data. The results are reported in Table 8. Interestingly, our model appears extremely resilient to this type of mismatch. From an architecture standpoint, the use of LayerNorm and 2D convolutions without additive bias makes the network insensitive to scaling. We hypothesize that this architectural choice paired with the small data augmentation during training results in remarkably stable performance with respect to this issue.