Deep learning enables fast and dense single-molecule localization with high accuracy

Artur Speiser[#,1,2,3,4], **Lucas-Raphael Müller**[#,5,6], **Philipp Hoess**[5], **Ulf Matti**[5], **Christopher J. Obara**[7], **Wesley R. Legant**[8,9], **Anna Kreshuk**[5], **Jakob H. Macke**[#,1,2,3,10,†], **Jonas Ries**[#,5,†], **Srinivas C. Turaga**[#,7,†]

[1]Machine Learning in Science, Excellence Cluster Machine Learning, Tübingen University, Germany

[2]Computational Neuroengineering, Department of Electrical and Computer Engineering, Technical University of Munich, Munich, Germany

[3]research center caesar, an associate of the Max Planck Society, Bonn, Germany

[4]International Max Planck Research School 'Brain and Behavior', Bonn/Florida

[5]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany

[6]Ruprecht Karls University of Heidelberg, Heidelberg, Germany

[7]HHMI Janelia Research Campus, Ashburn, VA, USA

[8]Joint Department of Biomedical Engineering, UNC, Chapel Hill, NC, USA, and NCSU Raleigh, NC, USA

[9]Department of Pharmacology, University of North Carolina, Chapel Hill, NC, USA

[10]Department Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany

[#] These authors contributed equally to this work.

Abstract

Single-molecule localization microscopy (SMLM) has had remarkable success in imaging cellular structures with nanometer resolution, but standard analysis algorithms necessitate the activation of only single isolated emitters which limits imaging speed and labeling density. Here, we overcome this major limitation using deep learning. We developed DECODE, a computational tool that can localize single emitters at high density in 3D with highest accuracy for a large range of imaging modalities and conditions. In a public software benchmark competition, it outperformed all other fitters on 12 out of 12 data-sets when comparing both detection accuracy and localization error, often by a substantial margin. DECODE allowed us to take fast dynamic live-cell SMLM data

with reduced light exposure and to image microtubules at ultra-high labeling density. Packaged for simple installation and use, DECODE will enable many labs to reduce imaging times and increase localization density in SMLM.

Introduction

Single-molecule localization microscopy (SMLM) (e.g. PALM[1] and STORM[2]) has become an invaluable super-resolution method for biology, as it can resolve cellular structures with nanometer precision. It is based on acquiring a large number of camera frames, in each of which only a tiny fraction of the emitters are activated into a bright 'on' state, so that their images do not overlap. This allows precise localization of the emitter coordinates by fitting a model of the Point Spread Function (PSF). A super-resolution image is then reconstructed from these coordinates. This principle of SMLM is at the same time one of its main limitations: the need for sparse activation leads to long acquisition times. This results in low throughput, poor time resolution when imaging dynamic processes, low labeling densities, and a reduced choice of fluorophores. Additionally, long acquisition times in combination with high excitation laser intensities needed for single-molecule imaging cause strong phototoxicity in live-cell SMLM.

All of these limitations can be mitigated by activating emitters at a higher density. In this 'multi-emitter' setting, PSFs are no longer well-separated but may overlap, making both the detection of multiple nearby emitters and their accurate localization computationally challenging. This is not adequately addressed by existing algorithms: Current 'multi-emitter' fitting algorithms[3–5] work reasonably well on two-dimensional samples where all emitters have the same z-coordinate and thus produce identical PSFs. These algorithms, however, have had limited success for realistic three-dimensional biological structures. In a software competition that benchmarked SMLM algorithms using realistic computer-generated data, simple single-emitter fitters outperformed dedicated high-density fitters on three-dimensional samples even in the high density regime[6].

Deep learning is revolutionizing biological image analysis[7–9]. For SMLM, deep learning holds promise to extract emitter coordinates and additional parameters under conditions and densities too complex for traditional fitters. With enough training data, deep networks are flexible function approximators which can be trained to recognize patterns in the image and thus transform images directly into predicted emitter configurations, even for challenging high densities of emitters. While groundtruth data to train the neural network is typically not available, synthetic training data can be generated by numerically simulating the imaging process[10, 11]. Convolutional neural networks (CNNs, a class of deep networks suitable for image data) have recently been used to extract parameters describing single isolated emitters such as color, emitter orientation, z-coordinate, background or aberrations[12–15] and to design optimized PSFs[16]. Two recent studies (DeepSTORM3D[16] and DeepLoco[17]) used CNNs for extracting emitter coordinates, and outperformed traditional single-emitter fitting algorithms at densities higher than the single-molecule regime. These studies illustrate the potential of deep learning for SMLM, however they have only been demonstrated either for exotic engineered point spread functions or on simulated data.

Here we present the DECODE (DEep COntext DEpendent) method for deep-learning based single-molecule localization that achieves high accuracy across a wide range of emitter densities and brightnesses. DECODE uses a novel deep network output representation, architecture, and cost function, which enable simultaneous detection and sub-pixel localization of single emitters. Uniquely, DECODE is able to predict both the probability of detection and the uncertainty of localization for each emitter. As the timing and duration of emitter activations are stochastic, they regularly persist over several imaging frames. The DECODE architecture can integrate information across neighboring frames ('temporal context'), which improves emitter detection and localization.

In the public SMLM challenge[6], DECODE outperformed all existing methods on 12 out of 12 datasets. Compared to previous deep learning based high-density fitters[16], DECODE is 10x faster and up to 2x more accurate, and it can be applied to a wide range of PSFs. We demonstrate on biological structures that DECODE allows for 5-fold higher labeling densities or 10-fold faster imaging compared to imaging in the single emitter regime, and thus enables fast live-cell SMLM with reduced light exposure and visualization of dynamic processes. We show the versatility of DECODE by re-analyzing a published Lattice Light-Sheet PAINT data set[18] for which we could substantially improve fluorophore detection and localization accuracy. DECODE is packaged for simple use and can be easily trained and used by non-expert users, without having to design new network architectures. Thus, it will enable the entire community to overcome the need of sparse activation as one of the main bottlenecks in SMLM.

Results

DECODE network

DECODE introduces a new output representation and architecture for detecting and localizing emitters. For each image frame it predicts multiple channels with the same dimensions as the input image (Fig. 1a). The first two channels indicate the probability p that an emitter exists near that pixel, as well as its brightness N (number of photons emitted by the emitter in the frame). The next three channels describe the coordinates of the emitter with respect to the center of the pixel, $xyz = [\ x,\ y,\ z]$. An additional channel predicts the background intensity B in each pixel.

This architecture overcomes limitations of current deep-learning[16, 17] and non-deep learning based high density approaches in three ways: First, DECODE predictions scale only with the number of imaged pixels (not super-resolution voxels as in DeepSTORM3D), resulting in over 20-fold improvement in prediction speeds, and the use of continuous sub-pixel coordinates eliminates a voxel size dependent limit on precision. The local output representation used by DECODE also avoids the potentially challenging non-local mapping of pixels to global coordinates used in DeepLoco.

Second, DECODE has four additional output channels that estimate the uncertainty of the localization along each coordinate given by $\sigma xyz = [\sigma x, \sigma y, \sigma z]$ and of the brightness σN. These predicted localization uncertainties can be used to filter out poorly localized detections to improve the rendering of super-resolution images. In addition, training the

network to additionally predict the localization uncertainty corresponding to each detection also helps to improve the quality of the detection probabilities p by implicitly grouping all the detections corresponding to the same emitter. In contrast, standard output representations which only indicate the probability of detecting an emitter on a per-voxel basis make it more challenging to correctly group detection probability voxels corresponding to the same emitter in high emitter density and high localization uncertainty scenarios.

Third, the DECODE network integrates information across multiple frames with a two-stage design: The first stage (frame analysis module) analyses single imaging frames using a 2D multi-resolution convolutional network based on the "UNet" architecture[19] to compute a feature representation of the single frame (1). The second stage (temporal context module) integrates the feature representations of the frame with those of the previous and next imaging frame using a second 2D UNet to produce the final predictions. As emitters persist over several frames, this improves detection and localization accuracy.

Training the DECODE network using simulator learning

We train DECODE to simultaneously detect and localize emitters in SMLM measurements. Ground truth data for supervised learning are not easily available for SMLM. However, it is possible to simulate realistic images of activated emitters as the physics of imaging single molecules is well understood[11]. We train the DECODE network by generating a large amount of simulated data. To avoid structural bias[7], we place emitters at random coordinates, and calculate simulated images with a realistic image-formation model that includes dye photo-physics, a measured PSF and camera noise (see Methods).

We trained the DECODE network to predict the probability of detection, along with the sub-pixel localization and localization uncertainty of each detected emitter. Our loss function has three terms: 1) a count loss that compares the true and detected number of emitters in the image; 2) a localization loss that trains the network to correctly localize the detected emitters and estimate the localization uncertainty and emitter brightness; 3) an optional background loss. The count and localization loss functions were derived together as an approximation to a spatial point process probability distribution. They work together to correctly train the DECODE network to predict one detection per emitter, and to correctly assign the localization uncertainty of each emitter to the corresponding detection. Together, they constitute a novel loss for counting, detecting, and localizing sets of discrete point-like objects.

The count loss first constructs a Gaussian approximation to the predicted number of emitters by summing the mean and the variance of the Bernoulli detection probability map, and then maximizes the probability of the true number of emitters under this distribution. Uncertain detections will lead to large predicted count variance, while confident detections will result in low variance. Thus, the count loss encourages a detection probability map with sparse but confident predictions. The localization loss models the distribution of sub-pixel localizations xyz with a coordinate-wise independent Gaussian probability distribution[20] with standard deviation σxyz. For imprecise localizations, this probability is maximized for large σxyz, for precise localizations for small σxyz. The distribution of all localizations over the entire image is approximated as a weighted average of individual localization distributions, where

the weights correspond to the probability of detection. By optimizing both the probability of detection, the sub-pixel localization xyz, and σxyz simultaneously, the network learns not only the best predictions for the coordinates of the emitters, but also the best estimate for their localization uncertainties. The emitter brightness predictions N and their uncertainties σN are optimized similarly. Finally, the optional background loss computes the mean squared error between the true and predicted background images B.

DECODE achieves high accuracy for a wide range of simulated data

Performance metrics—The quality of SMLM data analysis is commonly quantified by two factors: First, the detection accuracy quantifies the fraction of emitters that are detected. The metric we use here is the Jaccard Index (JI)[6], that sets the true positives (TP) in relation to the false positives (FP) and false negatives (FN), JI = TP/(TP + FN + FP). The second factor is the localization error, i.e. how close the measured coordinates are to the true coordinates, measured here as the RMSE averaged over the dimensions (see Methods). We matched the detected emitters to the ground truth emitters in 3D with a lateral threshold of 250nm and an axial threshold of 500nm.

There is a natural trade-off between JI and localization error: Discarding all but the brightest and best separated emitters will result in a good (low) localization error but a bad (low) JI. Conversely, including also poorly localized emitters might improve JI, but deteriorates the localization error. The optimal operating point between these two extremes will depend on the experimental conditions and the scientific question. Because DECODE also provides uncertainties for each localization, it offers a straightforward way to filter localizations and thus set the desired balance between the number of detected emitters and the localization error that can be tolerated.

The Cramér-Rao Lower Bound (CRLB) gives the minimum achievable localization error for an optimal fitter given a known PSF, background, and noise model[21]. Most commonly, it is calculated under idealized conditions (i.e. non overlapping PSFs, homogeneous background, assuming the chosen PSF model to be the true model) and we use it here for comparison as a best-case limit for localization error.

DECODE approaches the Cramér-Rao Lower bound for low densities—We simulated 100,000 frames with exactly one emitter per frame at random coordinates with a constant brightness and background and trained DECODE without temporal context. On this data with sparse activations, DECODE approaches the single emitter CRLB, i.e. the theoretical limit of precision (Fig. 2a). It thus performs as well as Maximum Likelihood Estimation (MLE) based fitters, which have also been shown to reach the CRLB[22] in this regime.

DECODE's uncertainty estimates are well calibrated—In the high density regime, DECODE's σ predictions correlate closely to the measured localization error (Fig. 2b), much better than the single emitter CRLB estimate that assumes isolated emitters (correlation coefficient 0.86 for σ vs. 0.07 for single emitter CRLB). For the low-density regime, the uncertainty estimates are in line with the measured error and the single emitter CRLB (Fig. 2a).

Temporal context improves localization error and detection—DECODE's temporal context module pools information across multiple (we used 3) frames, to model the fact that emitters can persist in multiple subsequent frames. Use of this context module improves both the detection accuracy (JI) and the localization error (Fig. 2c). The increase in JI is apparent for all densities and SNRs. In addition, the RMSE is reduced by up to 20nm. Overall, the temporal context has a large impact across imaging conditions, and is also more powerful than 'grouping' approaches which are often applied to localizations in a post-processing step (see Extended Data Fig. 2).

DECODE architecture outperforms a voxel based network architecture and a multi-emitter fitter—To assess how the DECODE network architecture performs against other deep learning based and iterative methods, we directly compared to DeepSTORM3D[16] and CSpline[3], a matching pursuit style multi-emitter fitter based on MLE, using the code provided by the authors. To minimise the risk of sub-optimal training, we trained DeepSTORM3D on data sampled from our generative model using the same parameters we used for the training of DECODE. For both DeepSTORM3D and CSpline we performed a parameter grid search over user-defined parameters to maximize their performance (measured as efficiency score[6]). To facilitate the comparison of localization precision, we filtered out DECODE localizations with the highest inferred uncertainties such that the remaining number match DeepSTORM3D. DECODE outperforms the other methods across all densities and SNRs (Fig. 2e, Extended Data Fig. 3) even without temporal context. When we use temporal context, DECODE reduces the localization error up to two-fold compared to DeepSTORM3D. Although both methods are based on deep learning, this performance improvement is based on the differences in output representation and loss function between DECODE and DeepSTORM3D. The localization error of DeepSTORM3D is limited by the super-resolution voxel size[16] (Extended Data Fig. 4), which prevents the method from achieving the single emitter CRLB, unlike DECODE which has no such limitation. Because DECODE has multiple output maps it is also able to provide accurate estimates of the signal photon counts and background values (Extended Data Fig. 9). Notably, DECODE performs favourably in fitting time (Extended Data Fig. 6), taking less then 1.5s to analyze 1000 frames of 64x64 pixels, while DeepSTORM3D requires between 34s and 54s and CSpline requires between 14s and 2680s, which is up to 1900-fold slower than DECODE. Training the DECODE network to convergence on a NVIDIA RTX2080Ti GPU requires around 10h while DeepSTORM3D takes around 50h.

DECODE outperforms all fitters on a public SMLM benchmark—The 2016 SMLM challenge is an on-going and continuously updated second generation comprehensive benchmark evaluation developed for the objective, quantitative evaluations of the plethora of available localization algorithms[6, 23]. It offers synthetic datasets for training, created to emulate various experimental conditions. To avoid overfitting, evaluations are carried out on data not shared with contestants. It calculates various quality metrics, among them RMSE lateral or volume localization error, as applicable for 2D and 3D data respectively, the Jaccard index JI quantifying detection accuracy and a single 'Efficiency' score that combines RMSE and JI. The performanceof DECODE in the SMLM 2016 challenge, including extensive evaluations and side by side comparisons, is available

online[†]. DECODE outperformed all 39 algorithms on 12 out of 12 datasets, often by a substantial margin (Fig. 3, data from challenge website, current as of Oct 1st, 2020). The datasets included high (N1) and low (N2, N3) signal to noise ratios (SNR), with low (LD) or high (HD) emitter densities, with 2D, astigmatism (AS) and double Helix (DH) PSF based imaging modalities.

DECODE achieves an average efficiency score of 66.6% out of the best possible score of 100% (achievable only by a hypothetical algorithm that accurately detects 100% emitters with 0nm localization error). This is compared to an average score of 48.3% and 45.6% for all second and third place algorithms, respectively. The difference is particularly large under difficult imaging conditions, when high emitter densities and low SNR can conspire to make detection and localization challenging, particularly so for the double helix PSF. For example, compared to the second best algorithm (SMAP2018) in the Low SNR/high density/double helix condition, DECODE improves the localization error from 75.2nm to 48.4nm and the JI from 30.0% to 67.5%.

DECODE enhances super-resolution reconstructions by improving both the detection and the localization of single molecules. An example of this can be seen in Fig. 3c, where we compare the reconstruction obtained with DECODE and CSpline[3] on a high-density 3D double-helix dataset (using settings provided by the authors, github.com/ZhuangLab/storm-analysis). Other deep learning based approaches have not yet submitted their results. However, we performed comparisons to DeepSTORM3D on low SNR high density training datasets and again achieved superior results (efficiency score of 51% against 32% on double helix and 45% against 31% on astigmatism data. See Supplementary Fig. S5, S6). Thus, DECODE is setting new quantitative standards for localization algorithms, across both low and high SNRs and densities.

Considerations—As with any fitter, DECODE relies on an accurate PSF model and proper parameters, otherwise artifacts will dominate the predictions. When the localization uncertainty is large, for very dim and dense localizations far from the focal plane, DECODE has a bias towards predicting localizations close to the pixel center. This effect can be overcome by filtering out localizations with large predicted uncertainty (Extended Data Fig. 8, Methods).

DECODE reduces imaging times by one order of magnitude

By enabling accurate emitter localization at high densities of more than $2.5\mu m^{-2}$ per frame (Fig. 2 c), DECODE can yield high-quality super-resolution reconstructions with much shorter imaging times. We demonstrate this by imaging and reconstructing the same sample of labeled microtubules at four different activation laser powers using STORM (stochastic optical reconstruction microscopy)[24, 25]. This results in different emitter densities per frame between 0.08 and $0.86\mu m^{-2}$. The imaging time was chosen to result in the same number of total localizations and decreased from 1120s to 460s, 250s and 93s for stronger activation.

[†] http://bigwww.epfl.ch/smlm/challenge2016/leaderboard.html

We trained and applied one common DECODE model to all four datasets (Fig. 4a). Whereas CSpline reconstructions quickly degrade with high emitter densities, DECODE consistently yields reconstructions with high accuracy even for the densest sample. We quantified the lateral resolution using Fourier Ring Correlation (FRC)[26], which estimates resolution by measuring the correlation of two different reconstructions of the same image across spatial frequencies. DECODE consistently improves the x,y resolution by 20 nm - 30 nm over CSpline across all imaging densities (Fig. 4b and c) while detecting around 30% more localizations.

DECODE enables fast live-cell SMLM with reduced light exposure

Fast imaging is especially relevant for live-cell SMLM where the dynamics of the biological system under investigation dictate the necessary time resolution. At the same time, fast imaging usually requires high laser powers, deteriorates resolution[27] and leads to substantial phototoxicity[28]. As DECODE allows activating emitters to high density, it enables faster imaging with decreased light dose for a given number of localizations. We were able to image dynamic changes of the Golgi apparatus (Fig. 4d) and the endoplasmic reticulum (Fig. 4e) with 7.5s temporal resolution. We imaged nuclear pore complexes in living cells[29] within only 3 seconds (Fig. 4f), 7 times faster than our previous speed-optimized live-cell SMLM[27] and with 70% reduced light dose and thus phototoxicity.

DECODE enables ultra-high labeling densities

Labeling densities in SMLM are fundamentally limited by the fraction of emitters that are in the bright state. For the best performing fluorophore Alexa Fluor 647, even without UV activation about 0.05% of the emitters are in the bright state[30] due to activation by the red imaging laser and spontaneous activation. For the single-emitter blinking regime (activated emitter density $< 0.1\ \mu m^{-2}$), this limits the number of total emitters to about $200 \mu m^{-2}$. For higher labeling, pre-bleaching can be employed to reduce the number of emitters to this regime, but the resulting low labeling limits the resolution[18] and in the superresolution reconstructions sparse individual emitters become dominant (Fig. 4g). With DECODE, we can now image densely labeled samples that previously were inaccessible. We demonstrated this on immuno-labeled microtubules that were labeled about 5-fold higher than compatible with single-emitter fitting, resulting in much smoother and denser decoration of the microtubules (Fig. 4g). In 50 nm thick orthogonal reconstructions, only the densely labeled microtubules were resolved as hollow cylinders, whereas after pre-bleaching to single-emitter blinking, these reconstructions only showed individual emitters (Fig. 4g,3,4). Additional comparisons with DeepSTORM3D highlight that the superior output representation and loss function of DECODE are critical to reach the optimal resolution for this dataset (Extended Data Fig. 5).

DECODE enables high fidelity reconstructions of 3D lattice light sheet PAINT

To illustrate the general applicability of DECODE, we applied it to 3D lattice light sheet (LLS) microscopy combined with the PAINT (point accumulation for imaging of nanoscale topography labeling) technique[18]. In PAINT microscopy, the fluorophore labeling a sample stochastically binds and unbinds from the sample, providing dense labeling. In

LLS microscopy, thick volumes are imaged at high resolution by scanning a thin (1.1 μm) light sheet, with axial localization within the sheet enabled by astigmatism.

Single-molecule localization in LLS-PAINT is usually performed frame-wise using MLE fitting[31]. However, an emitter is visible in several adjacent z-planes in the volumetric data set. Thus, similar to exploiting the temporal context, we now use the same spatio-temporal context by analyzing three adjacent frames in the z-stack at the same time to improve detection accuracy and localization error.

We reconstructed a previously reported dataset of a chemically fixed COS-7 cell with intracellular membranes labeled byazepanyl-rhodamine (AzepRh)[18, 31] consisting of 70,000 3D volumes comprising more than 10 million 2D images acquired in 270nm steps. DECODE detected 500 million emitters, compared to 200 million emitters detected by the original algorithm. Thus, for a comparable quality of the reconstruction, only half of the frames are needed, reducing imaging times by over a day from 2.7 days to 1.35 days (Extended Data Fig. 7). At the same time, improved accuracy of DECODE results in sharper reconstructions (Fig. 5).

Discussion

We presented DECODE, a new deep-learning based method for single molecule localization that performs exceptionally well on dense 3D data. DECODE differs from traditional localization algorithms by simultaneously performing detection and localization of emitters. It can be used in a flexible and general manner for a wide range of imaging parameters (including arbitrary Point Spread Functions and noise models) and imaging modalities such as 3D lattice light sheet PAINT imaging. In a publicly available benchmark challenge it is the best performing algorithm in every condition, and often improves both localization and detection accuracy by a large margin.

By making use of the temporal context, DECODE improves detection accuracy and localization error of emitters that are active across multiple imaging frames. Temporal context is also used by post-processing steps in SMLM relying on 'merging' or 'grouping' of localizations, in which localizations occurring in consecutive images that are closer to each other than a fixed threshold are assumed to belong to the same emitter and their coordinates are averaged, weighted by the uncertainty of each localization. However, grouping does not improve detection of emitters, and it fails for dense or dim emitters whose localizations cannot be linked unambiguously across frames.

DECODE not only predicts coordinates of emitters, but also their uncertainty. This is highly useful for filtering out imprecise localizations, for reconstruction of superresolution images in which every localizations rendered as a Gaussian with a size proportional to the coordinate uncertainty, and as weights for quantitative coordinate-based analysis of SMLM data.

We demonstrated the performance of DECODE on various experimental SMLM data sets. We could show that the excellent performance on high-density data can increase the achievable localization density or decrease imaging times by one order of magnitude. This

allowed us to perform live-cell measurements on nuclear pore complexes with high temporal resolution and reduced light exposure, and to achieve ultra-high labeling on microtubules. LLS-PAINT data analyzed with DECODE showed markedly improved resolution due to substantial improvements in emitter detection and localization error.

Prediction of coordinates with DECODE can be as fast as GPU-based MLE-fitters for sparse activation, but greatly outperforms those for high densities, as the computational complexity of DECODE depends only on the size of the image and not the number of emitters in each imaging frame. However, it requires the training of a new neural network whenever the optical properties of the microscope change. This training can currently take over 10 hours on a single GPU, but after just 2 hours of training time, the localization error is within 1 nm and the JI within 2% of the final value (Extended Data Fig. 10). To reduce training times further, one can likely take an existing network and fine-tune its parameters using a smaller number of simulations, rather than training it from scratch. Ultimately, it may be possible to train a single network across multiple parameter settings or even PSFs, so that the same network can 'amortize' inference across multiple experimental settings.

To make DECODE easily usable by the entire community, we distribute it as a Python-based open source software package based on the PyTorch[32] deep learning library. We provide pre-compiled, easily installable code, along with detailed tutorials and integration into the SMAP SMLM analysis software[33]. To enable anyone to directly use DECODE for training and prediction without relying on prior programming knowledge and dedicated local hardware, we deploy these Jupyter notebooks in Google Colab, complementing a recent initiative to make deep learning based image analysis tools accessible to non-experts at minimal cost[34]. Thus, DECODE will enable a large community to directly perform SMLM in a new high-density regime with greatly increased imaging speeds or localization densities and excellent localization and detection accuracy.

Methods

DECODE network architecture for probabilistic single molecule detection and localization

Our architecture consists of two stacked U-nets[19] (Extended Data Fig. 1), each with two up- and downsampling stages and 48 filters in the first stage. Each stage consists of three fully convolutional layers with 3×3 filters. In each downsampling stage, the resolution is halved, and the number of filters is doubled, vice versa in each upsampling stage. Upsampling is performed using nearest neighbor interpolation to avoid checkerboard artifacts[35]. For multi-frame DECODE, three consecutive frames are processed by the first frame analysis U-net (with parameters shared for every frame), and the outputs are concatenated and passed to the second temporal context U-net. The entire DECODE network is always trained end-to-end by gradient descent.

For each camera pixel k, the DECODE network predicts i) a Bernoulli probability map p_k that an emitter was detected near that pixel, ii) the coordinates of the detected emitter x_k, y_k, z_k relative to the center of the pixel x_k, y_k, z_k, iii) a non-negative emitter brightness ("photon count") N_k, and iv) the uncertainties associated with each of these predictions, σx_k, σy_k, σz_k, σN_k. For each of these outputs, we use two additional convolutional layers that

follow the second U-net. We used the Exponential Linear Unit (ELU) activation function36 for all hidden units, and the logistic sigmoid nonlinearity for the non-negative detection probability p, brightness N, and the uncertainty outputs σx, σy, σz, σN (scaled by a pre-factor of three). For the coordinate outputs x, y, z we use the hyperbolic tangent nonlinearity which limits their range to $[-1, 1]$ (i.e. to twice the size of a pixel). This way, even though the network can at most predict one emitter per pixel, when necessary the neighboring pixels can each contribute in order to place multiple localizations within a single pixel.

Novel loss function for simultaneous detection, localization, and uncertainty estimation

Given a set of E simulated emitters active in each imaging frame with locations for each emitter e given by x_e, y_e, z_e and brightness N_e, and a background image map B_k simulated as described below, we developed a loss function that trains the DECODE network to detect the correct number of emitters, to predict the sub-pixel localization and brightness for each detection (along with the uncertainty), and to predict the image background. Our loss function is a sum of three terms – a count loss $\mathcal{L}_{count}$, a localization loss $\mathcal{L}_{loc}$, and a background loss $\mathcal{L}_{bg}$,

$$\mathcal{L} = \mathcal{L}_{count} + \mathcal{L}_{loc} + \mathcal{L}_{bg}. \tag{1}$$

The **count loss** $\mathcal{L}_{count}$ is a function of the detection probability map p_k with K total pixels and the total number of true emitters E. Interpreting p_k as a Bernoulli detection probability for a single emitter, we can compute the mean and variance of the predicted total number of emitters detected, if we were to independently sample binary detections from each p_k. While the predicted count distribution $P(E \mid \{p_k\})$ over the number of emitters detected by this Bernoulli sampling procedure follows an intractable Poisson binomial distribution, we can approximate this predicted distribution as a Gaussian distribution,

$$P(E \mid \{p_k\}) \approx P\left(E \mid \mu_{count}, \sigma_{count}^2\right) = \frac{1}{\sqrt{2\pi}\sigma_{count}}\exp\left(-\frac{1}{2}\frac{(E - \mu_{count})^2}{\sigma_{count}^2}\right). \tag{2}$$

The mean of a sum of Bernoulli random variables is the sum of the means $\mu_{count} = \sum_{k=1}^{K} p_k$, and the variance is the sum of the variances of each independent Bernoulli random variable $\sigma_{count}^2 = \sum_{k=1}^{K} p_k(1 - p_k)$. This count loss maximizes the log probability of the true number of emitters E under the Gaussian approximation of the predicted count probability distribution. This loss is minimized when μ_{count} correctly matches E, sparsely predicting only one non-zero p_k per detected emitter, and when σ_{count}^2 is small, which happens when p_k are confident and so nearly binary,

$$\mathcal{L}_{count} = -\log P\left(E \mid \mu_{count}, \sigma_{count}^2\right) = \frac{1}{2}\frac{(E - \mu_{count})^2}{\sigma_{count}^2} + \log\left(\sqrt{2\pi}\sigma_{count}\right). \tag{3}$$

The **localization loss** $\mathcal{L}_{loc}$ is a function of the true emitter locations, and the predicted detection probability map, and the sub-pixel localizations x_k, y_k, z_k, brightness N_k, along with the associated uncertainties σx_k, σy_k, σz_k, σN_k for each detected emitter. For each pixel k, we predict a 4D Gaussian distribution $P\left(\vec{u}_k \mid \vec{\mu}_k, \Sigma_k\right)$ over the absolute position and brightness of an emitter $\vec{u} = [x, y, z, N]$ detected in pixel k corresponding to the mean and uncertainty in the sub-pixel localization and brightness of the emitter detected in pixel k, with mean $\vec{\mu}_k = [x_k + \Delta x, y_k + \Delta y_k, z_k + \Delta z_k, N_k]$, and diagonal covariance matrix $\Sigma_k = diag\left(\sigma x_k^2, \sigma y_k^2, \sigma z_k^2, \sigma N_k^2\right)$,

$$P\left(\vec{u} \mid \vec{\mu}_k, \Sigma_k\right) = \frac{1}{\sqrt{(2\pi)^4 \det(\Sigma_k)}} \exp\left(-\frac{1}{2}\left(\vec{\mu}_k - \vec{u}\right)^\top \Sigma_k^{-1}\left(\vec{\mu}_k - \vec{u}\right)\right). \tag{4}$$

Here, the x_k, y_k, and z_k are the absolute coordinates for the center of pixel k, so $x_k + x_k$ corresponds to the absolute coordinates of the emitter to sub-pixel precision. We note that the localization loss defined below ignores the predicted localization and brightness for pixels where no emitter is detected, i.e. p_k is zero.

At any given point in training, the true number of emitters will not necessarily match the detected number of emitters perfectly, and we will not have a perfect correspondence between predicted emitters and true emitters. A full probabilistic loss function would sum over all possible assignments of true emitters to detected emitters in order to correctly evaluate $P\left(\vec{u} \mid \vec{\mu}_k, \Sigma_k\right)$. And since p_k will not necessarily be sparse, the correct cost function would include an intractably large sum over $\binom{K}{E}$ terms. We approximate this by constructing a Gaussian mixture model over the predicted per pixel distributions $P\left(\vec{u}_k \mid \vec{\mu}_k, \Sigma_k\right)$ with mixture weights equal to $p_k / \sum_{j=1}^{K} p_j$ where the denominator is a sum of the detection probability over all pixels in the image.

The resulting approximation leads to the following localization loss function which maximizes the probability of the true absolute coordinates and brightness of each ground truth emitter $\vec{u}_e^{GT}$ under the weighted mixture of per pixel probabilities,

$$\mathcal{L}_{loc} = -\frac{1}{E}\sum_{e=1}^{E} \log \sum_{k=1}^{K} \frac{p_k}{\sum_j p_j} P(\vec{u}_e^{GT} \mid \vec{\mu}_k, \Sigma_k). \tag{5}$$

The **background loss** $\mathcal{L}_{bg}$ computes the simple squared error between the predicted and true background maps,

$$\mathcal{L}_{bg} = \sum_k \left(B_k^{GT} - B_k^{pred}\right)^2. \tag{6}$$

Obtaining localizations and post-processing

The DECODE network predicts the probabilities pk of an emitter being located at a specific pixel k. To get deterministic, fast and precise final localizations we use a variant of spatial integration. A detection is considered at pixel k if one of two conditions is met. 1) $p_k > 0.6$. 2) $p_k > 0.3$, and it is a local maximum of a 4-connected neighborhood. These candidates are then registered as detections if the cumulative probability of p_k and its 4 nearest neighbor pixels is > 0.7. Therefore, if the network predicts high confidence detection probability (> 0.6) in two adjacent pixels, two emitters will be considered to be detected. However, if a cluster of pixels has low predicted probability, their probabilities will be clustered toward the local maximum, if the local maximum has probability > 0.3, and an emitter will be considered to have been detected if the integrated probabilities of the cluster are > 0.7. The algorithm can be expressed purely in the form of pooling and convolution operations and therefore runs efficiently on a GPU.

For difficult imaging conditions when the predicted localization uncertainties are large, i.e. high densities, low SNR values, and large offsets from the focal plane, the sub-pixel coordinates x, y, and z can be biased towards the center of the pixels (Extended Data Fig. 8). This is because with large predicted localization uncertainty, the predicted mean location is poorly constrained. This bias towards 0 (pixel center) scales with the uncertainty of the predictions and can produce artifacts in the reconstructed image depending on how the reconstruction is performed. If a reconstruction uses only the coordinates while ignoring the uncertainty, poorly localized emitters will cluster towards the pixel centers. A more expensive rendering procedure which renders a Gaussian localization distribution with variance proportional to the estimated uncertainty corresponding to each emitter will reduce the impact of this artifact since the bias is usually small relative to the localization uncertainty. Also, filtering out localizations with high uncertainty removes this artifact (Extended Data Fig. 8).

Simulating training data

Training samples are continuously generated in an asynchronous fashion and each frame is only used once as a target. For this reason the network cannot overfit to specific frames. The performance of our approach will depend on an accurate generative model and could show reduced performance when there is a mismatch between the simulated and experimental data. Thus, we developed a realistic model for the image formation process that incorporates dye blinking behaviour, a realistic PSF model and realistic camera read noise.

Structural prior—While incorporating prior structural information has shown to be beneficial[37, 38], there are concerns that these priors could potentially bias the model to the training data, which could result in the presence of misleading structures after the fitting procedure. We therefore sample the coordinates of the emitters from a 3D homogeneous spatial Poisson point process distribution with density as specified in the text, limits corresponding to the size of the image and the z-range for which the PSF was calibrated.

Photophysical prior—In contrast to prior work, DECODE can directly incorporate temporal context into the detection and localization of emitters, rather than as a post-

processing step. We simulate the temporal dynamics of emitters, at least over the short time scale of three imaging frames corresponding to the temporal context of the DECODE network.

For each emitter, the time of initial appearance t_0 is sampled from a continuous random distribution. The on-time of the emitter follows an exponential distribution parametrized by λ. For each emitter, we draw a photon flux from a Gaussian distribution $N(\mu_{flux}, \sigma_{flux})$. Together with the amount of time the emitter is active in each frame this determines the total number of photons emitted in a frame. Since the input to our model is only a window of three frames, we argue that it is not necessary to model long range temporal correlations that are part of a more detailed photoactivation model[39], like an emitter in the dark state which reappears many frames later. The aforementioned parameters are estimated by a *prefit* procedure as described in *Estimating simulation parameters*.

Point spread function—The PSF is a fundamental characteristic of a microscope, specifying the image formed by a single point emitter, and we approximate it to be spatially invariant across the field of view. Given the object $O(\vec{r})$ in the object plane, and $PSF(\vec{r})$, the image $I(\vec{r})$ results as

$$I(\vec{r}) = O(\vec{r}) \circledast PSF(\vec{r}), \tag{7}$$

where $\circledast$ denotes the convolution operator. While Gaussian approximations of the PSF are frequently used for both 2D and 3D data[4, 5], (cubic) spline functions have been shown to achieve more accurate results and can mimic almost arbitrary PSF's,[3, 22]. Following *Li et al.*[22] and *Babcock et al.*[3] a three-dimensional PSF can be modelled as

$$f_{i,j,k}(x, y, z) = \sum_{m=0}^{3} \sum_{n=0}^{3} \sum_{p=0}^{3} a_{i,j,k,m,n,p} \left(\frac{x - x_i}{dx} \right)^m \left(\frac{y - y_j}{dy} \right)^n \left(\frac{z - z_k}{dz} \right)^p, \tag{8}$$

where i, j, k are the voxel indices, dx, dy are the pixel sizes; dz is the step size in the axial dimension; x_i, y_j, z_k are the corner coordinates of the voxel (i, j, k) in the respective directions and $a_{i,j,k,m,n,p}$ are the respective spline coefficients, which amounts to 64 coefficients per pixel and per z-slice. In a bead calibration routine, the coefficients $a_{i,j,k,m,n,p}$ are estimated and account for varying experimental conditions. Because of the simple form of equation 8, the CRLB with respect to the fitting parameters x, y, z can be calculated easily as the diagonal elements of the inverse of the Fisher information matrix[21].

Camera model—All real datasets presented in this work were recorded with an EMCCD camera, with the exception of the LLS data which was recorded with a sCMOS camera. The measured camera signal is subject to various noise sources, which we will discuss in the following:

Shot noise originates from the stochastic nature of photons when interacting with the camera chip. The expected number of detected electrons is

$$\lambda_k = \lambda_{0,k} \cdot qe + c_s. \tag{9}$$

Here, λo,k is the expected number of photons that are collected in pixel k, qe is the quantum efficiency, and c_s the spurious charge, measured in electrons. The probability $p_{\text{shot}}(s_k)$ of observing the signal s_k in pixel k follows a Poisson distribution,

$$p_{\text{shot}}(s_k) = \frac{\lambda_k^{s_k} e^{-\lambda_k}}{s_k!} . \tag{10}$$

EMCCD Amplification noise stems from the amplification of photo electrons that pass through the gain register and stochastically generate additional electrons. For our EMCCD camera noise model we follow *Huang et al.*[40]. EMCCD amplification noise can be described approximately by a Gamma distribution,

$$\rho_{\text{EM}}(x \mid s_k, \theta) = \frac{1}{\Gamma(s_k)\theta^{s_k}} x^{s_k - 1} e^{-\frac{x}{\theta}} . \tag{11}$$

$\rho_{\text{EM}}(x \mid s_k, \theta)$ denotes the probability that s_k input photo electrons in pixel k with an EM gain of θ create x output electrons after the gain register.

Read noise stems from the process of converting electrons into a digital signal. In this process, the signal is usually multiplied by a gain factor g and an offset o is added to avoid negative signal. In this work, we convert the input camera image to photon units prior to inference by subtracting o and dividing by g. In addition, when using EMCCD cameras we divide by the EM gain θ, thus the units of the read noise are photo electrons. We approximate the read noise (both for sCMOS and EMCCD cameras) by a zero mean additive Gaussian distribution with variance σ^2,

$$\rho_{\text{read}}(x \mid 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} . \tag{12}$$

Training details

Training was performed on 40×40 pixel sized regions that are directly simulated or randomly selected from larger simulated images at each iteration. We used the AdamW optimizer[41] with a group learning rate of $6 \cdot 10^{-4}$ for the network parameters. We reduce the learning rate by a factor of 0.9 after every 1500 iterations with a batch size of 64. To stabilize training we employ gradient norm clipping with a maximum norm of 0.03. Very dim emitters with less then 50 photons are excluded from the ground truth targets (but still rendered) so that the network is discouraged to make predictions for practically invisible emitters.

Estimating simulation parameters—For training DECODE, a proper parametrization of the simulation is needed to match the real data distribution. In a prefitting step, the main parameters, i.e. the emitter on-time, emitter brightness and background, can be determined. The prefitting can be performed with a single-emitter MLE fitter after filtering the log-

likelihood value to exclude data from overlapping PSFs. This step is incorporated in the SMAP software for the sake of ease of use[33]. We observed that the precise values of the simulation parameters of the emitters' photophysics (i.e. lifetime and brightness) and density are not crucial, as the stochastic nature of the emitters' position, brightness and appearance time presents the network with data that matches the real experiments under different conditions and effectively covers a broad range of these parameters. The camera parameters are usually given by the manufacturer. The given network architecture and training parameters are effective across different real and simulated datasets and in our experience do not have to be optimized by the end user.

Evaluating localization error and reconstruction resolution

To evaluate performance on the challenge datasets, as well as our own simulations, we use two metrics:

First, instead of the Euclidean distance, we use the **localization error**, measured in nm, which is the RMSE averaged over the dimensions:

$$\text{RMSE}_d = \left(\frac{1}{TP} \sum_{i=1}^{TP} \sum_{k=1}^{d} \left(x_{i,k} - x_{i,k}^{GT} \right)^2 / d \right)^{1/2} \tag{13}$$

TP is the number of localizations that are matched to ground truth coordinates, d is the dimension (2 for 2D data, 3 for 3D data), $x_k = x, y, z$ are the predicted coordinates and $x_k^{GT} = x, y, z$ the ground truth coordinates.

Second, the **detection accuracy** or Jaccard index JI, which quantifies how well an algorithm does at detecting all the emitters while avoiding false positives:

$$\text{JI} = TP/(FN + FP + TP) \tag{14}$$

TP are the true positives, FN the false negatives and FP the false positives.

Localizations are matched to ground truth coordinates when they are within a circle of 250 nm radius and the distance in z is less than 500nm. As a single metric that evaluates the ability to reliably infer emitters with high precision we use the efficiency metric as defined in Ref. 6:

$$E = 1 - \sqrt{(1 - \text{JI})^2 + \alpha^2 d \text{RMSE}_d^2} \tag{15}$$

Lateral and axial efficiency are calculated based on $RMSE_2$ and $RMSE_1$ with alpha values of $\alpha = 1 \times 10^{-2}$ nm^{-1} and $\alpha = 0.5 \times 10^{-2}$ nm^{-1} respectively and then averaged to obtain the overall efficiency. Detection accuracy is expressed in units of 0 to 1 (or 0% to 100%), the efficiency ranges up to 1 (or 100%) for a perfect fitting algorithm.

The Fourier ring correlation[26, 42] (FRC) in Fig. 4a was calculated by dividing the data in 10 blocks of equal number of frames and constructing super-resolution images from even and odd blocks (pixel size 5nm).

Simulating data for performance evaluation

To simulate data for performance evaluation and comparison shown in Fig. 2, we assumed an ideal camera without EMCCD or read noise and an image size of 64×64 pixels. We used the PSF model that was acquired for the data set in Fig. 4a. Data used to test the effect of the SNR and density were simulated using the structural and photophysical prior previously described with an average on-time of 2 frames. Precise simulation parameters can be found in Supplementary Table S1. The CRLB is evaluated as the diagonal elements of the inverse of the Fisher information matrix[21] with the simulated parameters and spline interpolated experimental PSF model and was calculated with the SMAP software[33]. A bootstrap estimate (N = 10000) of the RMSE was used to estimate the SEM on the localization error.

Comparison with DeepSTORM3D and CSpline

For both methods we used the software provided by the authors. For the DeepSTORM3D comparison instead of using their PSF fitting procedure and generative model we sampled ground truth coordinates and training images using our model so that it exactly matches the simulated test data. To minimize possible effects of overfitting we generated 22,500 images with a size of 121×121 pixels (22k for training and 500 for validation). DeepSTORM3D uses a fourfold super-resolved grid in the $x - y$ dimensions and we chose discretization of 15nm in z. As the camera we emulate in these experiments has a pixel size of 120nm, each voxel of the output representation has a size of $30 \times 30 \times 15$nm. For DECODE (with and without temporal context), and DeepSTORM3D we trained six networks each on training data generated with average emitter densities of 0.65 and 2.17μm^{-2} as well as low, medium and high SNRs (1000, 5000 and 20,000 average photons). We used the low density network for the CRLB evaluation (Fig. 2a, Extended Data Fig. 4) and the simulated data with densities between 0.04 and 2.4μm^{-2} (Fig 2c,d, Extended Data Fig. 3) and the high density networks for densities between 2.4 and 5.6μm^{-2}. DeepSTORM3D has two hyperparameters that control the post-processing and determine the balance between recall and localization error. We performed a sweep over combinations of radius = [5,6,7,8,10] and threshold = [5,8,12,20,30,40] and picked the values that maximized the efficiency score on the validation data for each of the six networks. We discovered and fixed a bug in the DeepSTORM3D post-processing software which led to poor localizations. All DeepSTORM3D results were reported with the fixed post-processing algorithm.

For the CSpline comparison we created a bright artificial bead with 500k photons using our PSF model, which we used to generate the CSpline PSF model. The most critical settings are the find-max-radius and threshold, which we again optimized by sweeping over values find-max-radius = [2,3,4,5], threshold = [6,7,8,9,10] to maximize efficiency for each of the three SNRs on data generated with an average emitter density of 0.9 μm^{-2}.

DECODE for LLS-PAINT microscopy

A DECODE model for lattice light sheet point accumulation for imaging of nanoscale topography (LLS-PAINT) microscopy[18] was trained by simulating the imaging of an angled light sheet being swept through a volume. This leads to the same emitter appearing with fixed shift in the x and z coordinates relative to the imaged plane between consecutive camera frames. The offset in emitter coordinates from frame to frame are given by the microscope geometry as described in[31]. We simulated data with a high emitter density of 1 μm^{-2} to match the densities seen in LLS-PAINT.

We analyzed a large dataset corresponding to a fixed COS-7 cell with intracellular membranes labeled with azepanyl-rhodamine (AzepRh) described in Legant et al.[31]. Over a period of 2.7 days (64.8 hours), LLS-PAINT imaging yielded 70,000 3D volumes comprising more than 10 million 2D images. Significant non-uniform swelling of the sample was observed over the course of the imaging, which was approximately corrected by non-rigid registration in Legant et al.[31]. We applied the same correction transformation estimated by Legant et al.[31] to DECODE localizations.

We introduced an additional simulation-free training step and loss function to the training of the LLS-PAINT DECODE network based on the Re-weighted Wake Sleep algorithm[43] for training variational autoencoders (VAE)[44,45]. This form of auto-encoder learning allowed us to further optimize the parameters of the PSF and improve the background predictions based on the real data, as opposed to the simulation.

Sample preparation

Sample seeding—Before seeding of cells, high-precision 24mm round glass coverslips (No. 1.5H, catalog no. 117640, Marienfeld) were cleaned by placing them overnight in a methanol:hydrochloric acid (50:50) mixture while stirring. After that, the coverslips were repeatedly rinsed with water until they reached a neutral pH. They were then placed overnight into a laminar flow cell culture hood to dry them before finally irradiating the coverslips by ultraviolet light for 30min. Cells were seeded on clean glass coverslips 2 days before fixation to reach a confluency of about 50 to 70 % on the day of fixation. They were grown in growth medium (DMEM; catalog no. 11880-02, Gibco) containing 1× MEMNEAA (catalog no. 11140-035, Gibco), 1× GlutaMAX (catalog no. 35050-038, Gibco) and 10% (v/v) fetal bovine serum (catalog no. 10270-106, Gibco) for approximately 2 days at 37 °C and 5% CO_2.

Transfection—The plasmids encoding calnexin (Addgene plasmid #57445; http://n2t.net/addgene:57445; RRID:Addgene_57445) and α- mannosidase II (Addgene plasmid #57467; http://n2t.net/addgene:57467; RRID:Addgene_57467) tagged on their C-termini with mEos3.2 were gifts from Michael Davidson. The plasmids were isolated by midi-prep (catalog no. 12143; QIAGEN, Hilden, Germany) and transfected into U-2 OS cells using Lipofectamine™ 2000 (catalog no. 11668019; Thermo Fisher, Waltham, MA, USA) according to the manufacturer's instructions. Briefly, cells were seeded on coverslips as described in the previous section, after 2 days the medium was replaced with OptiMEM™ (catalog no. 51985026, Thermo Fisher) and the transfection solution was added dropwise.

To prepare the transfection solution for 1 well (2mL of medium), in a first step 1 μg of plasmid was added to 50 μL of OptiMEM™ medium and 3 μL of Lipofectamine™ were added to 50 μL of OptiMEM™ medium, respectively. The two solutions were mixed individually by pipetting, incubated for 3min, and mixed together by pipetting to constitute the transfection solution after further incubation for 5 to 10 min. After 24 h, the OptiMEM™ medium was replaced by normal growth medium and the cells were grown for another 24 h before imaging.

Preparation of microtubule samples.—For microtubule staining, wild-type U-2 OS cells (ATCC HTB-96) were prefixed for 2min with 0.3 % (v/v) glutaraldehyde in cytoskeleton buffer (CB, 10mM MES pH 6.1, 150mM NaCl, 5mM EGTA, 5mM glucose, 5mM $MgCl_2$) + 0.25% (v/v) Triton X-100 and fixed with 2% (v/v) glutaraldehyde in CB for 10min. Fluorescent background was reduced by incubation with 0.1 % (w/v) $NaBH_4$ in PBS for 7min. After the samples had been washed three times with PBS, microtubules were stained with anti-α-tubulin (MS581; NeoMarkers, Fremont, CA, USA), and for ultra-high labeling (Fig. 4g) additionally with anti-β-tubulin (T5293; Sigma-Aldrich), each diluted 1:50 in PBS with 2% (w/v) BSA, overnight. After being washed three times with PBS, samples were incubated with anti-mouse Alexa Fluor 647 (A21236; Invitrogen, Carlsbad, CA, USA) 1:50 in PBS + 2% (w/v) BSA for 6h. After being washed three times with PBS, samples were imaged in blinking buffer as described below. The holder was sealed with parafilm.

Localization microscopy

Microscope setup—SMLM data were acquired on a custom built widefield setup described previously[46, 47]. Briefly, the free output of a commercial laser box (LightHub, Omicron-Laserage Laserprodukte) equipped with Luxx 405, 488 and 638 and Cobolt 561 lasers and an additional 640nm booster laser (iBeam Smart, Toptica) were coupled into a square multi-mode fiber (catalog no. M103L05). The fiber was agitated as described in Ref. 48. The output of the fiber was magnified by an achromatic lens and focused into the sample to homogeneously illuminate an area of about $700μm^2$. The laser was guided through a laser cleanup filter (390/482/563/640 HC Quad, AHF) to remove fluorescence generated by the fiber. The emitted fluorescence was collected through a high numerical aperture (NA) oil immersion objective (HCX PL APO 160×/1.43 NA, Leica), filtered with a 676/37 (catalog no. FF01-676/37-25, Semrock) bandpass filter (for imaging of Alexa Fluor 647) or with a 600/60 (catalog no. NC458462, Chroma) bandpass filter (for live-cell imaging of mMaple and mEos3.2) on an EMCCD camera (Evolve 512, Photometrics). Astigmatism was introduced by a cylindrical lens (f = 1.00m; catalog no. LJ1516L1-A, Thorlabs) to determine the z coordinates of fluorophores. The z focus was stabilized by an infrared laser that was totally internally reflected off the coverslip onto a quadrant photodiode, which was coupled into closed-loop feedback with the piezo objective positioner (Physik Instrumente). Laser control, focus stabilization and movement of filters was performed using a field-programmable gate array (Mojo, Embedded Micro). The pulse length of the 405nm laser could be controlled by a feedback algorithm to sustain a predefined number of localizations per frame.

Imaging conditions—Coverslips containing prepared samples were placed into a custom-built sample holder and 500 μL of blinking buffer (50 mM Tris/HCl pH 8, 10mM NaCl, 10% (w/v) d-glucose, 500μgmL^{-1} glucose oxidase, 40μgmL^{-1} catalase, 35mM MEA) was added for imaging of Alexa Fluor 647 samples.

For imaging of microtubules at different activation densities (Fig. 4 a), we used an exposure time of 15ms and an excitation intensity at 640nm of 15.5 kWcm^{-2}. We adjusted the UV pulse length to result in the desired density of activated fluorophores. As we started with the highest density, by the time we imaged the lowest density a large fraction of the fluorophores was bleached so that we could operate in the single-emitter regime.

For imaging microtubules with ultra-high labeling, we used an exposure time of 15ms and an excitation intensity at 640nm of 13.4 kWcm^{-2} and no UV activation.

For live-cell imaging of Calnexin-mEos3.2 and MannII-mEos3.2 (Fig. 4d and e), the coverslips were washed briefly in PBS and subsequently mounted in 50mM Tris/HCl pH 8 in 95 % (v/v) D$_2$O. The data were acquired with an exposure time of 15ms, an excitation intensity of 22.6kWcm^{-2} for the 561 nm laser, and a maximum intensity of 42 to 127Wcm^{-2} for the 405 nm laser. The pulse length of the 405nm laser was adjusted manually to maintain a high emitter density and to allow imaging of all fluorophores in the field of view in about 1 min.

For the acquisition of live-cell data of Nup96-mMaple (Fig. 4f), coverslips containing Nup96-mMaple cells[29] (catalog no. 300461; CLS Cell Line Service, Eppelheim, Germany) were rinsed twice with warm PBS before they were mounted in 1mL growth medium containing 20 mM HEPES buffer and imaged directly. During imaging, we used an excitation intensity at 561nm of 16.7kWcm^{-2} and a UV laser power of 80Wcm^{-2}. The exposure time was 12 ms and the pulse length of the UV laser was automatically adjusted from 1 to 12 ms to keep the density of localizations constant.

Extended Data
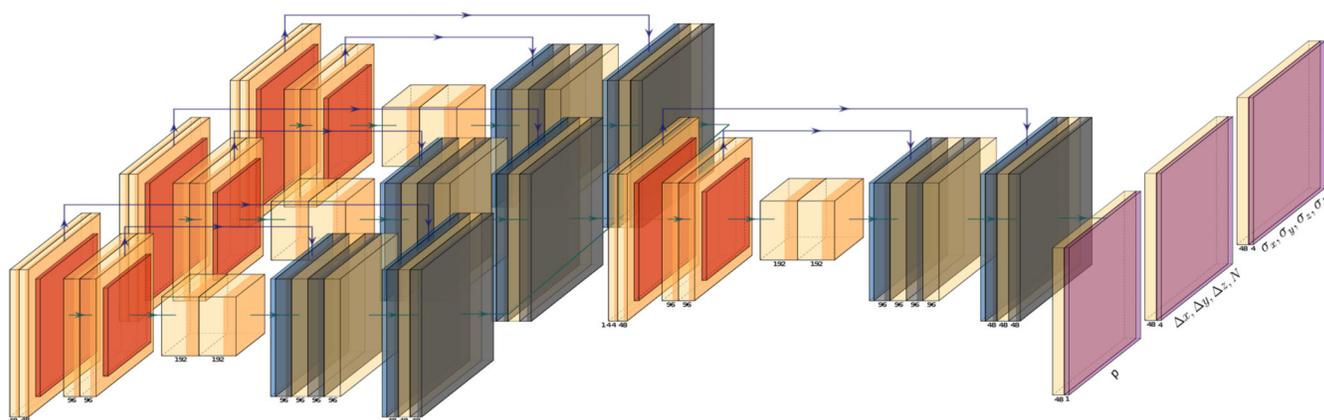
Extended Data Figure. 1. Architecture

The DECODE network consists of two stacked U-Nets19 with identical layouts (the three networks depicted on the left share parameters). The frame analysis module extracts informative features from three consecutive frames. These features are integrated by the temporal context module. Both U-Nets have two up- and downsampling stages and 48 filters in the first stage. Each stage consists of three fully convolutional layers with 3×3 filters. In each downsampling stage, the resolution is halved, and the number of filters is doubled, vice versa in each upsampling stage. Blue arrows show skip connections. Following the temporal context module three output heads with two convolutional layers each produce the output maps which have the same spatial dimensions as the input frames. The first head predicts the Bernoulli probability map p, the second head the spatial coordinates of the detected emitter x, y, z and its intensity N and the third head the associated uncertainties $\sigma x, \sigma y, \sigma z$, σN. An optional fourth output head can be used for background prediction.
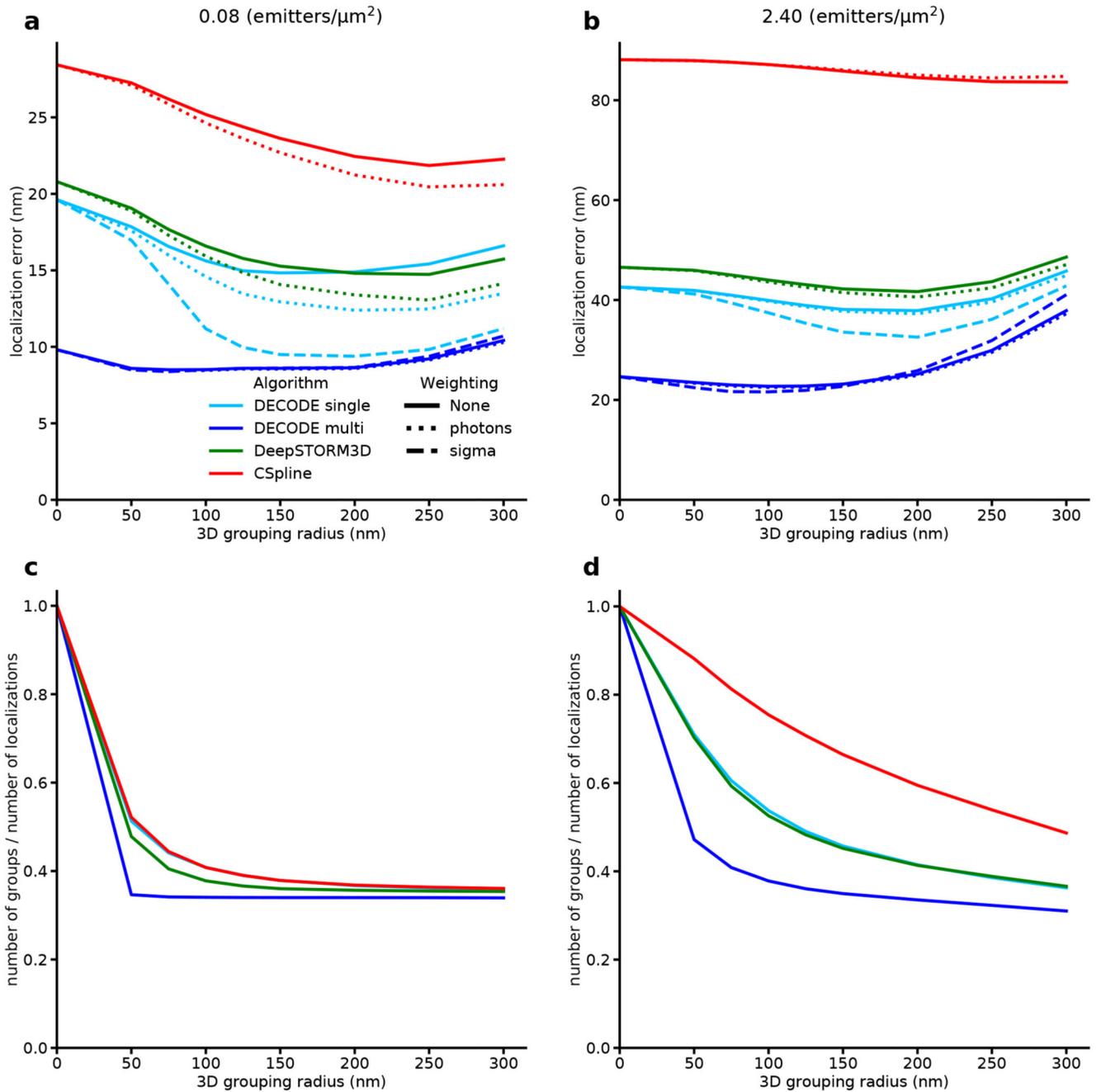
Extended Data Figure. 2. Impact of grouping across grouping radius for different averaging weights.

Predictions in consecutive frames are grouped when they are closer to each other than the given grouping radius. A grouping radius of 0nm corresponds to not performing any grouping. Predictions within a group are assigned a common set of emitter coordinates which is calculated as weighted average of their individual coordinates. We compare three different options for the weighted average: Uniform weighting ('None', solid lines); Weighting by the inferred number of photons for CSpline and DECODE or the inferred confidence for DeepSTORM3D ('photons', dotted line); Weighting by the predicted

DECODE σ values, where the x, y and z values are individually weighted by $\sigma_{x, y, z}^{-2}$.

a, b): 3D efficiencies across grouping radii. Grouping is especially useful in the low density setting (a) where DECODE without temporal context (DECODE single) with a correctly set grouping radius can match the performance of DECODE with temporal context (DECODE multi) without grouping. This is, however, only the case when weighting by the uncertainty estimates that DECODE provides. Using grouping on top of DECODE multi offers little additional benefit. **c, d**): Number of groups divided by the number of localizations. Detecting all emitters and correctly grouping them would result in a ratio of 1 : 3 as on average each emitter is visible in three consecutive frames. See methods and Supplementary Table S1 for additional details on training and evaluation.
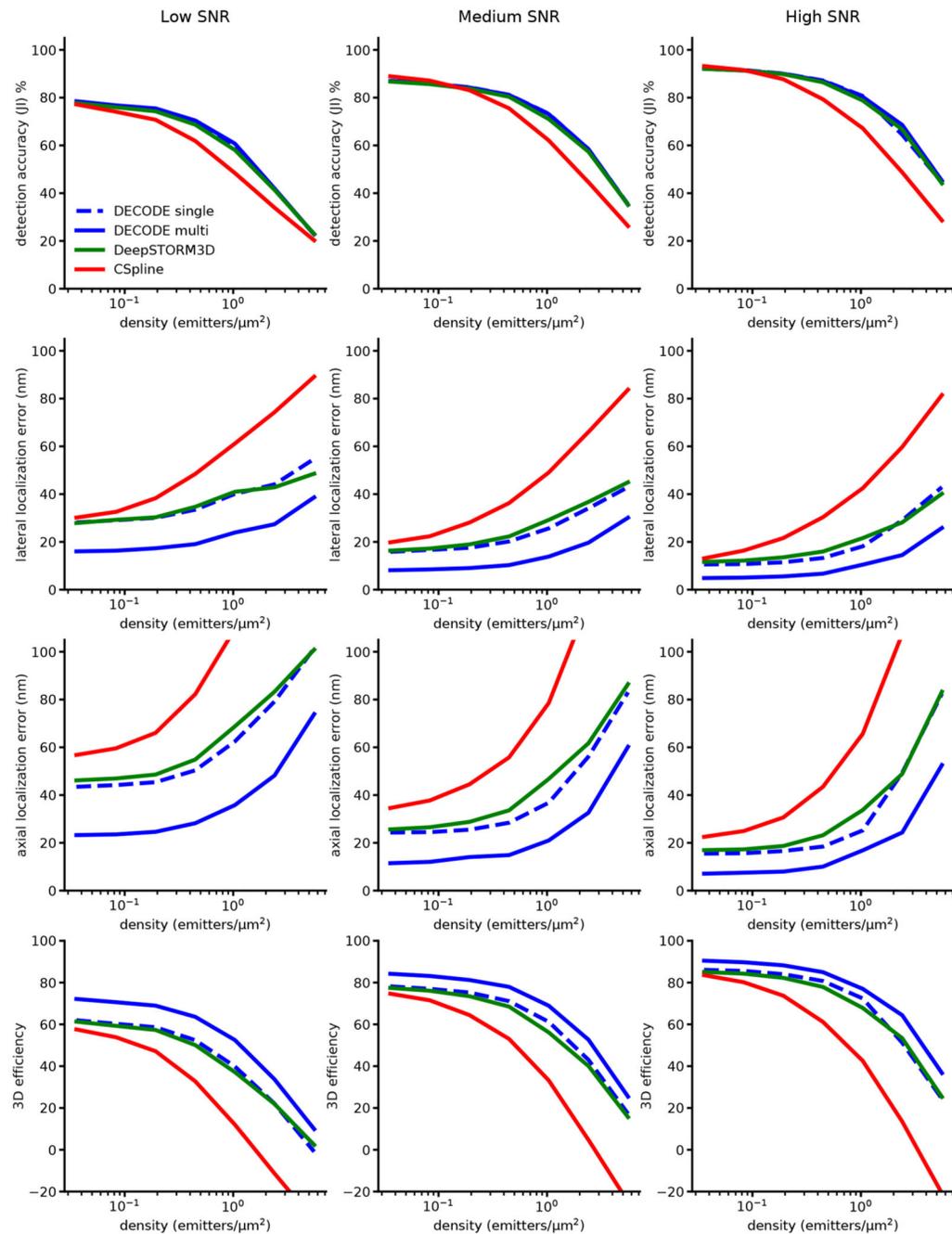
Extended Data Figure. 3. Comparison of performance metrics across densities and SNRs.
DECODE outperforms DeepSTORM3D and CSpline across densities and SNRs. See
methods and Supplementary Table S1 for additional details on training and evaluation.
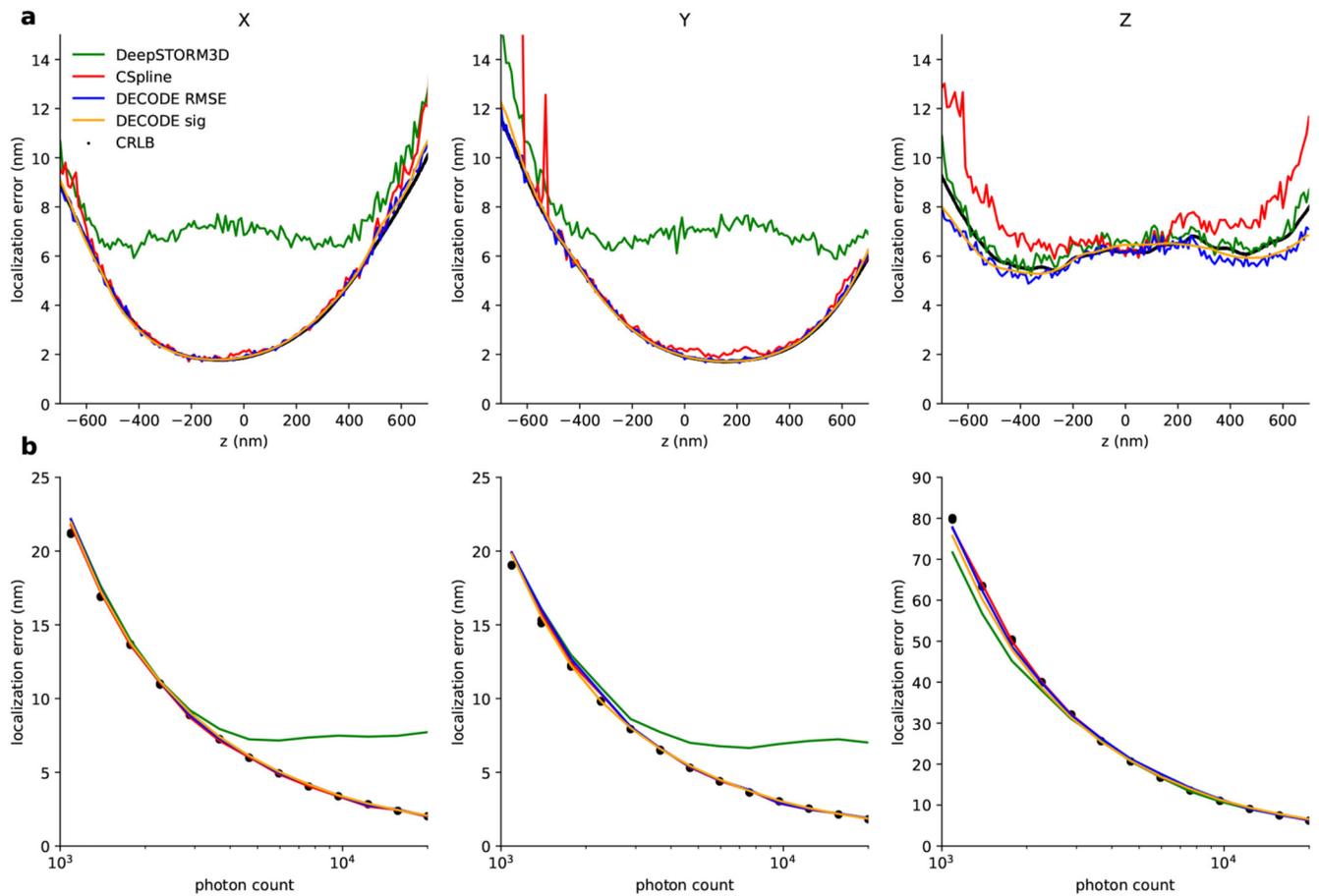
Extended Data Figure. 4. Comparison of localization error and CRLB for single emitter fitting.
The RMSE achieved by DECODE and its predicted σ values closely match the single
emitter CRLB in every dimension. CSpline is also able to achieve the CRLB, which has
been shown for iterative MLE fitters before. In contrast the resolution that DeepSTORM3D
can achieve is limited by its output representation and the size of the super-resolution voxels.
a): Data simulated with high SNR (20000 photons) and random z. RMSE and DECODE σ
averaged over 10 nm bins. **b**): Data simulated with fixed z (0nm) and varying SNR levels.
See methods and Supplementary Table S1 for additional details on training and evaluation.
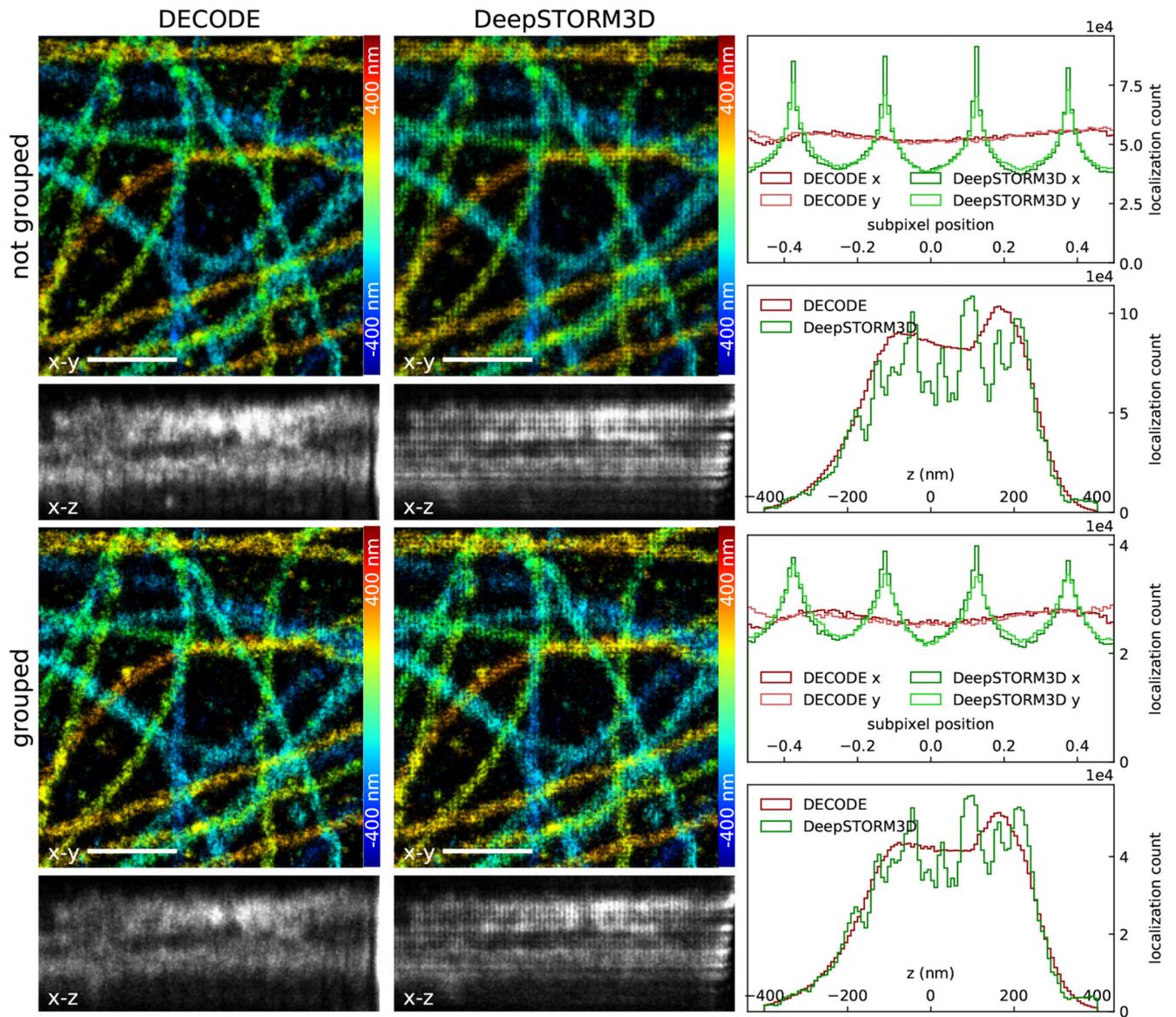
Extended Data Figure. 5. Comparison of reconstruction quality on on experimental STORM data.
Reconstructions by DECODE and the DeepSTORM3D on a subset of data shown in Fig. 4g. Histograms show within pixel distribution of localizations in x and y as well as the z coordinate in n. DeepSTORM3D has 4 significant peaks in the subpixel distribution, corresponding to the fourfold upsampling it uses for its network output. These are visible as grid artifacts in the reconstructions. In contrast the DECODE localizations are evenly distributed and no artifacts are visible. Scale bars 0.5 μm.
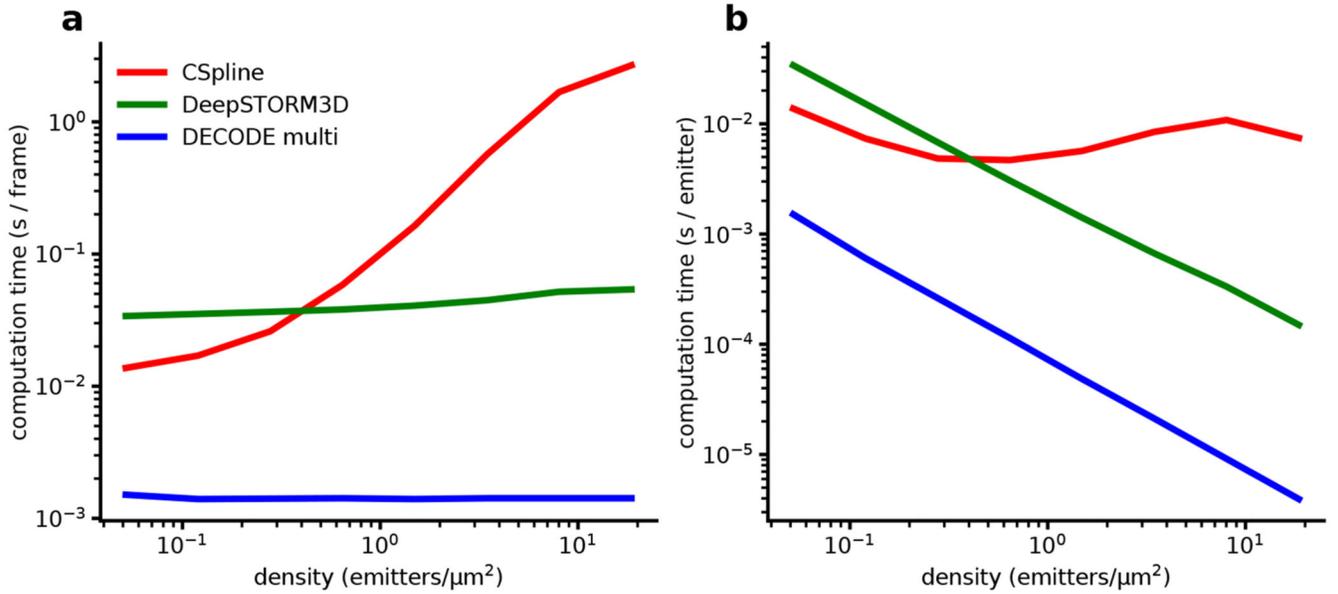
Extended Data Figure. 6. Comparison of computation times.
a). Measured as the time it takes to analyze a 64×64 pixel frame with varying emitter densities. Trained DECODE and DeepSTORM3D models were evaluated using a NVIDIA RTX2080Ti GPU. Computation time includes the network forward pass and post-processing and does not include training time. CSpline was evaluated on an Intel(R)Xeon(R) CPU E5-2697 v3. **b)** Computation time per simulated emitter. The computation time of CSpline scales with the number emitters while the two deep learning based approaches scale with the number (and size) of the analyzed frames. GPU-based DECODE is about 20 times faster than GPU-based DeepSTORM3D and outperforms CPU-based CSpline even at low densities.
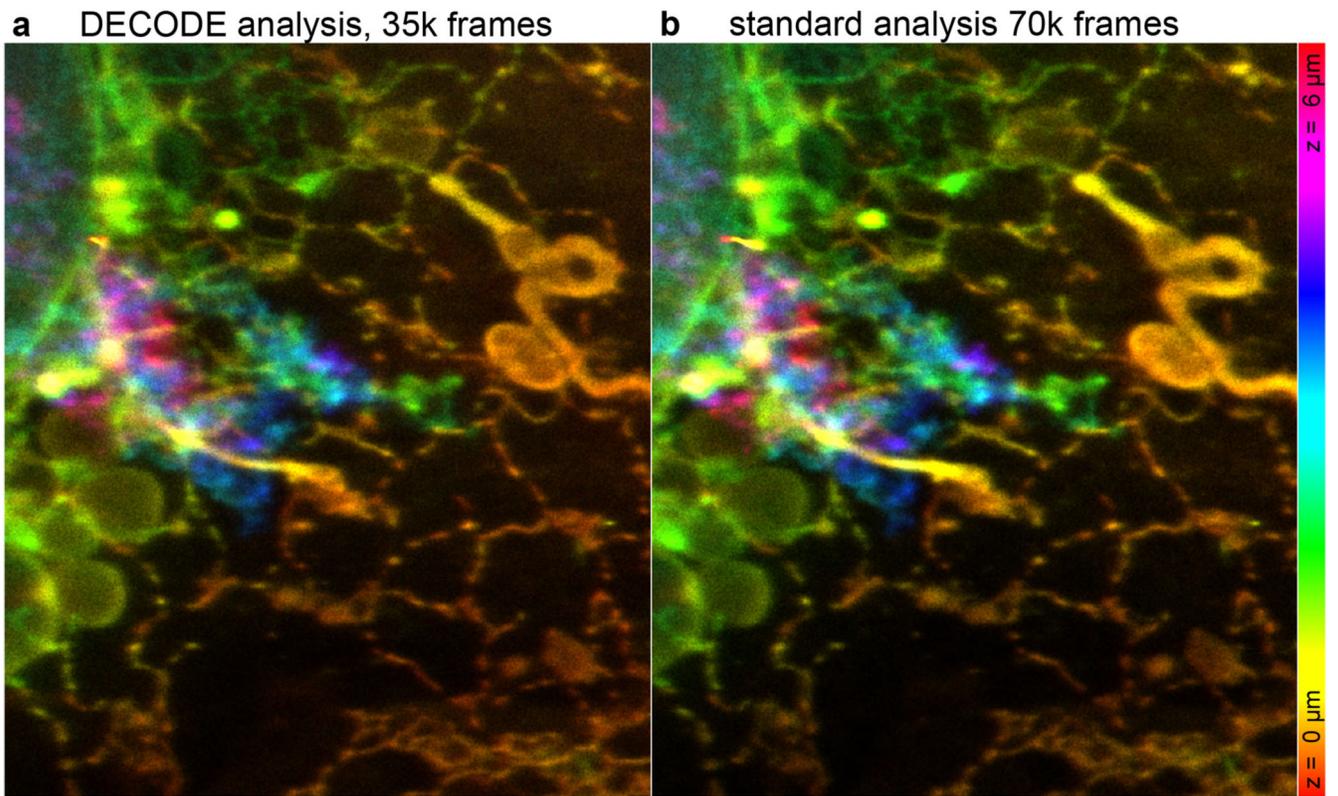
Extended Data Figure. 7. DECODE reduces acquisition times in LLS-PAINT.
DECODE reconstruction of 35,000 frames (**a**) results in the same number of localizations as the Standard reconstruction of 70,000 frames (**b**). As DECODE detects twice as many localizations as the traditional analysis, it needs only approx. half of the frames for a high-quality reconstruction.
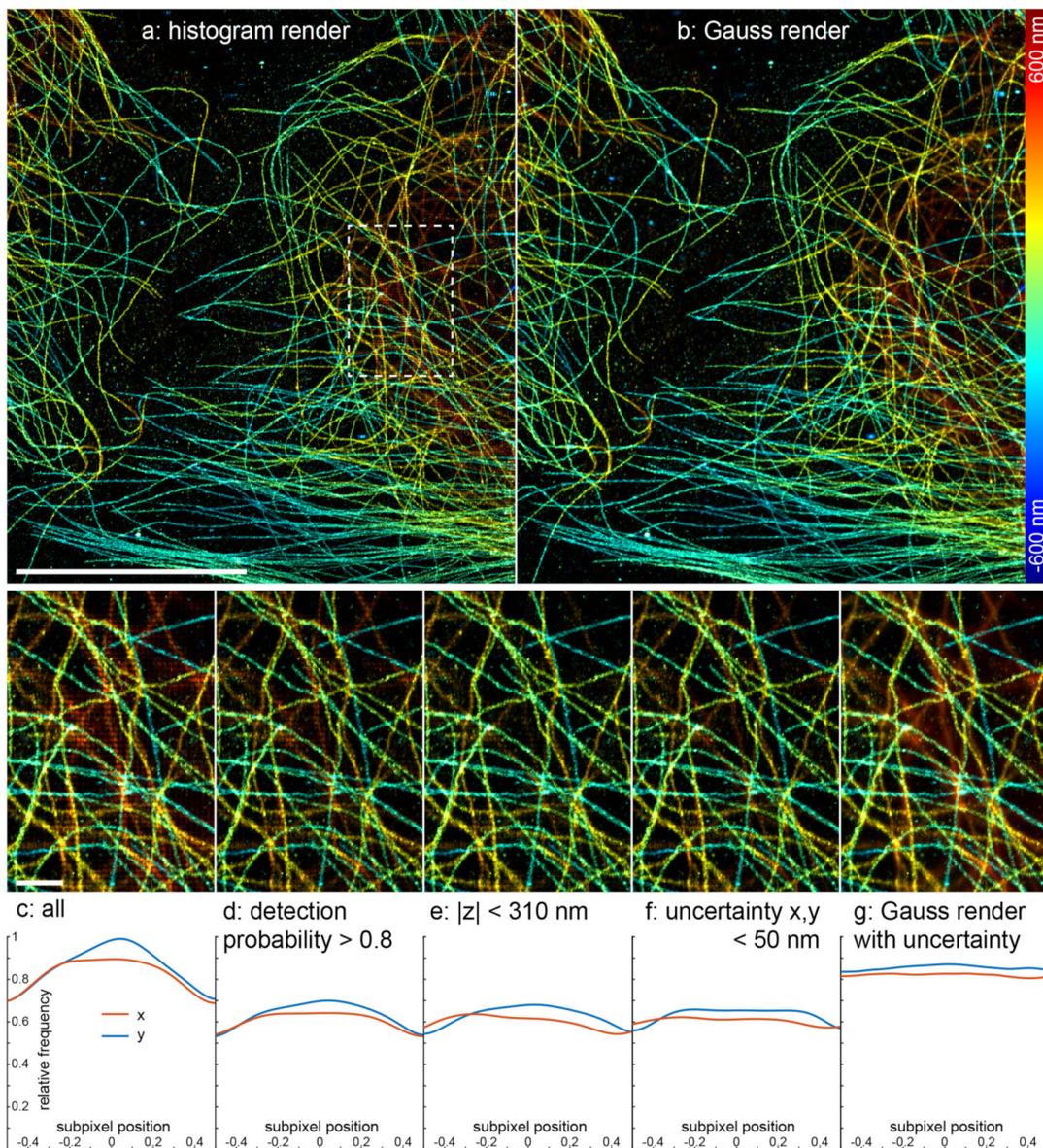
Extended Data Figure. 8. Removing Pixelation artifacts.
Dim, dense out-of-focus localizations have a bias towards the pixel center (a,c). This is apparent as a non-uniform distribution of the sub-pixel positions in x and y (bottom row). This bias is not visible if every localization is rendered as a Gaussian with a standard deviation equal to the predicted uncertainty σ (b,g). Filtering according to the detection probability reduces the artifact (d). Filtering according to the predicted uncertainty σ (f) or the fluorophore z-position (e) also removes the pixelation artifact. Scale bars 10μm (a,b) and 1 μm (c-g).
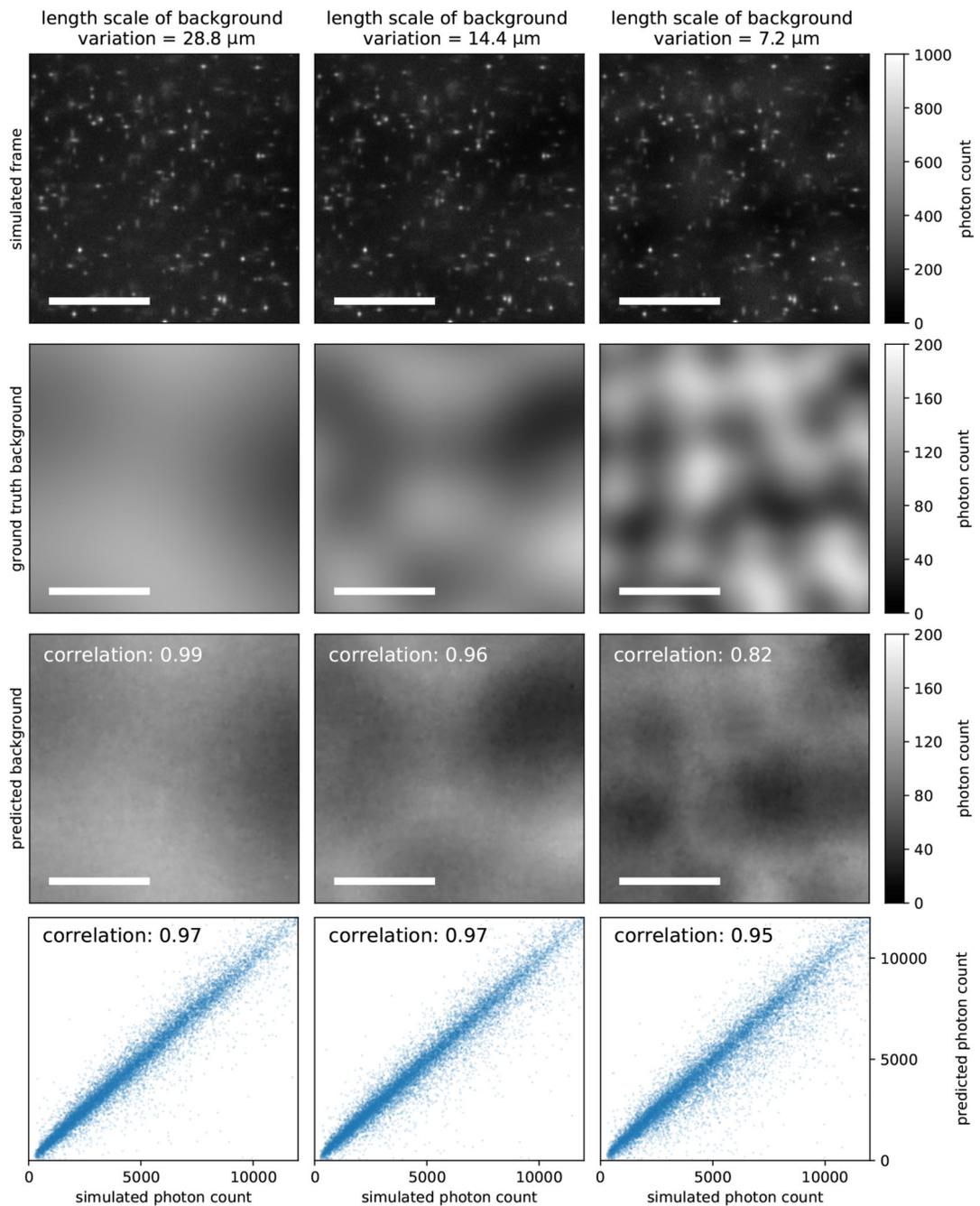
Extended Data Figure. 9. DECODE provides accurate background and signal predictions on simulated data with inhomogeneous background of various length scales.

First row: sample frames. Second row: background values simulated using Perlin noise[14, 49]. Third row: background values inferred by a DECODE network that was trained on 40x40 pixel sized simulations with uniform background. Fourth row: Scatter plot of inferred photon counts over simulated photon counts. Scale bars are 10 μm.
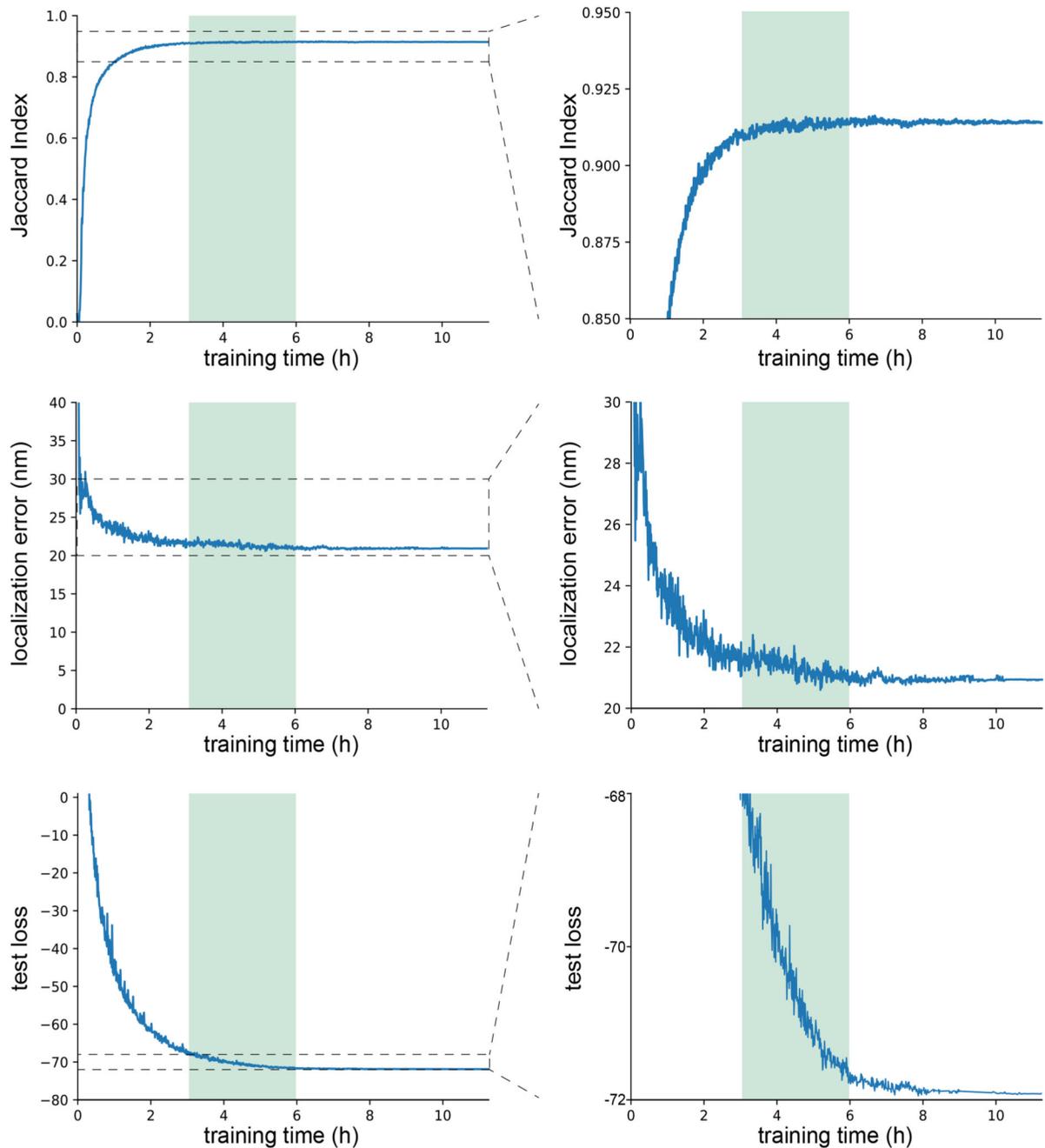
Extended Data Figure. 10. Performance as a function of deep network training time. Convergence of the accuracy of DECODE for several performance metrics. Runtimes are measured on a single nVidia RTX 2080 Ti GPU. The estimated training achievable with the maximum of 12 hours possible on the free tier of Google Colab is shown in green range (assuming that a Google Colab GPU is 2x-4x slower than the nVidia RTX 2080 Ti GPU). This suggests that acceptable performance is achievable using DECODE and Google Colab at minimal cost, no GPU needed. Metrics evaluated for prediction > 0.5 detection probability

estimate without sigma filtering. Training data was simulated at high SNR (as described in Figure 2c) at an average density of 1 μm^{-2}.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Data availability

All data can be publicly accessed from the DECODE webpage (see Code Availability).

Code Availability

DECODE is available at https://github.com/TuragaLab/DECODE.

References

1. Betzig E, et al. Imaging intracellular fluorescent proteins at nanometer resolution. Science. 2006; 313 :1642–1645. [PubMed: 16902090]

2. Rust MJ, Bates M, Zhuang X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). Nat methods. 2006; 3 :793. [PubMed: 16896339]

3. Babcock HP, Zhuang X. Analyzing single molecule localization microscopy data using cubic splines. Sci reports. 2017; 7

4. Babcock H, Sigal YM, Zhuang X. A high-density 3d localization algorithm for stochastic optical reconstruction microscopy. Opt Nanoscopy. 2012; 1 :6.

5. Ovesný M, K̆rížek P, Borkovec J, Švindrych Z, Hagen GM. Thunderstorm: a comprehensive imagej plug-in for palm and storm data analysis and super-resolution imaging. Bioinformatics. 2014; 30 :2389–2390. [PubMed: 24771516]

6. Sage D, et al. Super-resolution fight club: assessment of 2d and 3d single-molecule localization microscopy software. Nat methods. 2019; 16 :387–395. [PubMed: 30962624]

7. Belthangady C, Royer LA. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. Nat methods. 2019 :1–11. [PubMed: 30573832]

8. Ching T, et al. Opportunities and obstacles for deep learning in biology and medicine. J The Royal Soc Interface. 2018; 15 20170387

9. Weigert M, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. Nat methods. 2018; 15 1090 [PubMed: 30478326]

10. Le, TA; Baydin, AG; Zinkov, R; Wood, F. Using synthetic data to train neural networks is model-based reasoning; Neural Networks (IJCNN), 2017 International Joint Conference; 2017. 3514–3521.

11. Möckl L, Roy AR, Moerner W. Deep learning in single-molecule microscopy: fundamentals, caveats, and recent developments. Biomed Opt Express. 2020; 11 :1633–1661. [PubMed: 32206433]

12. Zhang P, et al. Analyzing complex single-molecule emission patterns with deep learning. Nat methods. 2018; 15 :913–916. [PubMed: 30377349]

13. Kim T, Moon S, Xu K. Information-rich localization microscopy through machine learning. Nat communications. 2019; 10 :1–8.

14. Möckl L, Roy AR, Petrov PN, Moerner W. Accurate and rapid background estimation in single-molecule localization microscopy using the deep neural network bgnet. Proc Natl Acad Sci. 2020; 117 :60–67. [PubMed: 31871202]

15. Zelger P, et al. Three-dimensional localization microscopy using deep learning. Opt express. 2018; 26 :33166–33179. [PubMed: 30645473]

16. Nehme E, et al. Deepstorm3d: dense 3d localization microscopy and psf design by deep learning. Nat Methods. 2020 :1–7. [PubMed: 31907477]

17. Boyd N, Jonas E, Babcock HP, Recht B. Deeploco: Fast 3d localization microscopy using neural networks. bioRxiv. 2018; doi: 10.1101/267096

18. Chen B-C, et al. Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. Science. 2014; 346 1257998 [PubMed: 25342811]

19. Ronneberger, O; Fischer, P; Brox, T. U-net: Convolutional networks for biomedical image segmentation; Int Conf on Med image computing computer-assisted intervention; 2015. 234–241.

20. Rieger B, Stallinga S. The lateral and axial localization uncertainty in super-resolution light microscopy. ChemPhysChem. 2014; 15 :664–670. [PubMed: 24302478]

21. Chao J, Ward ES, Ober RJ. Fisher information theory for parameter estimation in single molecule microscopy: tutorial. JOSA A. 2016; 33 :B36–B57. [PubMed: 27409706]

22. Li Y, et al. Real-time 3d single-molecule localization using experimental point spread functions. Nat methods. 2018; 15 :367. [PubMed: 29630062]

23. Small A, Stahlheber S. Fluorophore localization algorithms for super-resolution microscopy. Nat methods. 2014; 11 :267. [PubMed: 24577277]

24. Van de Linde S, et al. Direct stochastic optical reconstruction microscopy with standard fluorescent probes. Nat protocols. 2011; 6 :991. [PubMed: 21720313]

25. Rust MJ, Bates M, Zhuang X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). Nat methods. 2006; 3 :793–796. [PubMed: 16896339]

26. Nieuwenhuizen RP, et al. Measuring image resolution in optical nanoscopy. Nat methods. 2013; 10 :557–562. [PubMed: 23624665]

27. Diekmann R, et al. Optimizing imaging speed and excitation intensity for single-molecule localization microscopy. Nat Methods. 2020 :1–4. [PubMed: 31907477]

28. Wäldchen S, Lehmann J, Klein T, Van De Linde S, Sauer M. Light-induced cell damage in live-cell super-resolution microscopy. Sci reports. 2015; 5 15348

29. Thevathasan JV, et al. Nuclear pores as versatile reference standards for quantitative superresolution microscopy. Nat Methods. 2019; 16 :1045–1053. [PubMed: 31562488]

30. Dempsey GT, Vaughan JC, Chen KH, Bates M, Zhuang X. Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. Nat methods. 2011; 8 1027 [PubMed: 22056676]

31. Legant WR, et al. High-density three-dimensional localization microscopy across large volumes. Nat methods. 2016; 13 :359–365. [PubMed: 26950745]

32. Paszke A, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems. 2019 :8024–8035.

33. Ries J. Smap: a modular super-resolution microscopy analysis platform for smlm data. Nat Methods. 2020 :1–2. [PubMed: 31907477]

34. Chamier LV, et al. ZeroCostDL4Mic: an open platform to simplify access and use of Deep-Learning in Microscopy. bioRxiv. 2020; doi: 10.1101/2020.03.20.000133

35. Odena A, Dumoulin V, Olah C. Deconvolution and checkerboard artifacts. Distill. 2016; doi: 10.23915/distill00003

36. Clevert D-A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint. 2015 arXiv:1511.07289

37. Ouyang W, Aristov A, Lelek M, Hao X, Zimmer C. Deep learning massively accelerates super-resolution localization microscopy. Nat biotechnology. 2018

38. Weigert M, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. Nat methods. 2018; 15 1090 [PubMed: 30478326]

39. Annibale P, Vanni S, Scarselli M, Rothlisberger U, Radenovic A. Quantitative photo activated localization microscopy: unraveling the effects of photoblinking. PloS one. 2011; 6 e22678 [PubMed: 21818365]

40. Huang F, et al. Video-rate nanoscopy using scmos camera-specific single-molecule localization algorithms. Nat methods. 2013; 10 :653–658. [PubMed: 23708387]

41. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint. 2017 arXiv:1711.05101

42. Banterle N, Bui KH, Lemke EA, Beck M. Fourier ring correlation as a resolution criterion for super-resolution microscopy. J structural biology. 2013; 183 :363–367.

43. Bornschein J, Bengio Y. Reweighted wake-sleep. CoRR abs/1406.2751. 2014 1406.2751

44. Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint. 2014 arXiv:1401.4082

45. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint. 2013 arXiv:1312.6114

46. Mund M, et al. Systematic nanoscale analysis of endocytosis links efficient vesicle formation to patterned actin nucleation. Cell. 2018; 174 :884–896. [PubMed: 30057119]

47. Deschamps J, Rowald A, Ries J. Efficient homogeneous illumination and optical sectioning for quantitative singlemolecule localization microscopy. Opt express. 2016; 24 :28080–28090. [PubMed: 27906373]

48. Schröder D, Deschamps J, Dasgupta A, Matti U, Ries J. Cost-efficient open source laser engine for microscopy. Biomed Opt Express. 2020; 11 :609–623. [PubMed: 32206389]

49. Perlin K. An image synthesizer. ACM Siggraph Comput Graph. 1985; 19 :287–296.
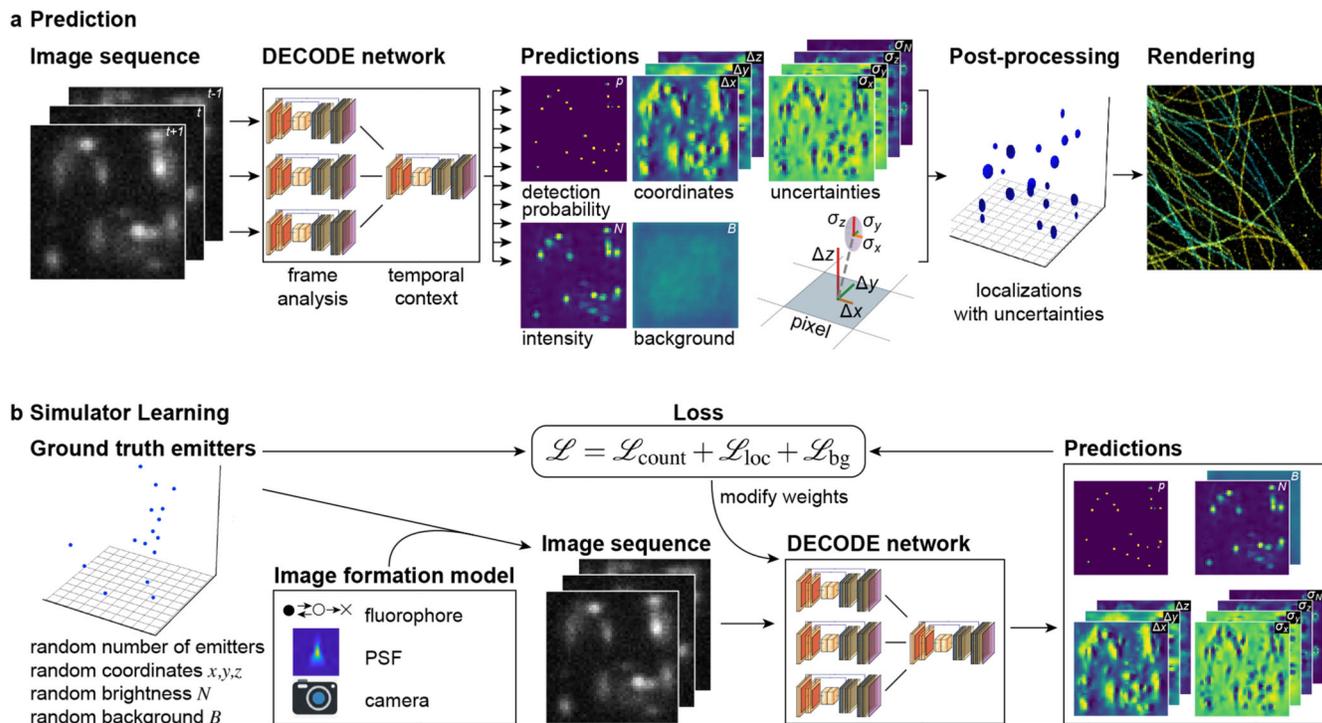
Figure 1. DECODE for high-density single molecule localization.

a) DECODE architecture. The DECODE network uses information from multiple frames to predict detection probabilities, coordinates and uncertainty estimates. The *frame analysis module* with a multi-scale 'U-Net' architecture[19] extracts informative features from each frame. These features are integrated by the *temporal context module* which produces 9 output maps: a map of emitter detection probabilities p, a map predicting the brightness of the corresponding detected emitter N, three maps of the three spatial coordinates of the detected emitter x, y, z (relative to the to the center of the detected pixel) and four maps of the associated uncertainties (standard deviations) σN, σx, σy, σz. In addition, we optionally predict a map with the background intensity B in each pixel. **b) Training DECODE.** The DECODE network is trained by simulator-learning: Ground truth (GT) emitter coordinates are generated randomly and a forward-model of the image formation process is used to simulate synthetic images. These simulated images are passed through the DECODE network. The loss quantifies the probability that the GT explains the output predictions. This probability is maximized during training. While the network only uses camera images to make predictions, the network training procedure does require PSF calibration measurements.
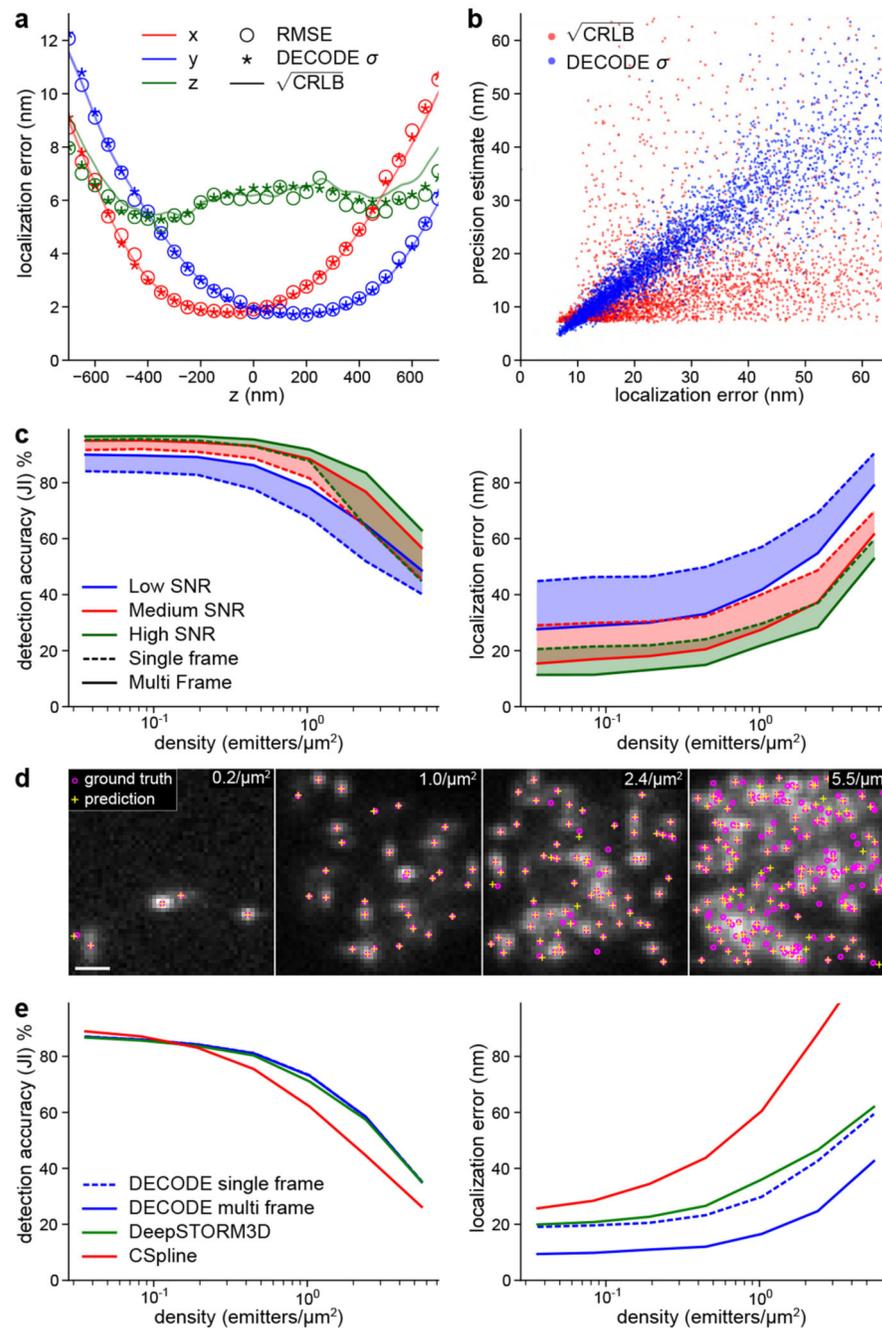
Figure 2. Performance of DECODE on simulated data

a) DECODE reaches the single emitter Cramér-Rao Lower Bound (CRLB) for isolated emitters. Root Mean Squared Error (RMSE) and DECODE σ averaged over 50 nm bins. See Extended Data Fig. 4 for additional comparisons with other methods. **b) The predicted localization uncertainty σ correlates well with the measured localization error for densely activated emitters.** We simulated the same dense emitter configuration 100 times and calculated the measured localization error as the RMSE of the predictions of the coordinates. In comparison, the (square-rooted) single-emitter-CRLB incorrectly under-

estimates the true localization error for high emitter densities. See Supplementary Fig. S1 for comparisons of individual dimensions. **c) Temporal context improves both detection performance and localization error.** We trained DECODE with (multi frame) and without temporal context (single frame) and compared detection accuracy and localization error for low, medium and high Signal to Noise Ratio (SNR). **d) Representative simulated frames** with ground truth coordinates (magenta circles) and predicted coordinates (yellow crosses) for the densities used in c and medium SNR. **e) Comparison of DECODE with CSpline and DeepSTORM 3D.** DECODE outperforms DeepSTORM and CSpline over a wide range of densities. See Extended Data Fig. 2 and 3 for additional comparisons with different conditions and metrics. The Standard Error of the Mean (SEM) on the localization error lies between 0.2 and 0.4 nm. See methods and Supplementary Table S1 for additional details on training and evaluation.
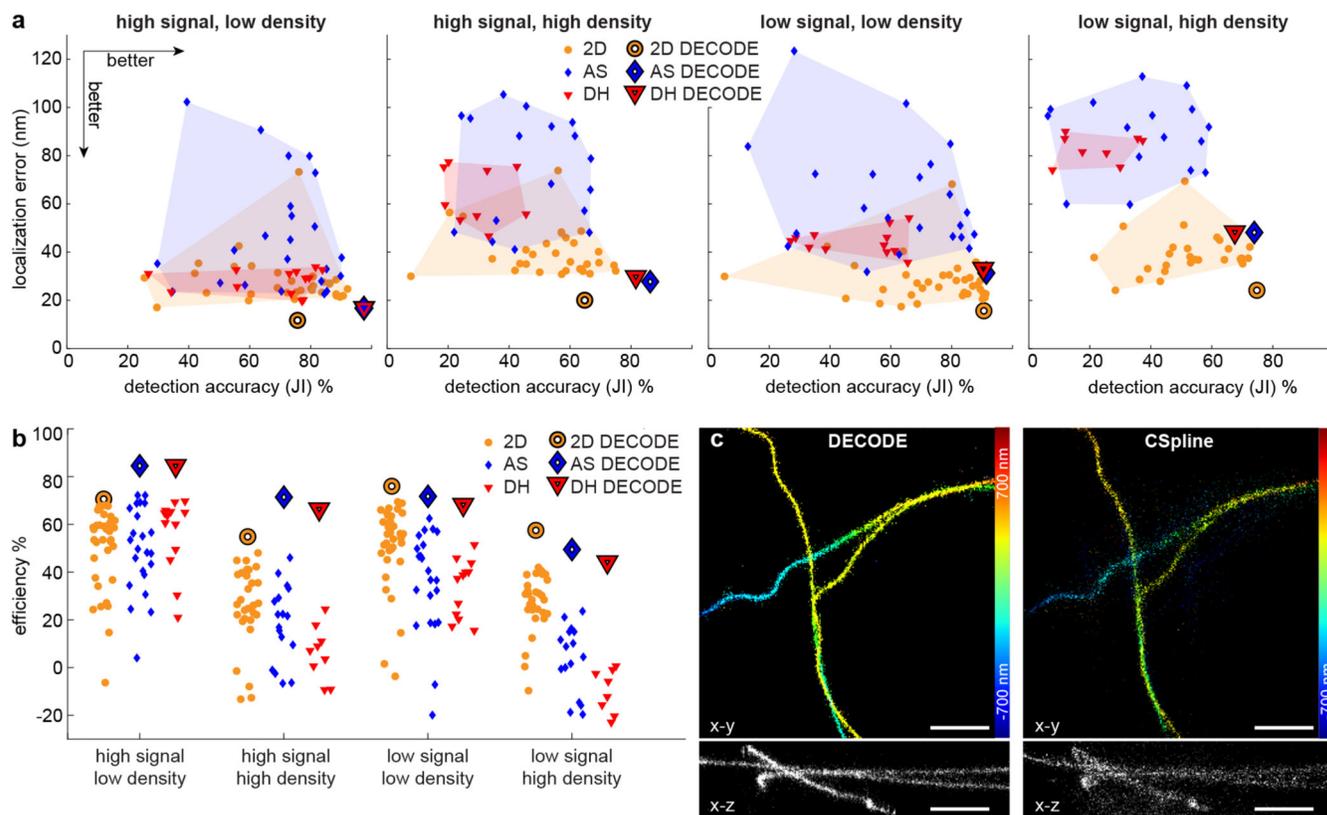
Figure 3. Performance comparison on the SMLM 2016 challenge.
a) Performance evaluation on the twelve test datasets with low/high density, low/high SNR and different modalities (2D, AS: astigmatic, DH: double helix) using the detection accuracy (Jaccard Index, JI, higher is better) and localization error (lower is better) as metrics. Each marker indicates a benchmarked algorithm, large markers indicate DECODE. **b)** Efficiency scores (higher is better) for DECODE compared to other algorithms. Colored dots indicate performance numbers for other methods. All metrics were calculated by the SMLM 2016 challenge and downloaded from the challenge website[†]. **c)** Reconstructions by DECODE and the CSpline algorithm on the high density, low signal double helix challenge training data. Upper panels x-y view, color coded by z coordinate, lower panels x-z reconstructions. Scale bars 1 μm. See Extended Data Fig. S5 and S6 for additional comparisons with DeepSTORM3D on training datasets.
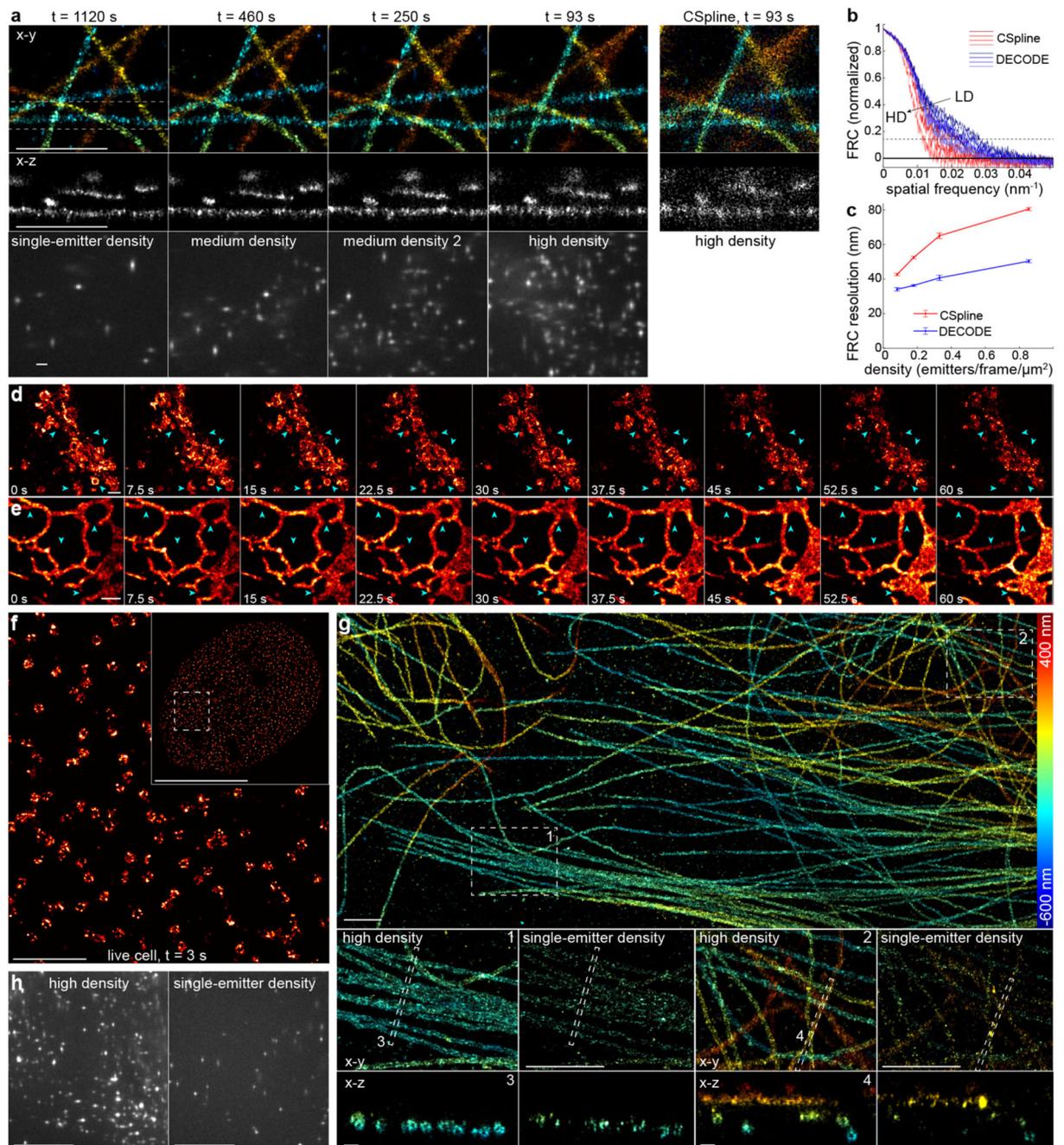
Figure 4. DECODE enables high-speed and live-cell SMLM and ultra-high labeling densities.
a) DECODE can reduce acquisition times by one order of magnitude. The same sample of microtubules, labeled with anti-α-tubulin primary and AF647 secondary antibodies, imaged with different UV activation intensities to result in different emitter densities per frame, between 0.08 and $0.86 \mu m^{-2}$ and acquisition times between 93 and 1120 s, while keeping the total number of localizations the same. For high-density activation, we show a comparison with CSpline. **b)** Fourier Ring Correlation curves for DECODE and CSpline for different emitter densities. **c)** Resolution estimates obtained using the Fourier Ring

Correlation and 0.143 criterion across densities for both methods. **d) Fast live-cell SMLM on the Golgi apparatus** labeled with *a*-mannosidase II-mEos3.2. See Supplementary Movie 1. **e) Fast live-cell SMLM on the endoplasmic reticulum** labeled with calnexin-mEos3.2. See Supplementary Movie 2 and Supplementary Fig. S3. **f) Fast live-cell SMLM on the nuclear pore complex protein** Nup96-mMaple acquired in 3 seconds. **g) DECODE enables ultra-high labeling densities.** Microtubules labeled with a high concentration of anti-α and anti-β-tubulin primary and Alexa Fluor 647 secondary antibodies. **g1, g2)** Magnified regions as indicated in g. Data acquired with high-density labeling shows continuous structures. As a comparison, the same sample was acquired after pre-bleaching of the fluorophores to reach the single-molecule blinking regime. Here, single labels are resolved in the superresolution reconstruction and lead to a sparse decoration of the microtubules. **g3, g4)** Side view reconstructions of regions as indicated in g1, g2 resolving the hollow, cylinder-like structure of immunolabeled microtubules. **h)** Representative raw camera frames for the high-density and single-emitter acquisitions, respectively. Scale bars: 10μm (f inset, h), 1 μm (a, d, e, f, g, g1, g2), 100nm (g3, g4).
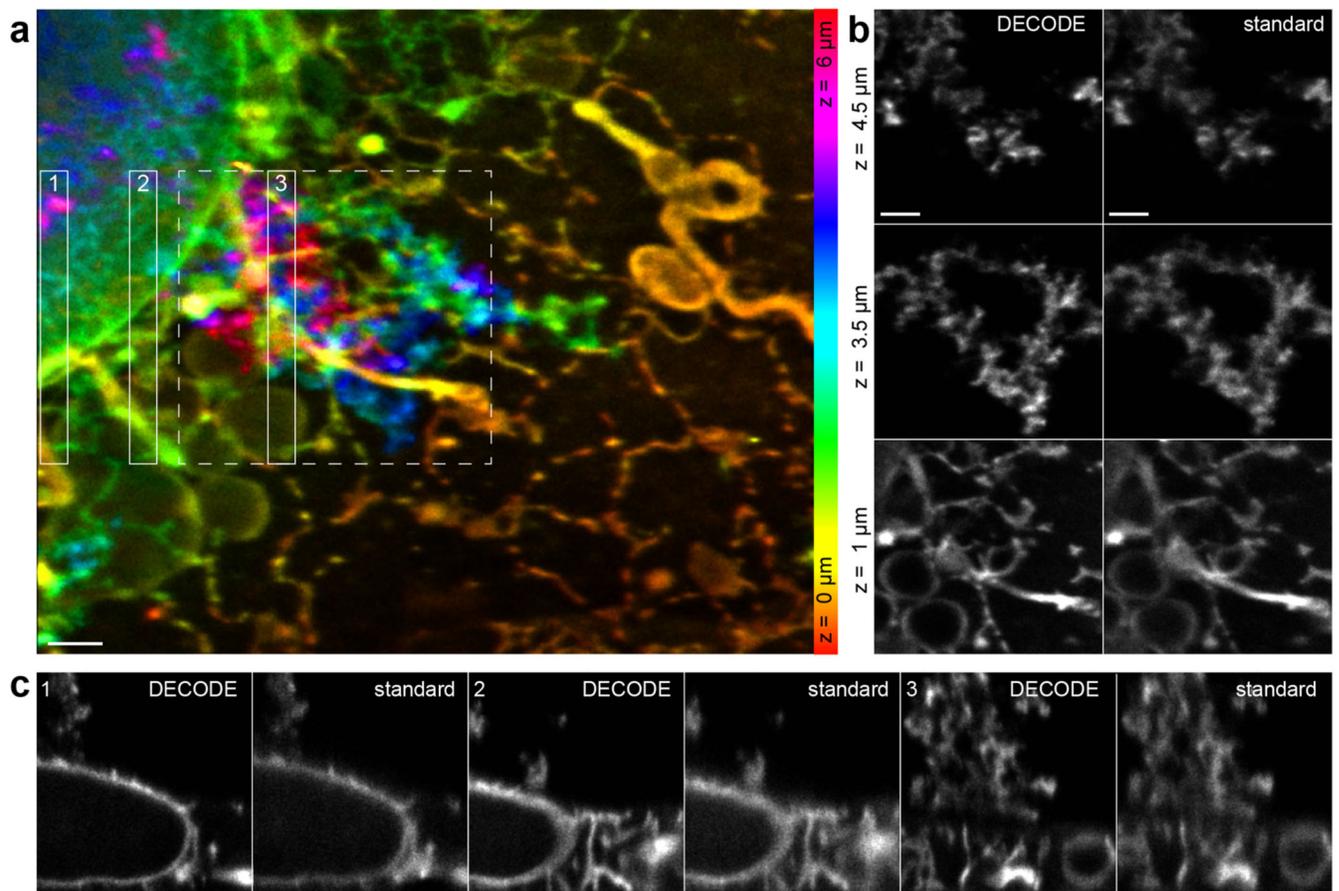
Figure 5. DECODE improves resolution in LLS-PAINT.
a) COS-7 cell imaged with LLS-PAINT microscopy, overview. Data from Legant et al.[31], 70,000 volumes imaged over 2.7 days. **b)** 500nm thick slices of the region indicated in a (dashed line), comparing DECODE analysis and the original analysis using MLE fitting (standard analysis). **c)** Perpendicular (side-view) reconstructions of 500nm thick regions as indicated in a comparing DECODE and standard analysis. Scale bars 1 μm. See Extended Data Fig. 7 for additonal comparisons.