LocNet: Deep learning-based localization on rotating point spread function with applications to telescope imaging

LINGJIA DAI,[1] MINGDA LU,[1] CHAO WANG,[2,3,6] SUDHAKAR PRASAD,[4] AND RAYMOND CHAN[1,5,7]

[1]*Department of Mathematics, City University of Hong Kong, Hong Kong SAR, China*
[2]*Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518005, Guangdong Province, China*
[3]*National Centre for Applied Mathematics Shenzhen, Shenzhen 518055, Guangdong Province, China*
[4]*School of Physics and Astronomy, University of Minnesota, USA*
[5]*Hong Kong Centre for Cerebro-Cardiovascular Health Engineering, Hong Kong SAR, China*
[6]*wangc6@sustech.edu.cn*
[7]*raymond.chan@cityu.edu.hk*

Abstract: Three-dimensional (3D) point source recovery from two-dimensional (2D) data is a challenging problem with wide-ranging applications in single-molecule localization microscopy and space-debris localization telescopy. Point spread function (PSF) engineering is a promising technique to solve this 3D localization problem. Specifically, we consider the problem of 3D localization of space debris from a 2D image using a rotating PSF where the depth information is encoded in the angle of rotation of a single-lobe PSF for each point source. Instead of applying a model-based optimization, we introduce a convolution neural network (CNN)-based approach to automatically localize space debris in full 3D space. A hard sample training strategy is proposed to further improve the performance of CNN. Contrary to the traditional model-based methods, our technique is efficient and outperforms the current state-of-the-art method by more than 11% in the rate of precision with a comparable improvement in the rate of recall.

1. Introduction

Three-dimensional localization of point sources is an indispensable part of applications in different fields. One example is 3D single-molecule localization microscopy (SMLM), which localizes individual fluorophores in 3D structures to render super-resolution images or facilitate analyses of fluorescent molecules [1–6]. When the images of individual fluorophores do not overlap, the coordinates of each fluorophore can be located with high precision. By activating and imaging subsets of fluorophores with non-overlapping images from which a composite image is reconstructed, SMLM overcomes the diffraction limit and can image biological structures near the molecular scale. Another example is detecting and localizing space debris in the vicinity of a space asset, such as a satellite, by using active illumination and 3D imaging modules mounted on it. The inclination and altitude of operational satellite orbits can be changed, if necessary, to mitigate collisions with space debris if the latter can be 3D-tracked accurately. Active space debris removal will, in fact, become increasingly critical as space technology becomes increasingly widely used. According to the NASA Orbital Debris Program Office, currently, there are more than 26,000 objects in orbit around Earth, including both operational and defunct satellites and other human-made debris [7]. The natural decay of space debris can take months to years, a rate that is dwarfed by the typical rates at which fresh debris are being generated [8]. Radar systems can sometimes detect such space objects, but can, at best, localize them with lower precision than the shorter-wavelength optical systems. A stand-alone optical system based on the use of a light-sheet illumination and scattering concept [9] for spotting debris within meters

of a spacecraft has been proposed. A second system can localize all three coordinates of an unresolved, scattering debris [10, 11] by utilizing either the parallax between two observations, a pulsed laser ranging system, or a hybrid system. However, to the best of our knowledge, there is no other proposal of either an optical or an integrated optical-radar system to perform full 3D debris localization and tracking in the range of tens to hundreds of meters. Prasad [12] has proposed engineering point spread functions for 3D localization that exploits off-center image rotation.

Point spread function (PSF) engineering is a promising technique for solving the 3D localization problem, particularly in microscopy. The PSF morphology in a single 2D snapshot can be used to encode the depth information of a point source by using a selected phase pattern. The phase mask makes defocused images of point sources depth-dependent without excessive blurring. Different kinds of phase masks yield different kinds of depth-dependent PSFs, including the astigmatic PSF [13], tetrapod PSF [14], double-helix PSF [15] and single-lobe rotating PSF [12, 16, 17]. Our work only focuses on the single-lobe rotating PSF proposed by Prasad [12], which exploits off-center image rotation to encode both the range ζ and lateral coordinates (x, y) of point sources. Such idealized sources can model small, sub-centimeter class space debris, which, when actively illuminated, scatter a fraction of laser irradiance back into the imaging sensor.

Model-based approaches [18–21] have been explored to recover 3D coordinates of point sources via a rotating PSF imager. In general, these algorithms recover source locations by solving an optimization problem with an objective function consisting of a data-fitting term and a regularization term. For images corrupted by Poisson noise, the KL-NC algorithm, proposed in [18], consists of Kullback-Leibler (KL) divergence based data-fidelity term and a non-convex (NC) regularization term. It outperforms other combinations, such as ℓ_2-ℓ_1, the least-squares fitting-term model with the convex ℓ_1 regularization model, which is also considered in [18]. For the case of Gaussian noise, the CEL0 algorithm [20] using an ℓ_2 fitting term and an approximate ℓ_0 regularization term is proposed. Other multi-emitter fitting algorithms [1, 2, 22] exist in the field of super-resolution microscopy localization. However, these methods may require considerable computational time and careful adjustment of parameters in different situations.

Data-driven methods [3, 5, 6, 23] are also used for the localization problem in the field of microscopy, with fewer custom parameters. Over the past decade, deep learning-based data analytic methods have gained considerable attention in various fields. In recent years, this trend has reached the single-molecule imaging community. Generating a sufficiently large training dataset is very fast for SMLM experiments compared to other deep learning applications. To do this, related works usually employ a well-characterized forward model of the specific PSF to simulate the desired image pattern. Models trained on simulated data are then applied to real data, such as images acquired from microscopy. DeepSTORM3D [5] is a typical example of using a well-defined tetrapod PSF model to train a neural network with simulated data, which is then validated using both simulated and experimental data. It is applied to localization microscopy to render a super-resolution image of micro-tubules from images composed of single or overlapping PSFs. In [3], different representations of ground-truth labels are used. Given a 2D image, the neural network directly outputs a collection of fluorophore coordinates. In [6], recurrent layers are used to replace convolution layers to extract features more efficiently and save computational costs. A deeper framework called DECODE is proposed in [23]. DECODE allows input of multiple consecutive frames and concatenates features from multiple frames, based on the fact that emitters can persist in multiple subsequent frames. Compared with conventional model-based optimizations, data-driven methods require minimal refinements of parameters. However, the lack of interpretability prevents us from discussing the trade-off between precision and recall.

In order to dispense with careful and expensive adjustments of parameters that are specific to each situation, we introduce here a localization network to localize space debris using a rotating PSF. As the performance of the localization network shows a certain bias, a hard-sample

strategy [24, 25] is additionally integrated into the network structure to refine the dataset and improve performance by adjusting the trade-off between the precision and recall evaluation metrics. This also improves the interpretability of the network. To the best of our knowledge, our algorithm is the first developed so far for snapshot 3D localization and tracking of space debris via a rotating PSF approach within the deep learning framework. Our technique is efficient, and outperforms the current state-of-the-art model-based KL-NC method by more than 11% in precision with a comparable improvement of the recall rate. In addition, the proposed learning pipeline can be easily adapted to 3D SMLM applications.

The rest of this article is organized as follows: In Section 2, we first introduce the physical-optics model underlying the rotating PSF and subsequently calculate the minimum variance of unbiased estimation of the position coordinates of a point source by inverting the Fisher information matrix with respect to those coordinates. A localization network that incorporates a hard-sample strategy is proposed in Section 3. We then present a series of computer-simulation-based results in Section 4 to illustrate our approach and finally provide conclusions in Section 5.

2. Rotating point spread function

In this section, the physics model for the single-lobe rotating PSF is formulated, and the Cramer-Rao lower bound (CRLB) analysis is used to calculate the minimum variance of unbiased estimation of the source coordinates using such PSF. We use the CRLB as a criterion for choosing the only adjustable parameter in the original rotating-PSF design, namely the Fresnel zone count L, and for evaluating the performance of the rotating PSF.

2.1. Physics model for single-lobe rotating PSF

The PSF describes the image of a point source created by an imaging system. Here we specifically consider the single-lobe rotating PSF, which encodes the depth coordinate of the point source into the amount of PSF rotation [12]. In the paraxial scalar-field approximation, which is accurate for low-NA microscopy and telescopic imaging being considered here, the rotating PSF $\mathcal{A}_\zeta$ for a point source with unit flux $f = 1$, source lateral location $\mathbf{r}_0 = (x_0, y_0)$, and defocus parameter ζ is given by

$$\mathcal{A}_\zeta(\mathbf{s}) = \frac{1}{\pi}\left|\int P(\mathbf{u})\exp\left[\iota(2\pi\mathbf{u}\cdot\mathbf{s} + \zeta u^2 - \psi(\mathbf{u}))\right] d\mathbf{u}\right|^2, \tag{1}$$

where $\iota = \sqrt{-1}$ and $\mathbf{s} = \frac{\mathbf{r}}{\lambda z_I/R}$ is the scaled version of image plane position vector $\mathbf{r}$ as measured from the Gaussian image point that is located at $\mathbf{r}_I = \frac{z_I \mathbf{r}_0}{l_0+\delta_z}$ and about which the PSF rotates in the transverse image plane. Here λ is the imaging wavelength, and δ_z, l_0, z_I are the distances from the object to the in-focus object plane, in-focus object plane to the object-side principal plane, and image-side principal plane to the image plane, respectively. We denote the indicator function for the telescope exit pupil as $P(\mathbf{u})$, with $\mathbf{u}$ being the scaled pupil-plane position vector obtained from the physical pupil-plane position vector $\boldsymbol{\rho}$ by dividing it by the pupil radius R. In addition, $\psi(\mathbf{u})$ is the spiral phase profile defined in terms of the polar coordinates $\mathbf{u} = (\phi_\mathbf{u}, u)$ in the pupil plane as

$$\psi(\mathbf{u}) = l\phi_\mathbf{u}, \text{for } \sqrt{\frac{l-1}{L}} \le u \le \sqrt{\frac{l}{L}}, l = 1, \cdots, L, \tag{2}$$

where L represents the number of annular zones in the phase mask.

The rotating PSF performs one complete rotation in the depth-misfocus range $\zeta \in [-\pi L, \pi L]$, before it disintegrates unacceptably. In the paraxial-imaging regime, the physical depth misfocus distance, δ_z, from the plane of Gaussian focus is related to the dimensionless parameter, ζ, by

the relation,

$$\zeta = -\frac{\pi \delta_z R^2}{\lambda l_0 (l_0 + \delta_z)}. \tag{3}$$

For microscopy, typically $\delta_z \ll l_0$, but for remote sensing applications of interest here, δ_z may be of comparable order to l_0 or even much larger than l_0. In the latter case, as Eq. (3) shows, ζ becomes essentially independent of δ_z and the rotating PSF no longer carries any signature of source depth. This would tend to limit the range of performance of a practical rotating-PSF system for 3D localization of space debris under active illumination to depths of a few meters to a few hundreds of meters.

Figure 1 shows the comparison of the Gaussian PSF with the single-lobe rotating PSF. We generate the former PSF by using a Gaussian phase mask. The dimensionless parameter, which is proportional to the physical depth in the small misfocus limit, $\delta_z \ll l_0$, is changing in the range $[-\pi L, \pi L]$, where L is the number of Fresnel zones used in rotating PSF, which is set to be 7 here. It can be seen that rotating PSF maintains, on average, a smaller footprint while encoding depth (δ_z) information in its angle of rotation throughout this misfocus range. Its smaller average footprint allows it to continuously concentrate its intensity near the center of rotation, the latter being the (x, y) position of the source. By contrast, the peak brightness of the Gaussian PSF decreases and its width increases rapidly as the point source moves away from the focus.

The observed image I with M point sources is then generated as

$$I(x, y) = \mathcal{P}\left(\sum_{i=1}^{M} \mathcal{A}_i(x - x_i, y - y_i) f_i + b\right), \tag{4}$$

where (x_i, y_i, z_i), and f_i are, respectively, the 3D coordinates and the radiant flux of the ith point source. The information about the source depth, z_i, is encoded in the rotating PSF $\mathcal{A}_i$ via the dimensionless defocus parameter ζ_i. Here b is the spatially uniform mean background count per pixel, and $\mathcal{P}$ is the operator for adding data-dependent Poisson noise to the image.

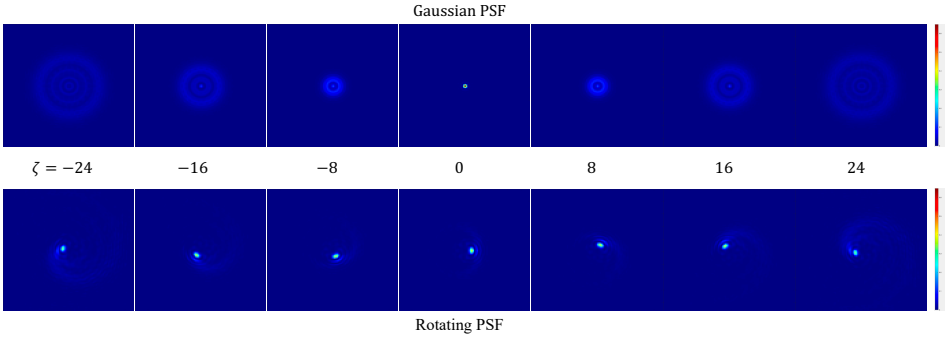

Fig. 1. Images of a single point source generated using the Gaussian PSF and the rotating PSF. The shape of PSFs is a function of the axial position of the point source. The Gaussian PSF is generated by inserting a Gaussian phase mask. The axial distance represented by the dimensionless parameter ζ of the two rows of images is in the range $[-\pi L, \pi L]$, where the number of zones is $L = 7$.

2.2. Cramér-Rao lower bounds for rotating PSF

The minimum possible error variance for unbiased estimation of a parameter from statistical data is called the Cramér-Rao lower bound (CRLB) [26, 27]. We now calculate the CRLB for

estimating the coordinates of a point source by considering its rotating PSF image, $h(\mathbf{r})$, where $\mathbf{r}$ denotes the position vector with respect to the location of the Gaussian image of the source. For notational brevity, we omit the ζ coordinate of the source from the list of arguments in its image h.

Let the square pixel pitch be a, pixel array size $N \times N$, and the average total signal and background photon counts distributed over the entire array be K and B, respectively. Let the background counts be uniform on average, with $b = B/N^2$ being the average background count per pixel. The mean photon count at the (i, j) pixel then has the value,

$$\mathbb{E}(K_{ij}) = K h_{ij}(\mathbf{r}) a^2 + b, \tag{5}$$

where we have assumed that the PSF is properly normalized over the image plane,

$$\int dA_I \, h_I(\mathbf{r}) \approx \sum_{i,j} h_{ij}(\mathbf{r}) a^2 = 1, \tag{6}$$

and the sampling pixel size is fine enough that $h(\mathbf{r})$ when integrated over the ij^{th} pixel may be accurately replaced by its value at the pixel center, h_{ij}, times the pixel area, a^2. Note that the sum condition in Eq. (6), when combined with Eq. (5), implies the sum rule,

$$\sum_{i,j} \mathbb{E}(K_{ij}) = K + B. \tag{7}$$

We will henceforth use a lexicographic single-index remapping of the pixels, $(i, j) \mapsto n$, as the actual square arrangement of the pixel array is irrelevant for our subsequent calculations.

The probability of detection of a count K_n at the n-th pixel follows the Poisson distribution,

$$P(K_n|h, K, B) = \frac{[\mathbb{E}(K_n)]^{K_n}}{K_n!} \exp[-\mathbb{E}(K_n)], \;\; \mathbb{E}(K_n) = K h_n \, a^2 + b. \tag{8}$$

Thus under the assumption of pixels performing statistically independent detections, the joint probability of detection of a set of counts, $\{K_1, K_2, \ldots, K_{N^2}\}$, has the product form,

$$\begin{aligned} P(\{K_n\}|h, K, B) &= \prod_{n=1}^{N^2} \frac{[\mathbb{E}(K_n)]^{K_n}}{K_n!} \exp[-\mathbb{E}(K_n)] \\ &= \exp[-(K + B)] \prod_{n=1}^{N^2} \frac{(K h_n(\boldsymbol{r}) a^2 + b)^{K_n}}{K_n!}. \end{aligned} \tag{9}$$

Hence its logarithm has the following form:

$$\ln P = -(K + B) + \sum_n \left[K_n \ln(K h_n(\mathbf{r}) \, a^2 + b) - \ln K_n! \right]. \tag{10}$$

The μ, ν element of the Fisher information matrix, $J_{\mu\nu}$, is defined [26, 27] as the statistical expectation,

$$J_{\mu\nu} = \mathbb{E}\left(\frac{\partial \ln P}{\partial x_\mu} \frac{\partial \ln P}{\partial x_\nu} \right), \;\; \mu, \nu = 1, 2, 3, \tag{11}$$

in which x_μ is the μ-th component of the source location vector $\mathbf{r}$. Since $\ln P$ given by Eq. (10) depends on $\mathbf{r}$ only through the dependence of h on $\mathbf{r}$, we can see that the above expression for FI simplifies to the following expectation of a double sum over the pixels:

$$J_{\mu\nu} = K^2 a^4 \sum_n \sum_m \frac{\mathbb{E}(K_n K_m) \partial_\mu h_n \, \partial_\nu h_m}{(K a^2 h_n + b)(K a^2 h_m + b)}, \tag{12}$$

in which $\partial_\mu, \partial_\nu$ are each a shorthand symbol for the partial derivative of the quantity that immediately follows it with respect to x_μ, x_ν, respectively. Since the detections by different pixels, indexed by $m \neq n$, are statistically independent, while for the nth pixel, under Poisson statistics, $\mathbb{E}(K_n^2) = [\mathbb{E}(K_n)]^2 + \mathbb{E}(K_n)$, we may see that by dividing the double sum in expression (12) into a double sum over all $m \neq n$ terms and a single sum over $m = n$ terms, we have

$$\begin{aligned} J_{\mu\nu} =& K^2 a^4 \sum_n \sum_m \frac{\mathbb{E}(K_n)\,\mathbb{E}(K_m)\partial_\mu h_n\,\partial_\nu h_m}{(Ka^2h_n + b)(Ka^2h_m + b)} + K^2 a^4 \sum_n \frac{\mathbb{E}(K_n)\,\partial_\mu h_n\,\partial_\nu h_n}{(Ka^2h_n + b)^2} \\ =& K^2 a^4 \left[\left(\sum_n \partial_\mu h_n \right) \left(\sum_m \partial_\nu h_m \right) + \sum_n \frac{\partial_\mu h_n\,\partial_\nu h_n}{Ka^2h_n + b} \right] \\ =& K^2 a^4 \sum_n \frac{\partial_\mu h_n\,\partial_\nu h_n}{Ka^2h_n + b}. \end{aligned} \tag{13}$$

To arrive at the second equality in Eq. (13), we used the relation in Eq. (5) to cancel out all factors except for the partial derivatives of the PSF in the double sum and to simplify the single sum. The final equality in Eq. (13) follows from the fact that the sum $\sum_n h_n$ is fixed at $1/a^2$ according to the normalization condition (Eq. (6)) on the PSF and thus its derivative must vanish, i.e.,

$$\partial_\mu \sum_n h_n = \sum_n \partial_\mu h_n = 0, \forall \mu. \tag{14}$$

The CRLBs on the estimation of the x, y, z coordinates of the point source are the corresponding diagonal elements of the inverse of the 3×3 FI matrix, i.e.,

$$\mathrm{CRLB}(x) = J_{11}^{-1}, \quad \mathrm{CRLB}(y) = J_{22}^{-1}, \quad \mathrm{CRLB}(z) = J_{33}^{-1}. \tag{15}$$

Let us write the rotating PSF as

$$h(\mathbf{r}) = E(\mathbf{r})\,E^*(\mathbf{r}), \tag{16}$$

where E denotes the complex amplitude PSF that may be expressed as the pupil area integral

$$E(\mathbf{r}) = \frac{1}{\sqrt{\pi}} \int dA\, P(\boldsymbol{u})\, \exp[i(2\pi \boldsymbol{u} \cdot \mathbf{r} + \zeta u^2 + \Psi(\boldsymbol{u}))], \tag{17}$$

in which $P(\boldsymbol{u})$ is the pupil indicator function, taking the value 1 inside the circular pupil of normalized radius 1 and 0 outside, $\boldsymbol{u} \cdot \mathbf{r} = (u_1 x + u_2 y)$ is the 2D inner product in the transverse plane, ζ is the defocus coordinate, and $\Psi(\boldsymbol{u})$ is the $L-$zone spiral phase mask that produces the PSF rotation. The dimensionless position coordinates x, y, ζ are related to the physical image-plane position coordinates of the source by the following scaling transformations (for $\delta_z \ll l_O$):

$$x \mapsto x(\lambda z_I/R),\ y \mapsto y\,(\lambda z_I/R),\ \delta_z = -\zeta\,(\lambda l_O^2/(\pi R^2)), \tag{18}$$

as we already know.

From the form of the rotating PSF in Eq. (16), it follows that

$$\partial_\mu h = E^* \partial_\mu E + E \partial_\mu E^* = 2\,\Re(E^* \partial_\mu E), \tag{19}$$

where $\Re$ denotes the real part of the quantity that follows it. The amplitude PSF has the following partial derivatives with respect to x, y, ζ:

$$\partial_\mu E = \frac{1}{\sqrt{\pi}} \begin{cases} i2\pi \displaystyle\int dA\, P(\boldsymbol{u})\, u_\mu \exp[i(2\pi \boldsymbol{u} \cdot \mathbf{r} + \zeta u^2 + \Psi(\boldsymbol{u}))], & \mu = 1, 2 \\ i \displaystyle\int dA\, P(\boldsymbol{u})\, u^2 \exp[i(2\pi \boldsymbol{u} \cdot \mathbf{r} + \zeta u^2 + \Psi(\boldsymbol{u}))], & \mu = 3. \end{cases} \tag{20}$$

Each of the area integrals in Eqs. (17) and (20) is readily evaluated using the MATLAB function "integral2", which, in view of Eq. (19), thus computes the three partial derivatives needed inside the sum (12) that represents the FI matrix elements.

The associated CRLBs for the rotating PSF, given by Eq. (15), then yield the minimum variances for any unbiased estimation of the dimensionless source coordinates (x, y, ζ). They are used as a criterion to evaluate the localization error performance when comparing different localization methods.

3. Localization network

In this section, we propose a supervised localization network for obtaining the point source positions. In addition, we explore a hard-sample strategy in the training set preparation to improve the interpretability of results.

3.1. Architecture and the pipeline of LocNet

Inspired by recent developments in deep learning for the 3D SMLM [5, 23], we propose a CNN-based method for our specific single-lobe rotating PSF with applications to telescope imaging. We implement the network structure on the deep learning platform *PyTorch*. We will henceforth refer to our CNN-based localization framework as LocNet.

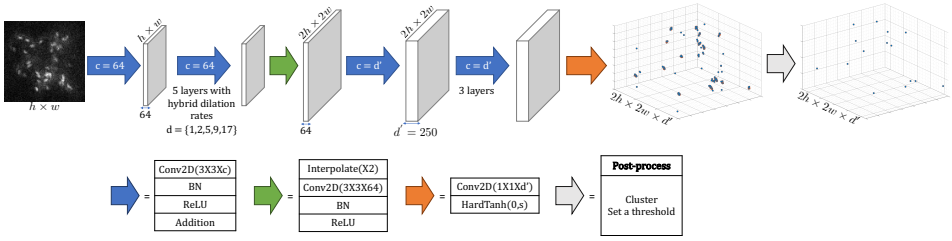

Fig. 2. A visualization of the LocNet framework.

A schematic of LocNet is shown in Figure 2. The framework consists of a network that is followed by a post-processing part. The network architecture comprises feature extraction parts, one interpolation layer, and one final prediction layer. For feature extraction parts, the first residual convolution layer [28] is used to increase the number of channels from the input gray image. We then set dilation rates for the subsequent five residual convolution layers following the hybrid dilated scheme [29], namely, $1, 2, 5, 9, 17$, to avoid the gridding issue. Residual convolution layers are represented by blue arrows in Figure 2. Each of these layers consists of a 2D convolution layer (Conv2D) with the filter size being $3 \times 3 \times c$, where c is the channel number, a batch normalization (BN) layer, a rectified linear unit (ReLU) working as activation layer and an addition operator to implement the summation of input and output that estimate the residual. The interpolation layer is represented by a green arrow that first upsamples the input features by two times using nearest neighbor interpolation, followed by Conv2D, BN, and ReLU operations. The final prediction layer is represented by an orange arrow consisting of Conv2D and the HardTanh function as the activation layer. The HardTanh function limits each entry to the range $[0, s]$, which is consistent with the entry values of point sources in ground truth labels mentioned below. The white blocks represent the intermediate features. The output of the network is a 3D discretized grid. The post-processing block, which follows the prediction layer, is represented as a gray arrow. Comprised of clustering and thresholding protocols, it controls the sparsity of the network output, producing a 3D grid with fewer nonzero entries.

Given the observed image $I \in \mathbb{R}^{h\times w}$ as the input, the LocNet outputs the corresponding up-sampled 3D grid $\hat{\mathcal{X}} \in \mathbb{R}^{2h\times 2w\times d'}$, with each entry value indicating the possibility of the existence of a point source. Both the width and height of $\hat{\mathcal{X}}$ are upsampled by a factor of 2, and d' denotes the number of slices with evenly distributed ζ. We adopt the mean square error as our loss function for localizing objects in our task. To evaluate the accuracy of our predictions, we compare them with the simulated ground truth using the l_2 distance between their respective heatmaps. Specifically, we compute the following expression,

$$l(\hat{\mathcal{X}}, \mathcal{X}_{\mathrm{GT}}) = \|\mathcal{G} * \hat{\mathcal{X}} - \mathcal{G} * (s\mathcal{X}_{\mathrm{GT}})\|_F^2, \tag{21}$$

where $\mathcal{X}_{\mathrm{GT}} \in \mathbb{R}^{2h\times 2w\times d'}$ is denoted as the ground truth, and $\mathcal{G}$ is a 3D Gaussian kernel with a standard deviation of 1 voxel. The Frobenius norm of a 3D tensor $\mathcal{A}$ is defined as $\|\mathcal{A}\|_F = \sqrt{\sum_{ijl} |a_{ijl}|^2}$. The ground-truth grid $\mathcal{X}_{\mathrm{GT}}$ indicates the existence of point sources,

$$(\mathcal{X}_{\mathrm{GT}})_{uvw} = \begin{cases} 1, & (u, v, w) = (x_i, y_i, z_i), \\ 0, & \text{otherwise}, \end{cases} \tag{22}$$

where (x_i, y_i, z_i) are the 3D coordinates of the ith point source. When we consider the 3D localization problem as a classification problem of entries, it is highly unbalanced, with only a few entries having point sources. We use a large value $s = 800$ as the weight of those entries with existing point sources in the ground truth grid $\mathcal{X}_{GT}$ to prevent gradient clipping [5]. The last activation layer HardTanh also limits the entries of output $\hat{\mathcal{X}}$ to the same range. Since the output of our model $\hat{\mathcal{X}}$ is unsampled, we rescale the ground-truth coordinates by the same factor and round them up and down to entries that match the discretized grid $\mathcal{X}_{\mathrm{GT}}$. Additionally, we add a small blur, via $\mathcal{G}$, to each ground-truth point source to compensate for minor shift errors that may occur while transforming actual point source coordinates onto a 3D lattice.

Since we focus on the rotating PSF and the loss function does not require sparsity of prediction, we use Algorithm 4.2 in [18] for post-processing, which contains two steps. The first step clusters point sources within a certain distance into a single point source, significantly reducing false positives. The second step removes point sources with intensity lower than 5% of the highest value. In this way, a list of coordinates $\{(x'_1, y'_1, z'_1), (x'_2, y'_2, z'_2), ..., (x'_{n'}, y'_{n'}, z'_{n'})\}$ is obtained as the final set of predicted point-source locations.

3.2. Data preparation and network training

We simulate both datasets used to train the LocNet and test the performance using the well-calibrated forward model of rotating PSF via Eq. (4). For comparison with KL-NC, we set the same number of zones $L = 7$ in rotating PSF and generate images with size 96×96. Considering the image center as the origin, all point sources in images have lateral coordinates $(x, y) \in (-34, 34) \times (-34, 34)$ in pixel units, and the dimensionless parameter $\zeta \in [-7\pi, 7\pi]$, which is proportional to the depth misfocus of the point sources. The magnitude of the maximum lateral coordinate relative to the image center is set to be smaller than half of the image size, which prevents the PSF from being cropped by the boundaries of the images. In the test set, 9 different source-density values following the uniform distribution in the range of 5 to 45 point sources, are considered. The photon numbers emitted by each point source follow a Poisson distribution with a mean of 2000 photon counts, which follows the setting in [18], as shown in Figure 3. It does not depend on different density cases. The uniform background noise is set to 5. For each density case, we generate 100 test images and take the average precision and recall rate to evaluate localization performance. In the training set, 10,000 images are simulated, with 90% used for training and the remaining 10% used for validation during the training of LocNet. The number of point sources in these 10,000 images follows a uniform distribution from 5 to 45, which covers all of the source densities tested.

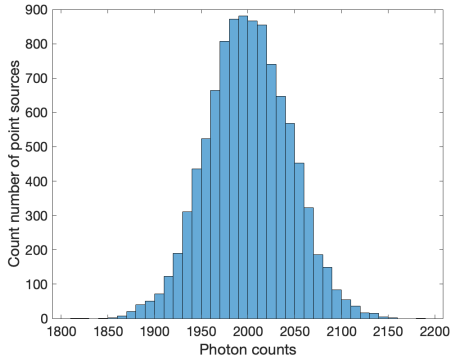

Fig. 3. The count number of point sources in each range of photon numbers out of 10,000 point sources, when the photon numbers follow the Poisson distribution with a mean of 2000.

The model is optimized by the Adam optimizer with an initial learning rate being 5×10^{-4}, which is found in the middle of the descending loss curve. The learning rate decays by a factor of 2 after every three epochs that the loss does not improve. The training stops when the learning rate is lower than 1×10^{-7} or the loss does not improve within 15 epochs. The training ran over 300 epochs with a total of 9,000 images, which took about one day on a computer equipped with an NVIDIA GeForce RTX 2080 Ti GPU and an Intel Xeon Silver Processor 4210 (2.20 GHz).

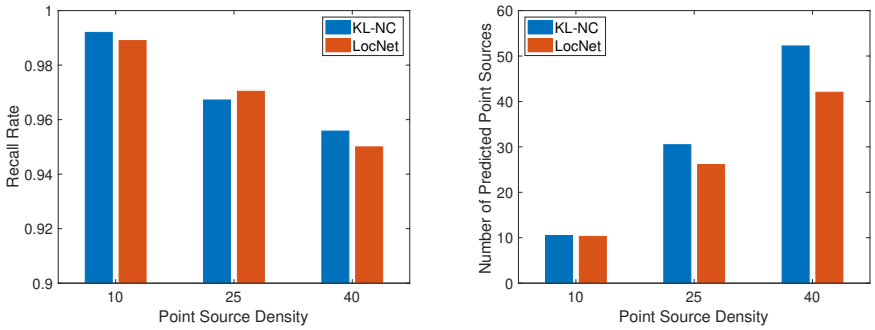

Fig. 4. Simulation-based results at different point source densities. **(a)** Average recall rate for 100 test images from KL-NC and LocNet. **(b)** Average number of final predicted point sources for 100 test images from KL-NC and LocNet.

3.3. Hard sample strategy

When considering telescope imaging, we want to guarantee that ground-truth point sources are predicted while accepting a small number of false positives. However, we found that the results of CNN-based methods exhibit specific biases, see Figure 4. For all the different densities, LocNet tends to generate fewer point sources than KL-NC. In some cases, LocNet has a lower recall than KL-NC.

Hard sample mining [30–33] is a promising approach to improve the performance of CNN based on considering the hardness of each sample. Resampling is one of the hard sample methods. It is widely used in dealing with highly unbalanced datasets, which helps prevent the network bias

toward learning information from the majority class and categorizing it into the majority class in the image classification task. It consists of adding more examples in the minority class by data augmentations such as rotating and flipping and/or removing some examples in the majority class. Inspired by these techniques, we propose a hard sample strategy customized for LocNet. The hard sample selection criteria focus on the index that is of the greatest concern in an application. As shown in Table 1, LocNet has lower recall rates than KL-NC [18] for most of the density cases. Hence it is natural to set the recall rate as a criterion to evaluate the hardness of each sample.

Instead of resampling from the existing training data, we obtain hard samples from a mock set Λ, newly generated from the forward model, since our images can be generated quickly based on the forward model of rotating PSF engineering [12]. Therefore, we enlarge the training data with some hard samples from a new mock set in each iteration. In general, for the kth iteration, we train the LocNet on a training set $\Omega^{(k)}$ and validate it on a mock set $\Lambda^{(k)}$. The recall rate is then calculated for each sample. Those samples whose recall rate is lower than the given threshold τ and thus prove to be difficult ("hard") to be predicted by the network, are moved into a hard-sample set $\Lambda_\tau^{(k)}$. The training set $\Omega^{(k)}$ is updated by adding those hard samples into it, namely, $\Omega^{(k+1)} = \Omega^{(k)} \cup \Lambda_\tau^{(k)}$. After sufficient training iterations, a refined dataset for training and a trained model can be obtained. Figure 5 and Algorithm 1 summarize the workflow for training LocNet using a hard sample strategy. The model is then used for the prediction of test images with different point source densities to evaluate the performance.

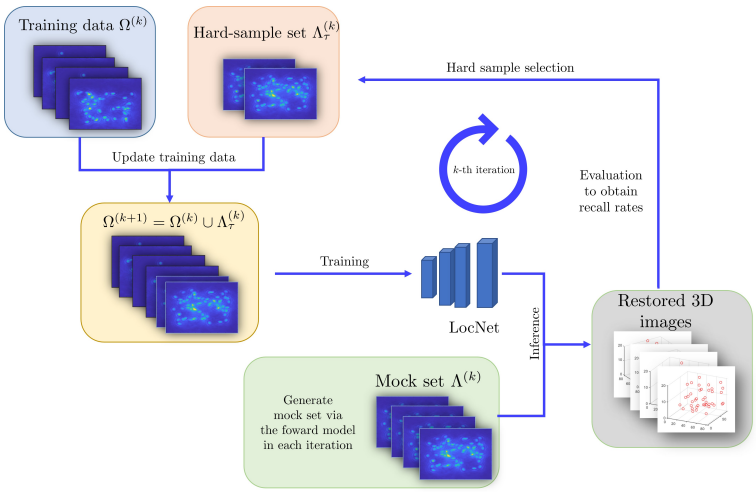

Fig. 5. Workflow of LocNet training with hard sample strategy. After each training iteration, a mock set Λ is validated. Samples with lower recall rates are selected into a hard sample set Λ_τ. The training set Ω is updated by adding those hard samples.

4. Results

In this section, we apply our CNN-based approach on rotating PSF for localizing point sources and compare results to KL-NC [18], which uses a variational optimization method.

We use recall and precision rate as metrics to judge the quality of 3D localization of point sources on 2D observed images. The recall and precision rate are calculated by,

$$\text{Recall rate} = \frac{\text{Number of identified true positive point sources}}{\text{Number of all true point sources}},$$

$$\text{Precision rate} = \frac{\text{Number of identified true positive point sources}}{\text{Number of all point sources identified by algorithm}}.$$

Algorithm 1: LocNet with Hard Sample Strategy for 3D Localization.

Input: Training set $\Omega^{(0)}$, the maximum iteration number kMax, threshold τ, and set $k = 0$

while $k \leq$ kMax **do**
- Train LocNet with $\Omega^{(k)}$
- Generate a mock set $\Lambda^{(k)}$
- Estimate 3D localization for each sample in $\Lambda^{(k)}$
- Do the evaluation to get recall rates
- Select hard-sample set $\Lambda_{\tau}^{(k)}$ corresponding to the recall rate being lower than τ
- $\Omega^{(k+1)} = \Omega^{(k)} \cup \Lambda_{\tau}^{(k)}$
- $k = k + 1$

end

True positive point sources are determined by considering the distance threshold between pairs of predicted and ground-truth point-source locations. According to our choice of the pupil-plane side length used in our FFT-based simulation of the rotating PSF, point sources within two-pixel units in the transverse dimensions from the ground-truth locations meet the Abbe-Rayleigh criterion of minimum transverse-resolvability threshold. Sources meeting this threshold still have to be within one unit of ζ in the axial dimension of the ground-truth source locations before we regard them as accurate estimations.

Table 1. Evaluation results of KL-NC [18], LocNet, and its variants. The result of LocNet with hard sample strategy is shown in columns of LocNet-HS. A control group of LocNet is trained on the same volume of samples as LocNet-HS but without using a hard sample strategy.

Number of Point Sources	KL-NC [18]		LocNet		LocNet-HS		Control Group	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
5	96.40%	99.80%	98.80%	99.80%	**99.33%**	**100.00%**	98.80%	99.60%
10	95.00%	**99.20%**	96.28%	98.90%	**98.19%**	99.10%	97.22%	98.20%
15	89.18%	98.80%	95.54%	98.87%	**96.74%**	**99.20%**	96.23%	98.67%
20	85.02%	97.55%	94.45%	98.00%	**95.17%**	**98.10%**	95.07%	97.80%
25	82.55%	96.72%	93.17%	97.04%	**94.48%**	**97.60%**	93.87%	97.00%
30	79.54%	97.30%	93.97%	96.87%	**94.37%**	**97.47%**	94.01%	96.70%
35	77.78%	95.26%	92.06%	95.80%	**92.15%**	**96.14%**	92.00%	95.66%
40	73.64%	**95.58%**	90.59%	95.00%	**90.89%**	95.20%	90.21%	94.98%
45	72.42%	94.31%	88.59%	93.93%	89.14%	**94.51%**	**89.15%**	93.89%
Average	83.50%	97.17%	93.71%	97.13%	**94.49%**	**97.48%**	94.06%	96.94%

4.1. Comparison with the model-based method

Table 1 shows the performance on test sets for 9 different point source densities, where the number of point sources is uniformly distributed between 5 and 45. The best precision and recall rates in each case are marked in bold. It can be seen that LocNet achieves higher precision than the KL-NC algorithm [18], but with lower recall, especially in high-density cases. However,

using a hard sample strategy, our method yields comparable recall by adjusting the training dataset. Since the hard sample strategy adds new samples to the training dataset, we also generate a control group to demonstrate that the increase in training dataset samples is not the main reason for the improved performance. The training dataset for the control group is the same size as LocNet-HS but uses randomly generated new data instead of the hard sample strategy. The performance of the control group is comparable to LocNet without significant improvement.

Figure 6 shows an example of the 25-point-source case. A comparison of the first two columns illustrates how the estimates from LocNet incurred fewer false positives but missed a ground-truth point source, the latter pointed out via a red arrow in Figure 6. This missed point source was recovered with the hard sample strategy, as seen in the third column. Examples of two other source densities are also shown in supplementary material.

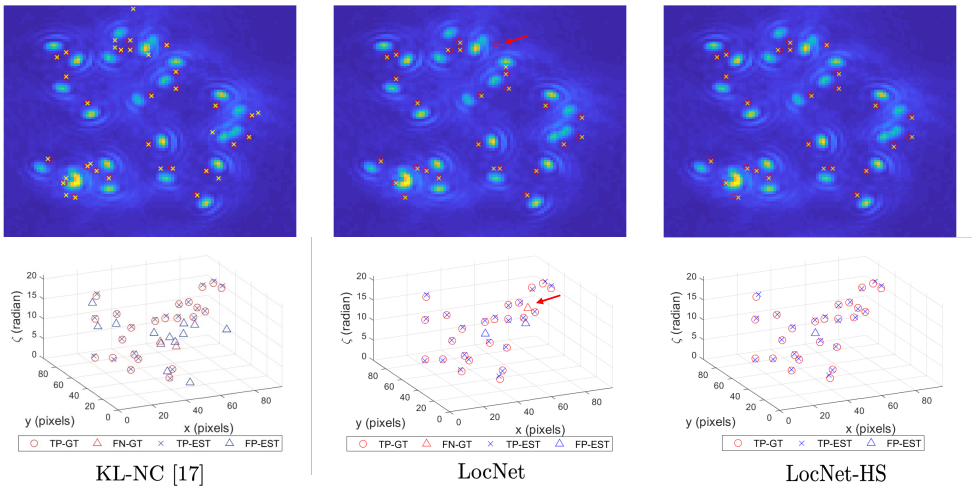

Fig. 6. 3D localizations for the 25-point-source case. The first row is 2D snapshots where "o" is the ground-truth point source, and "x" is the estimated point source. The second row is the locations shown in 3D grids. where the ground-truth point sources are in red markers with red "Δ" being false-negative and red "o" being true-positive. The estimated ones are in blue, with blue "Δ" being false-positive and blue "x" being true-positive.

4.2. Localization error

We next analyze the localization error for a single-point source using models pre-trained in Section 4.1. We sample the value of ζ in the whole range $[-\pi L, \pi L]$ with step size =1. For each sampled value of ζ, a test set of 100 images are generated. Each image contains only one point source with random x and y coordinates and the fixed ζ. Figure 7 shows the root mean square error (RMSE) for KL-NC, LocNet, and LocNet-HS based localizations, compared with the theoretical lower bound from unbiased estimation. It can be seen that using a CNN-based framework significantly reduces the localization error, while LocNet-HS achieves still smaller localization errors in most sampled zeta values. In addition, since the output of KL-NC and LocNet are discretized 3D grids, the accuracy of their output will be limited by the spacing of the grids. However, after post-processing clustering, the center obtained from the cluster does not have to be on the grid, and the localization error can be much lower than the grid spacing, as can be seen in Figure 7.

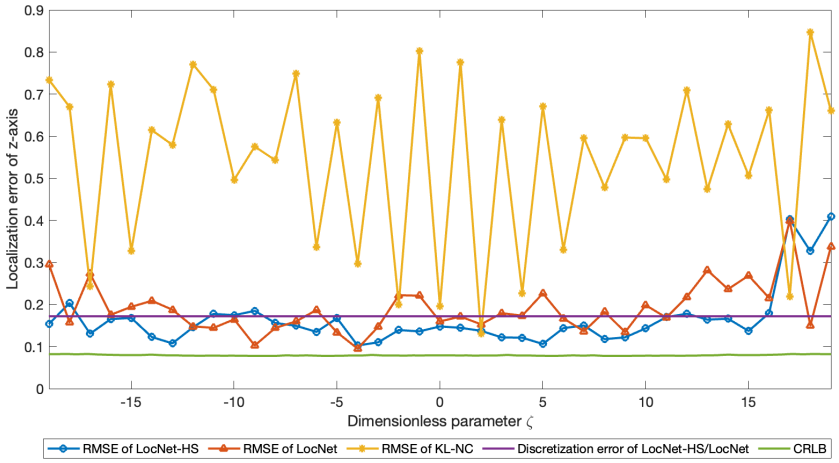

Fig. 7. Localization error of single point source in pixel unit computed from LocNet-HS, LocNet, and KL-NC [18], compared with CRLB. The discretization error of LocNet-HS/LocNet (the purple curve) is the grid spacing in the discretized output lattice, which shows the limitation of network prediction.

4.3. Higher noise level

To assess the robustness of our approach, we have also examined datasets with a higher noise level. In particular, we set the uniform background noise per pixel at the mean value of $b = 10$, as opposed to 5 in the previous experiments. The results of the experiment, compared with KL-NC, are presented in Table 2. These findings reveal that in a more noisy and challenging scenario, our LocNet-HS model yields an even greater improvement, about 1.85% in precision and 1.25% in the recall rate.

5. Conclusions

In [18], KL-NC was shown to outperform other variational methods. In this work, we use a localization network with a hard sample strategy to localize positions of 3D point sources from a 2D snapshot generated using rotating PSF. Our new approach further enhances the performance by removing false-positive point sources.

Our future work will be focused on further improving both the performance and interpretability, using other tools such as physics-informed neural networks and associated loss terms and unrolling in combination with the hard-sample strategy of the present work. In addition, we will consider multi-frame images [34] to track the motion of space debris from the perspective of deep learning.

Funding. HKRGC Grants Nos. N_CityU214/19, CityU11301120, C1013-21GF, and CityU11309922; the Natural Science Foundation of China No. 12201286; Shenzhen Fundamental Research Program JCYJ20220818100602005; CityU Grant 9380101.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Table 2. Evaluation results of KL-NC [18], LocNet, and its variants when background noise has the mean value, $b = 10$. The result of LocNet with hard sample strategy is shown in columns labeled as LocNet-HS. A control group of LocNet is trained on the same volume of samples as LocNet-HS but without using a hard sample strategy.

Number of Point Sources	KL-NC [18]		LocNet		LocNet-HS		Control Group	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
5	95.62%	98.40%	97.87%	**99.60%**	**99.50%**	98.60%	96.45%	99.40%
10	90.86%	97.60%	95.53%	97.50%	**97.05%**	**98.60%**	94.91%	97.50%
15	87.24%	97.27%	93.05%	96.93%	**96.14%**	**98.13%**	94.21%	96.40%
20	79.79%	96.55%	91.33%	96.20%	**94.00%**	**97.65%**	92.80%	93.25%
25	75.59%	96.12%	90.48%	94.60%	**92.49%**	**96.20%**	91.31%	91.72%
30	76.26%	94.73%	89.25%	94.57%	**91.89%**	**95.43%**	89.15%	95.13%
35	75.28%	**94.94%**	89.27%	92.47%	**90.34%**	94.54%	87.47%	94.10%
40	70.04%	**93.90%**	87.60%	91.83%	**88.33%**	93.60%	86.74%	93.10%
45	70.11%	92.56%	86.44%	90.69%	**87.69%**	**92.84%**	85.81%	92.49%
Average	80.09%	95.79%	91.20%	94.93%	**93.05%**	**96.18%**	90.98%	94.78%

References

1. R. J. Marsh, K. Pfisterer, P. Bennett, L. M. Hirvonen, M. Gautel, G. E. Jones, and S. Cox, "Artifact-free high-density localization microscopy analysis," Nat. Methods **15**, 689–692 (2018).
2. J. Min, C. Vonesch, H. Kirshner, L. Carlini, N. Olivier, S. Holden, S. Manley, J. C. Ye, and M. Unser, "FALCON: fast and unbiased reconstruction of high-density super-resolution microscopy data," Sci. Reports **4**, 1–9 (2014).
3. N. Boyd, E. Jonas, H. Babcock, and B. Recht, "DeepLoco: fast 3D localization microscopy using neural networks," BioRxiv p. 267096 (2018).
4. M. Lelek, M. T. Gyparaki, G. Beliu, F. Schueder, J. Griffié, S. Manley, R. Jungmann, M. Sauer, M. Lakadamyali, and C. Zimmer, "Single-molecule localization microscopy," Nat. Rev. Methods Primers **1**, 1–27 (2021).
5. E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, T. Naor, R. Orange, T. Michaeli, and Y. Shechtman, "DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning," Nat. Methods **17**, 734–740 (2020).
6. J. Li, G. Tong, Y. Pan, and Y. Yu, "Spatial and temporal super-resolution for fluorescence microscopy by a recurrent neural network," Opt. Express **29**, 15747–15763 (2021).
7. N. O. D. P. Office, "Monthly object type charts by number and mass," Orbital Debris Q. News **27**, 1–14 (2023).
8. J. F. Dargin III, "Removing orbital space debris from near earth orbit," (2019). US Patent 10,501,212.
9. C. R. Englert, J. T. Bays, K. D. Marr, C. M. Brown, A. C. Nicholas, and T. T. Finne, "Optical orbital debris spotter," Acta Astronaut. **104**, 99–105 (2014).
10. D. Hampf, P. Wagner, and W. Riede, "Optical technologies for the observation of low earth orbit objects," arXiv preprint arXiv:1501.05736 (2015).
11. P. Wagner, D. Hampf, F. Sproll, T. Hasenohr, L. Humbert, J. Rodmann, and W. Riede, "Detection and laser ranging of orbital objects using optical methods," in *Proc. Remote Sens. Sys. Eng.*, vol. 9977 (SPIE, 2016), pp. 66–76.
12. S. Prasad, "Rotating point spread function via pupil-phase engineering," Opt. Lett. **38**, 585–587 (2013).
13. B. Huang, W. Wang, M. Bates, and X. Zhuang, "Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy," Science **319**, 810–813 (2008).
14. Y. Shechtman, L. E. Weiss, A. S. Backer, S. J. Sahl, and W. Moerner, "Precise three-dimensional scan-free multiple-particle tracking over large axial ranges with tetrapod point spread functions," Nano Lett. **15**, 4194–4199 (2015).
15. S. R. P. Pavani, M. A. Thompson, J. S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. E. Moerner, "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function," Proc. Natl. Acad. Sci. **106**, 2995–2999 (2009).
16. M. D. Lew, S. F. Lee, M. Badieirostami, and W. E. Moerner, "Corkscrew point spread function for far-field three-dimensional nanoscale localization of pointlike objects," Opt. Lett. **36**, 202–204 (2011).
17. R. Kumar and S. Prasad, "PSF rotation with changing defocus and applications to 3D imaging for space situational awareness," in *Proc. AMOS Tech. Conf., Maui, HI,* (2013).
18. C. Wang, R. Chan, M. Nikolova, R. Plemmons, and S. Prasad, "Nonconvex optimization for 3-dimensional point

source localization using a rotating point spread function," SIAM J. Imag. Sci. **12**, 259–286 (2019).
19. C. Wang, R. Plemmons, S. Prasad, R. Chan, and M. Nikolova, "Novel sparse recovery algorithms for 3D debris localization using rotating point spread function imagery," (2018).
20. C. Wang, R. H. Chan, R. J. Plemmons, and S. Prasad, "Point spread function engineering for 3D imaging of space debris using a continuous exact ℓ_0 penalty (CEL0) based algorithm," in *Int. W. Imag. Proces. & Inverse Probl.*, (Springer, 2018), pp. 1–12.
21. C. Wang, G. Ballard, R. Plemmons, and S. Prasad, "Joint 3D localization and classification of space debris using a multispectral rotating point spread function," Appl. Opt. **58**, 8598–8611 (2019).
22. B. Shuang, W. Wang, H. Shen, L. J. Tauzin, C. Flatebo, J. Chen, N. A. Moringo, L. D. Bishop, K. F. Kelly, and C. F. Landes, "Generalized recovery algorithm for 3D super-resolution microscopy using rotating point spread functions," Sci. Reports **6**, 1–9 (2016).
23. A. Speiser, L.-R. Müller, P. Hoess, U. Matti, C. J. Obara, W. R. Legant, A. Kreshuk, J. H. Macke, J. Ries, and S. C. Turaga, "Deep learning enables fast and dense single-molecule localization with high accuracy," Nat. methods **18**, 1082–1090 (2021).
24. Y. Suh, B. Han, W. Kim, and K. M. Lee, "Stochastic class-based hard example mining for deep metric learning," in *Proc. Conf. CVPR,* (2019), pp. 7251–7259.
25. A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. Conf. CVPR,* (2016), pp. 761–769.
26. H. L. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory* (John Wiley & Sons, 2004).
27. S. M. Kay, *Fundamentals of statistical signal processing: estimation theory* (Prentice-Hall, Inc., 1993).
28. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. CVPR,* (2016), pp. 770–778.
29. P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *IEEE Conf. WACV,* (IEEE, 2018), pp. 1451–1460.
30. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell **32**, 1627–1645 (2009).
31. H. Sun, Z. Chen, S. Yan, and L. Xu, "Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision,* (2019), pp. 6737–6747.
32. K. Chen, Y. Chen, C. Han, N. Sang, and C. Gao, "Hard sample mining makes person re-identification more efficient and accurate," Neurocomputing **382**, 259–267 (2020).
33. H. Sheng, Y. Zheng, W. Ke, D. Yu, X. Cheng, W. Lyu, and Z. Xiong, "Mining hard samples globally and efficiently for person reidentification," IEEE Internet Things J. **7**, 9611–9622 (2020).
34. J. Tao, Y. Cao, and M. Ding, "SDebrisNet: A spatial–temporal saliency network for space debris detection," Appl. Sci. **13**, 4955 (2023).