Enabling Real-Time Fluctuation-Based Super Resolution Imaging

Jelle Komen[1], Miyase Tekpınar[1], Ran Huo[1], Klarinda de Zwaan[1], Nergis Tomen[2], Kristin Grußmayer[1*]

[1*]Department of Bionanoscience and Kavli Institute of Nanoscience, Delft University of Technology, Delft, Netherlands.
[2]Department of Intelligent Systems, Delft University of Technology, Delft, Netherlands.

*Corresponding author(s). E-mail(s): k.s.grussmayer@tudelft.nl;
Contributing authors: j.j.m.komen@tudelft.nl; m.tekpinar@tudelft.nl; r.huo@tudelft.nl; b.k.d.zwaan@tue.nl; n.tomen@tudelft.nl;

Abstract

Live-cell imaging captures dynamic cellular processes, but many structures remain beyond the diffraction limit. Fluctuation based super-resolution techniques reach beyond the diffraction limit by exploiting correlations in fluorescence blinking. However, the methods require the acquisition of hundreds of frames and involve computationally intensive post-processing, which can take tens of seconds, thereby limiting the suitability for real-time sub-diffraction imaging of fast cellular events. To address this, we use the recursive neural network model termed MISRGRU, which integrates sequential low-resolution frames to extract spatio-temporally correlated signals. Using super-resolution optical fluctuation imaging as a target, our approach not only significantly improves temporal resolution but also doubles the spatial resolution. We benchmark MISRGRU against the widely used U-Net architecture, evaluating performance in terms of resolution enhancement, training complexity, and inference latency using both synthetic data and real microtubule structures. After optimizing MISRGRU, we demonstrate a 400-fold reduction in computational latency compared to SOFI, highlighting its promise as an efficient, real-time solution for live-cell super-resolution imaging.

Keywords: Real time imaging, AI, convolutional gated recurrent unit (convGRU), fluctuation imaging, super-resolution

1 Introduction

Live-cell fluorescence imaging captures dynamic cellular behaviors and aims to maximize both spatial and temporal resolution while minimizing sample damage [1], enabling advancements in fundamental cell biology. However, visualizing these cellular dynamics is constrained by the diffraction limit, which restricts the spatial resolution to around 200-300 nm. Super-resolution (SR) techniques can overcome this barrier and achieve higher spatial resolutions. Well-known techniques include Stimulated Emission Depletion (STED) microscopy [2], Structured Illumination Microscopy (SIM) [3], and Single Molecule Localization Microscopy (SMLM) [4, 5]. While STED and SIM require a complex microscopy setup, SMLM can be performed with a wide-field microscope. By using the stochastic and sparse blinking

of fluorophores, individual molecules are localized from diffraction-limited image sequences and their positions are used to generate a super-resolution image with a spatial resolution of a few tens of nanometers. However the reconstruction of a super-resolution image needs many thousands of frames, typically corresponding to data acquisition over minutes to hours. This elongated experimental duration leads to relatively low temporal resolution and speeding-up the imaging usually entails higher laser illumination. Thus, this prevents observation of fast processes and can lead to high phototoxicity [6, 7]. On the other hand, fluctuation-based super-resolution techniques, which also utilize blinking emitters, typically require only hundreds of frames to reconstruct an SR image as they can handle high densities of blinking fluorophores. Widely used examples of fluctuation-based methods are super-resolution optical fluctuation imaging (SOFI) [8–10] as well as the related super-resolution imaging based on autocorrelation with two-step deconvolution (SACD) and super-resolution Radial Fluctuations (SRRF) microscopy [11, 12]. Other options include the Multiple signal classification algorithm (MUSICAL) [13], Bayesian analysis of blinking and bleaching (3B)) and entropy-based super-resolution imaging (ESI).

SOFI and SACD both leverage the correlation of blinking from a fluorophore over time and space and the independence of different fluorophores [8, 12]. Dertinger et al. showed that n^{th} order cumulants of a diffraction limited image sequence of blinking emitters result in a point spread function (PSF) raised to the power of n, which brings a $\sqrt{n}$-fold better resolution. This resolution enhancement can be further increased to n-fold in n^{th}-order SOFI by applying linearization and deconvolution or Fourier reweighting after the cumulant calculation [14]. Higher orders require better statistics, thus more frames and longer acquisition time than second-order SOFI. This can be challenging due to limited signal-to-noise ratio (SNR) as a result of the sample bleaching. Therefore, there have been previous attempts to reduce the number of frames. One such approach is SACD (super-resolution imaging based on autocorrelation with two-step deconvolution), which applies background subtraction and deconvolution as a pre-processing step to conventional SOFI to reduce noise and improve signal quality prior to cumulant analysis [12]. Another related technique is SRRF (Super-Resolution Radial Fluctuations), which operates by computing gradient maps and applying a radiality transform that exploits the radial symmetry of the PSF. A final cumulant-based analysis of the image sequence produces a super-resolved, noise-suppressed image. The enhanced version, eSRRF, further assists users in selecting parameter settings that minimize the risk of artifacts [15].

Even-though the above-mentioned methods require shorter imaging time than SMLM, their temporal resolution are typically in the range of seconds to minutes. Under low laser illumination necessary for prolonged imaging and thus for low SNR conditions, they require longer measurements to reach high SNR image reconstruction. In addition, the SR image is mostly reconstructed offline, after the experiment and the computational latency of these methods is limited to the range of seconds even with GPU support. Therefore they are not suitable for long-duration, real-time, live cell imaging. In this study, we aim to offer a network architecture to reconstruct a super-resolution image in real-time from few, low-SNR microscopy image frames by using deep learning.

Deep learning has been extensively explored to increase image quality and imaging speed [16]. Most methods employ the U-Net architecture [17] because of its high performance in image segmentation, denoising, and deconvolution[18–22]. Despite the growing number of studies applying deep learning in microscopy, few models focus on extracting the temporal information. For instance, unlike standard convolutional neural network (CNN) designs, such as U-Net, recurrent neural networks (RNNs) are rarely used in microscopy. Nevertheless, some studies have shown the potential of RNNs to accelerate SMLM image reconstruction [23, 24]. A notable example is DBlink, which represents one of the first applications of RNNs, specifically, a long-short-term memory (LSTM) network, for video reconstruction from SMLM data [25]. By feeding localization maps into the LSTM, DBlink generates super-resolution videos at rates of up to 66 frames per second (FPS). However, this method remains constrained to SMLM sample conditions and relies on localization-based post-processing, limiting its suitability for real-time image computation.

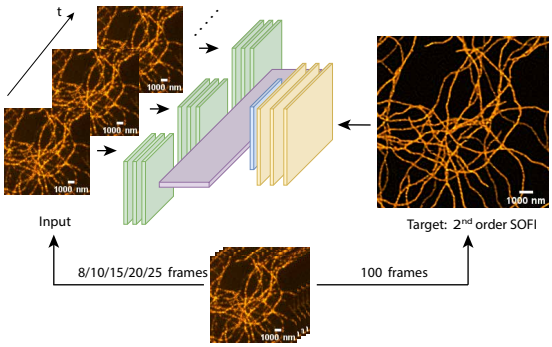


Fig. 1 Training MISRGRU by feeding with diffraction-limited image sequences for accelerating second-order SOFI with optimization based on varying the number of input frames. Input sequence lengths of 8, 10, 15, 20, and 25 consecutive frames are evaluated. Green indicates the encoder blocks, purple the fusion layer, blue the global average pooling, and yellow the decoder. The scale bar is 1000 nm.

In this study, we propose to use the multi-image super-resolution gated recurrent unit (MISRGRU) architecture for real-time super-resolution screening [26]. Originally designed to predict single super-resolution images from multiple low-resolution images in remote sensing, MISRGRU utilizes a convolutional gated recurrent unit (convGRU), a powerful RNN architecture capable of leveraging both temporal and spatial information, making it compatible with any fluctuation-based super-resolution technique. MISRGRU is flexible in that it supports training and inference with varying number of input images. By design, the output can be easily interpolated to any image size, unlike commonly used U-Net which requires pre-processing, allowing for straightforward training to produce n-fold larger output images for n-fold resolution improvement in higher-order SOFI.

Here, we first test MISRGRU on synthetic data, where we have full access to ground truth structures for evaluation. Our synthethic dataset features microtubule-like structures with plausible curvature and a realistic sCMOS camera noise model. In this study, we focus on reconstructing second-order SOFI images as the target output in order to limit computational demands and maximize imaging speed. We tested different input frame counts with the MISRGRU model to determine the temporal resolution. Notably, the 8-frame model performed similarly to the 20-frame model in resolution enhancement and reconstructing the true signals. Similar improvements were observed on real-world data. Compared to the

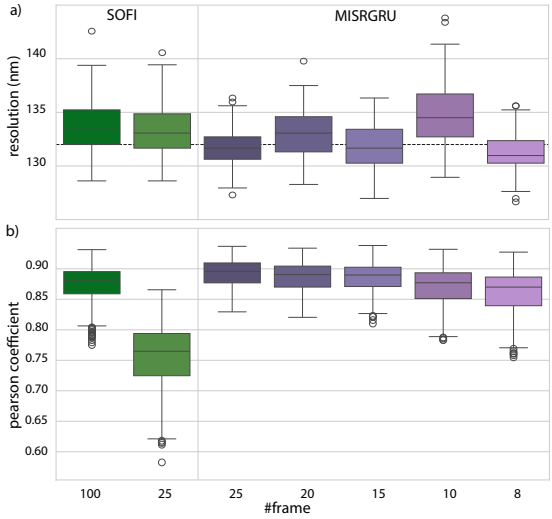


Fig. 2 Evaluation of our MISRGRU model and SOFI on simulation data. (a) Spatial resolution estimated via image decorrelation analysis and (b) resolution-scaled Pearson correlation coefficient for SOFI (green) and MISRGRU (purple), using different numbers of input frames. Each boxplot depicts the median value (center line), 25th and 75th percentile (box edges), and the 1.5 IQR (interquartile range) (whiskers). The dashed line in (a) marks the theoretical half-resolution limit of the widefield average image.

hundreds of frames typically required for second-order SOFI, MISRGRU achieved a two-fold spatial resolution improvement for microtubules in cells using significantly fewer frames and reduced computational latency. Furthermore, we optimize MISGRU and show that it has reduced overall prediction latency and training complexity with respect to a U-net model. Lastly, we assessed the dynamic limits of the networks by simulated movements and showed that 8-frame model can still reconstruct non-blurry structures from the movements with a speed up to 2 μm/s. Overall, the proposed method offers great potential for real-time sample screening by enhancing spatio-temporal resolution during experiments.

2 Results

2.1 SOFI for real-time imaging

SOFI typically requires several hundred to thousands of frames—depending on fluorophore blinking characteristics and the signal-to-noise ratio (SNR)—to reach its theoretical two-fold resolution enhancement and to produce high-SNR second-order SOFI reconstructions [27]. However,

this frame count and the associated computational time limit SOFI's ability to achieve high temporal resolution for real-time super-resolution sample screening. To address this limitation, we leverage deep learning to reduce the number of required input frames and enable real-time analysis, while maintaining image quality comparable to high-SNR SOFI reconstructions.

The deep-learning network architecture is based on the MISRGRU model, which comprises an encoder to extract feature representations of low-resolution frames, followed by a convolutional gated recurrent network (convGRU) based fusion layer to process the features sequentially. Global average pooling (GAP) combines them into a single representation, and a decoder reconstructs the high-resolution image by upsampling the fused feature maps (see 2.2.1). The input of the network consists of a time series of diffraction-limited frames, with the target output being second-order SOFI reconstructions of the structures as illustrated in Figure 1. To validate the network, we created a realistic synthetic dataset by generating structural patterns of microtubules randomly decorated with fluorophores using binary masks. We simulated fluorophore blinking dynamics, including on- and off-states, and incorporated an sCMOS camera noise model along with a non-stationary background to closely mimic experimental imaging conditions (see Methods). The dataset includes a range of SNR levels and includes varying off-times (600, 1200, 2000 ms), while maintaining a constant on-time of 10 ms. For generating target images, we used 100 diffraction-limited frames to compute second-order cross-cumulants, followed by linearization and Richardson-Lucy deconvolution for post-processing [14]. These SOFI images were then employed as targets during network training. Following an approach similar to Chen et al., we trained separate models using 8, 10, 15, 20, and 25 input frames to assess the network's ability to reconstruct super-resolved images under different temporal constraints[22].

We determined the minimum number of input frames necessary to achieve a two-fold resolution enhancement by applying a set of established evaluation metrics, as described in the Performance evaluation section. Since higher spatial resolution alone does not ensure accurate structural reconstruction, we also evaluated the preservation

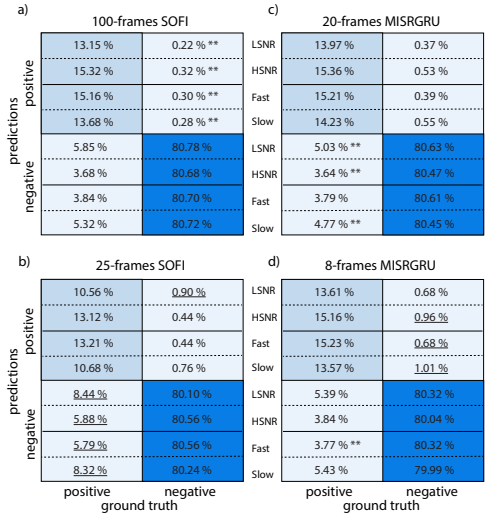


Fig. 3 Evaluation of our MISRGRU model and SOFI on simulation data via confusion matrices. (a) SOFI reconstruction using 100 frames, (b) and using 25 frames. (c) MISRGRU model using 20 frames input, and (d) 8 frames input. Each box of the confusion matrix divided in four rows to represent the individual datasets including, low-SNR , high-SNR (Ion: 40 and 100 photons/frame), fast and slow blinking (off-time: 600 vs 2400 ms). The highest percentages of misclassified pixels among different models are underlined and the lowest percentages are marked with two star signs.

of microtubule integrity. Our evaluation process consisted of the following steps:

1. We estimated the resolution of the predicted super-resolution (SR) images using decorrelation analysis, a widely adopted technique in super-resolution microscopy for resolution quantification [28].
2. We measured the similarity between the predicted and ground-truth (GT) images using resolution-scaled Pearson correlation (RSP) via SQUIRREL analysis, a standard method in SR imaging for quantitative comparison [29].
3. To evaluate the accuracy of SR reconstructions, we binarized the images and computed the True Positive Rate (TPR), True Negative Rate (TNR), and constructed confusion matrices. These metrics are commonly used in machine learning for performance evaluation.

In figure 2a, we compare the resolutions of reconstructed images using different models, including the SOFI method. All models approximate the theoretical two-fold spatial resolution improvement by second-order SOFI. The

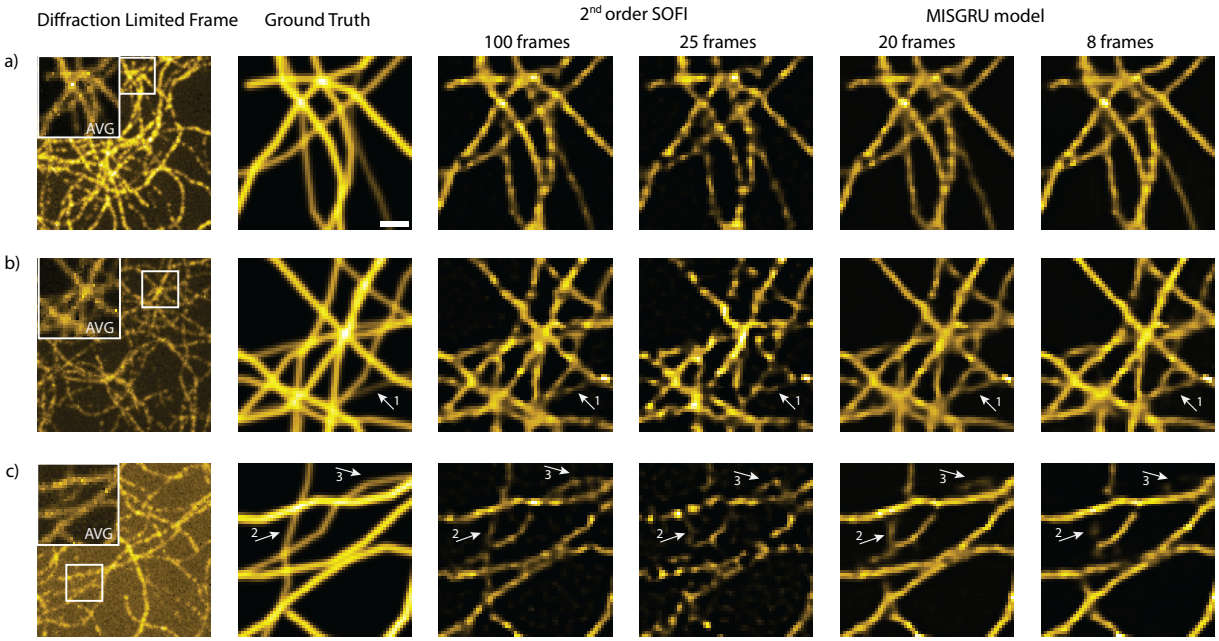


Fig. 4 Comparison of super-resolution (SR) reconstructions under varying simulated signal-to-noise ratios (SNRs) for the ground truth (GT), second-order SOFI, and the MISRGRU models. Each row corresponds to a different SNR condition, defined by the photon count per frame (100, 60, and 40 photons/frame, from top to bottom). From left to right: a single diffraction-limited input frame of a simulated microtubule structure, the ground truth image demonstrating two-fold resolution enhancement, second-order SOFI reconstructions using 100 and 25 frames, and model-based SR reconstructions using 20 and 8 frames. The scale bar represents 500 nm. White arrows show the degradation in structural detail.

resolution-scaled Pearson correlation coefficient, shown in figure 2b, quantifies the structural similarity between super-resolution reconstructions and the simulated ground-truth images. The analysis reveals that all MISRGRU models surpass SOFI in preserving structural integrity. While the Pearson correlation coefficient slightly decreases with fewer input frames, we find that the 8-frame model offers a favorable trade-off between temporal resolution and spatial accuracy.

Figure 3 presents confusion matrices used to assess pixel-level classification accuracy by comparing model predictions to binarized ground truth images. We evaluated true positive and true negative rates for SOFI-based super-resolution reconstructions using 100 and 25 input frames, as well as for predictions from the 8-frame and 20-frame MISRGRU models. These evaluations were conducted across different signal-to-noise ratio (SNR) conditions—by varying the mean photon count per frame—and under different blinking regimes, achieved by adjusting the off-times while keeping the on-time fixed.

The results show that the 20-frame MISRGRU model outperforms SOFI across all conditions by exhibiting the lowest rate of missed structures (false negatives) and, in most cases, generates fewer artifacts (false positives) than SOFI with 25 frames. At high SNR, both MISRGRU models tend to produce slightly more false positives compared to SOFI. This behavior suggests that the network learns to infer missing signal, potentially leading to minor over-prediction. However, these false positives remain under 1 % of total pixels, indicating that such errors do not result in significant hallucinations. Notably, the 8-frame model achieves true positive rates comparable to SOFI with 100 frames, though it is more prone to false positives than the 20-frame model, particularly under slow-blinking conditions—likely due to reduced emitter activity and insufficient temporal sampling in the shorter image sequence.

Visual examples of reconstructions under varying SNR levels are provided in Figure 4, and under different blinking conditions in Figure 5. At lower SNRs, SOFI reconstructions using 100 or

25 frames fail to resolve some structures, whereas both MISRGRU models recover a greater proportion of the signal (Figure 4b). While SOFI performance can be improved with longer acquisitions, the 8-frame MISRGRU model maintains a clear advantage in low-SNR settings, even with fewer frames. Although it exhibits some loss of structural detail under extreme noise conditions (Figure 4c), it still outperforms SOFI with 25 frames, offering a practical benefit for live-cell imaging by allowing lower laser intensities—reducing phototoxicity—or extended acquisition durations.

As expected, under the slowest blinking condition (Figure 5c), the 8-frame model fails to recover certain structures. Since the models were trained with a fixed on-time of 10 ms, we anticipate improved performance in cases involving faster blinking dynamics or longer on-times.

2.2 Optimizing the network design

2.2.1 MISRGRU vs U-net

Multi-image super-resolution (MISR) methods have been extensively studied in satellite imaging [30]. MISRGRU was first introduced in 2020 to the computer vision community to improve the quality of satellite imagery [26]. This model leverages multiple low-resolution images and an encoder-decoder framework to generate a super-resolved output. As illustrated in Figure 1, the architecture is composed of an encoder, a convolutional gated recurrent unit (convGRU) serving as a fusion layer, and a decoder.

The encoder processes each input frame independently to extract and encode relevant features. These features are then passed to the ConvGRU fusion layer, which is specifically designed to handle sequential data. At each step, the ConvGRU updates its hidden state based on the current input and the preceding state. This recurrent mechanism enables the model to learn relationships both within and across the input frames by effectively capturing both temporal and spatial correlations. One of the strengths of this approach is its flexibility to accommodate varying numbers of input frames of any size without needing architectural changes. Following the sequential fusion, Global Average Pooling (GAP) is applied along the temporal dimension to create a fixed-size aggregated feature representation. Finally, the decoder upsamples this fused representation to reconstruct the high-resolution output image.

By using convGRU as a fusion layer, a notable strength of MISRGRU is its flexibility to allow the model to be trained on datasets containing different numbers of input images, enabling the integration of heterogeneous experimental data. Therefore, the model is readily extendible to different number of input frames, and different magnification factors using higher order SOFI targets. We believe that this adaptability holds promise for training a more generalized model capable of capturing diverse biological structures.

In real-time imaging applications, minimizing the latency required to generate a super-resolution (SR) image is critical, as lower latency directly contributes to higher temporal resolution. There are several strategies to improve the latency of the MISRGRU model. The first involves reducing the number of input frames, as discussed in the previous section. Since the network processes frames sequentially, minimizing the required input frames is key to reducing inference time. Other important factors include the number of convolutional layers and the number of hidden states per layer. To keep the model's complexity manageable, we maintained the original number of layers and focused on varying the number of hidden states. Each hidden state captures distinct features from the input data, and increasing the number of hidden states enhances the network's capacity to learn more complex or detailed representations required by challenging tasks. While increasing the number of hidden states can improve accuracy and generalization, it also increases computational latency. Therefore, it is good practice to limit the number of hidden states to reduce both inference and training time and to mitigate the risk of overfitting. The results shown in Figure 2 highlight the performance of models with 24 hidden states using different numbers of input frames. To further accelerate the inference time of the 20-frame model, we also evaluated configurations with varying numbers of hidden states, as illustrated in Figure 6. Our results show that reducing the number of hidden states from 24 to 8 decreases latency from 118 ms to 27 ms on our benchmarking system, while still achieving performance comparable to the model with 24 hidden states.

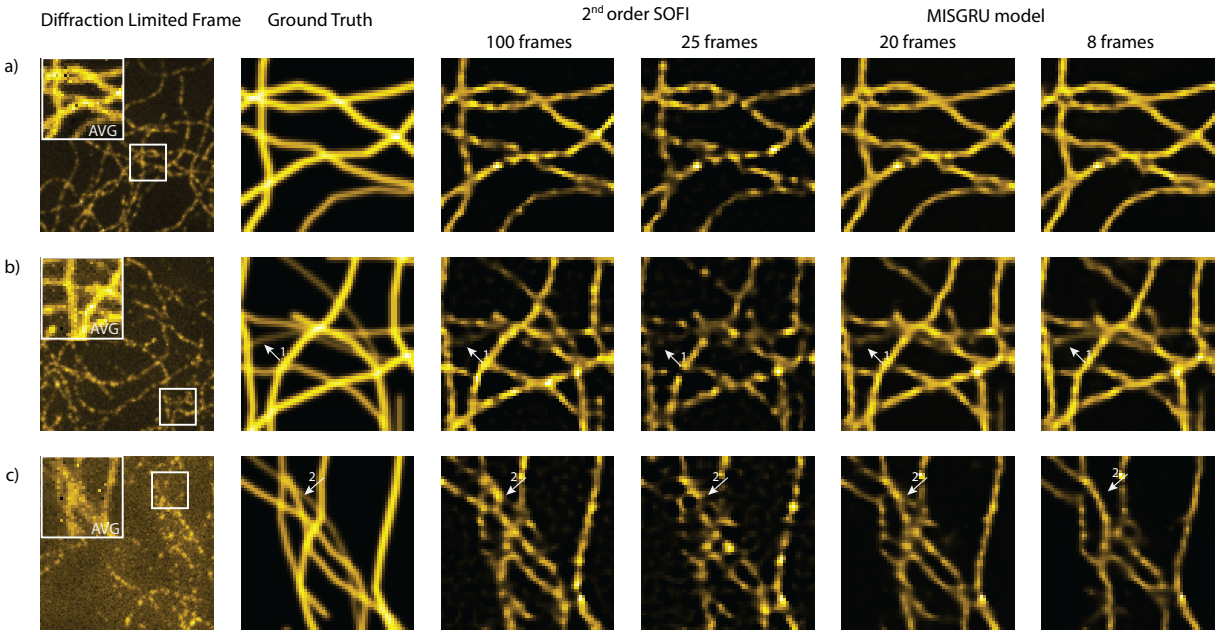


Fig. 5 Comparison of SR reconstructions for different simulated blinking conditions between the GT, second-order SOFI, and the MISRGRU models. Each row represents different off times (on time: 10 ms, off time: 600, 1200, 2400 ms). Left to right: single diffraction-limited frame of a simulated microtubule, ground truth for two-fold resolution improvement; SOFI SR image based on 100 frames and 25 frames, model-based SR image based on 20 frames and 8 frames. White arrows show the degradation in structural detail.

An alternative to MISRGRU is the widely used U-Net architecture, which has become a cornerstone in biomedical image analysis. Originally introduced in 2015 for image segmentation tasks [17], U-Net features a fully convolutional, U-shaped architecture with symmetric encoder-decoder pathways and skip connections that preserve high-resolution spatial features. This design significantly enhances segmentation accuracy and has contributed to U-Net's widespread adoption across various image processing tasks, including denoising, deconvolution, and image registration [18, 21, 31]. Given its success in related domains, U-Net provides a valuable point of comparison for evaluating the performance, efficiency, and flexibility of MISRGRU in super-resolution imaging applications.

Although U-Net can process multi-channel inputs—where each channel may correspond to a low-resolution image from a different time point [18, 22]—its architecture is inherently designed to capture spatial features rather than temporal dynamics since U-Net only learns 2D convolutional filters. As a result, it lacks the capability to model temporal dependencies, which are often crucial for tasks involving fluctuation-based super-resolution imaging. Additionally, U-Net requires a fixed number of input channels and an input image size that matches the target super-resolution output, which limits its adaptability and general use across variable data formats.

In our benchmarking setup, U-Net achieves an inference runtime of 46 ms for an input field of view (FOV) of 1024×1024 pixels. However, to satisfy the Nyquist criterion for two-fold resolution enhancement, the input data must first be upsampled through a Fourier-based interpolation step prior to training — a common practice in other super-resolution applications of U-Net [22]. When this GPU-accelerated preprocessing step is included, the total latency increases to 73 ms for 20 input frames and 54 ms for 8 input frames.

The results, summarized in Extended Figure 5, report average latency across different models, providing a more detailed comparison between MISRGRU, U-Net and SOFI method. Despite the inherent complexity of sequential data processing

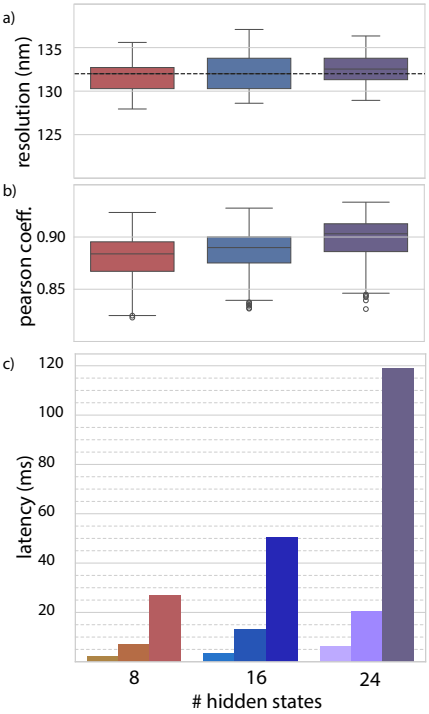


Fig. 6 Performance evaluation of MISRGRU models configured with different numbers of hidden states: 8 (red), 16 (blue), and 24 (purple), all using a 20-frame input. Panels (a) and (b) show boxplots of the resolution-scaled Pearson correlation coefficient and spatial resolution, estimated via image decorrelation analysis. The dashed line in (a) indicates the theoretical half-resolution limit of the wide-field average image. Panel (c) displays the average inference latency over 500 repetitions; from left to right, each bar represents a different field of view (FOV) size: 128×128, 256×256, and 512×512 pixels.

in recurrent neural networks, the optimized MISRGRU model achieved a significantly lower end-to-end latency of 27 ms. This represents an approximate 400-fold acceleration in super-resolution reconstruction compared to the SOFI method. These findings highlight the strong potential of MISRGRU for real-time, live-cell super-resolution imaging.

2.2.2 Generalization of the Model

Along with filament-like structures, we evaluated the network on circular patterns, incorporating varying blinking characteristics and higher background levels. As illustrated in Extended Figure 2, both the 8-frame and 20-frame models successfully reconstruct the circular structures, despite not being explicitly trained under high background conditions. Notably, the 20-frame model more effectively mitigates reconstruction artifacts that are more evident in the 8-frame output. These findings highlight the MISRGRU model's strong potential for generalization across different biological structures.

2.2.3 Synthetic Movement

In addition to static structures, we evaluated the network's performance under dynamic conditions to better reflect live-cell imaging scenarios. Microtubules can exhibit a wide range of movement speeds depending on the cellular context [32, 33]. For our simulations, we selected a fast-motion scenario, modeling a vertically aligned filament moving horizontally at a speed of 2 μm/s, equivalent to 20 nm per frame. As shown in Figure 7, averaging over time leads to increased motion blur in the reconstructed images. This effect is especially evident in the 20-frame model, where the filaments appear thicker and more blurred due to accumulated displacement over time. In contrast, the 8-frame model yields a sharper reconstruction, as the shorter temporal window minimizes the impact of motion. These findings suggest that limiting the input to a few number of frames could be a practical solution for high-quality, high-frame-rate movie reconstruction in dynamic conditions.

2.3 Evaluating the Network on Experimental Microtubule Imaging

The 20-frame model, previously validated on synthetic datasets, was further tested on fixed-cell microtubule samples to evaluate its performance on experimental data and examine its consistency with simulation results. We recently showed that the DNA-PAINT method can be used to tailor fluorescent blinking to achieve both SMLM and SOFI imaging towards high-throughput multimodal super-resolution imaging [34]. The tunable blinking based on the hybridization of a single-stranded, fluorescently labeled DNA imager to a complementary target-bound DNA strand also provides a flexible benchmarking platform for training and testing the network under varying fluorescence dynamics. As no ground-truth (GT) image is available in this setting, we employed resolution-scaled Pearson correlation (RSP) to

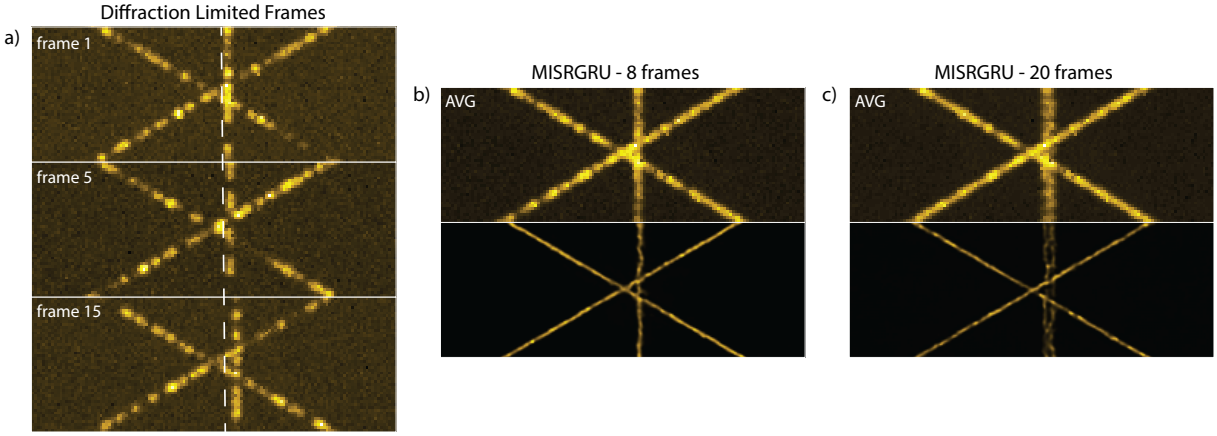


Fig. 7 Testing the models with synthetic movement of a vertical line with a speed of 2 $\mu m/s$. From left to right : Movement in diffraction-limited frames (a), averaged frames and corresponding predictions using MISRGRU models based on 8 frames (b), and 20 frames input(c)

assess structural fidelity, using the standard deviation projection of the input frames as the reference image [29]. Spatial resolution was evaluated using image decorrelation analysis [28], following the same procedure applied in the simulation studies. In Extended Figure 4, we present a comparison of the RSP performance between SOFI reconstructions—using different numbers of input frames—and the 20-frame MISRGRU model. The results show that both the SOFI and the 20-frame model closely approximate the theoretical two-fold resolution enhancement.

Using pre-trained weights from synthetic data allows us to leverage learned features and patterns, speeding up the training process and improving generalization, especially when constrained by limited microscope data [35] [25, 36]. We train the model on a dataset consisting of fixed-cell microtubules organized into six sets. Three of these sets have relatively high SNR and dense fluorophores, while the other three have lower SNR and sparser fluorophores (see Methods). We created ground truth SOFI images by using all the frames that the experiment provides. As DNA-PAINT conditions, we chose fast-blinking probe and buffer conditions, and most of the probes reach the on-state at least once shortly after the experiment starts. Using this assumption, we trained the network to learn well-sampled structures and the highest reachable SNR condition.

The results are presented in Figure 8. Similar to the simulations, the model effectively approximated the filaments. Because of the training scheme we used for the model, it estimates the connected structures in the filaments, filling the smaller gaps between fluorophores. This behavior can be attributed to the fact that half of the dataset has a sparser fluorophore distribution, while the other half is very dense. This mix encourages the model to make more approximations during training to minimize sampling loss, particularly when using only 20 frames, where some labels are never in the on-state and thus "lost". As a result, there are more filament approximations, yet larger gaps between disconnected filaments are preserved (see average widefield in figure 8(a-b)).

3 Discussion

Live-cell imaging aims to achieve high spatial and temporal resolution while minimizing phototoxicity and sample damage. Although single-molecule localization microscopy (SMLM) is widely employed for both imaging and tracking at the molecular scale, it compromises temporal resolution and imposes stress on the sample due to the prolonged acquisition times. As a faster alternative, fluctuation-based imaging techniques enhance resolution using only tens to hundreds of frames, offering a more favorable trade-off between resolution and imaging speed while better preserving sample health throughout the experiment.

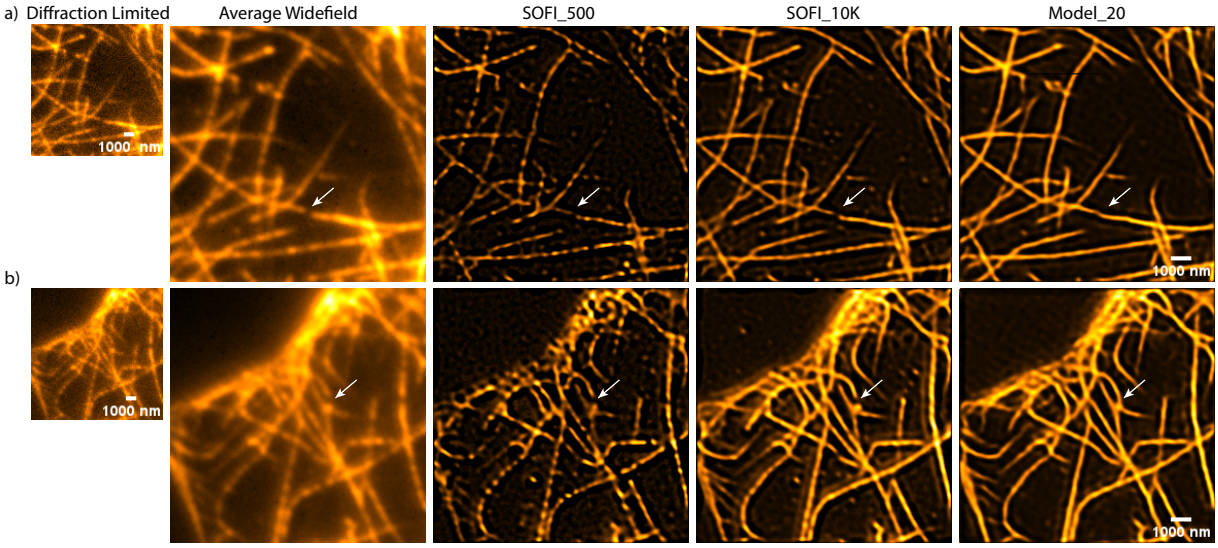


Fig. 8 Comparison of the 20-frame MISRGRU model and SOFI reconstructions using varying numbers of frames in DNA-PAINT microtubule experiments. Panels (a–b), from left to right, show: a single diffraction-limited frame of a microtubule; the average widefield image based on 20 frames, upsampled via bilinear interpolation to match the resolution of the SOFI and model outputs; SOFI-based super-resolution (SR) reconstruction using 500 frames; SOFI-based SR reconstruction using 10,000 frames; and the MISRGRU model's SR reconstruction using 20 frames. White arrows highlight regions where the model reconnects filaments that appear disconnected in the average widefield. Scale bar: 1000 nm.

To advance this approach, we propose the MISRGRU architecture, a deep learning model designed to enhance both spatial and temporal resolution while remaining adaptable to a range of imaging conditions. We evaluated its performance using SOFI targets, as SOFI offers advantages for live-cell imaging, including compatibility with denser fluorophore labeling, operating under lower SNR conditions, and reduced frame requirements compared to SMLM. Despite these strengths, instantaneous SOFI is still limited in temporal resolution due to its computational demands. To overcome this, we optimized and trained MISRGRU using synthetic and real datasets to reconstruct second-order SOFI images with approximately 400-fold acceleration in processing time. The developed deep-learning model not only enables real-time imaging but also reduces background artifacts while approaching the theoretical limits of spatial resolution.

The MISRGRU model leverages temporal information to effectively capture correlated fluorophore blinking events. Using synthetic microtubule datasets, we observed that MISRGRU achieves a favorable balance between spatial resolution, structural fidelity, and temporal performance. Under low signal-to-noise ratio (SNR) conditions, MISRGRU outperforms SOFI, suggesting its potential for live-cell imaging with reduced laser intensity, thereby minimizing photodamage and allowing for longer imaging sessions. In dynamic imaging scenarios, using a smaller number of input frames, in this case 8, offers a practical advantage, as increasing the frame count can lead to motion blur due to structural displacement over time. This enables the reconstruction of high-quality, high-frame-rate movies that preserve spatial clarity in fast-moving scenes. With further optimization, particularly by incorporating dynamic training data, MISRGRU could be extended for longer input sequences while maintaining temporal precision. Additionally, its reliable reconstruction performance could be utilized as a valuable pre-processing tool for downstream applications such as segmentation, classification, or rare event detection.

Moreover, the models were able to reconstruct hollow circles, showing the potential for generalizing to different structures. However, the generalizability of the model across a wider array

of structural geometries and different background levels should be quantitatively assessed. To further improve the model's robustness, especially under challenging imaging conditions such as those encountered in live-cell cytosolic environments, several strategies can be pursued. One promising direction is to train the network with simulations that incorporate higher background levels, more accurately reflecting the noise characteristics typical of in vivo imaging. Alternatively, the quality of SOFI targets used during training can be enhanced through advanced post-processing techniques—such as improved linearization—and through sophisticated pre-processing methods, including denoising and background subtraction. These improvements are especially critical in ultralow-SNR regimes, where SOFI reconstructions tend to degrade in quality. Together, these strategies would pave the way toward applying MISRGRU for live-cell imaging scenarios involving fast-bleaching, such as samples labeled with fluorescent proteins, offering a promising route for high-speed, high-fidelity super-resolution microscopy.

We first pre-trained the model on synthetic datasets and then fine-tuned it using a small number of fixed-cell samples containing microtubule structures. Leveraging pre-trained weights from synthetic data enabled the model to retain useful feature representations, thereby accelerating convergence and enhancing generalization. This is particularly valuable when only limited experimental microscopy data is available. The model demonstrated comparable performance across real and synthetic data in terms of Pearson correlation, spatial resolution, and suppression of background artifacts. However, in half of the experimental datasets, slower blinking dynamics led to approximation of filament structures by filling the small gaps. To address this, future trainings should include optimizing blinking kinetics—favoring faster blinking conditions—to enable more effective sampling of structural features within short acquisition times and thereby improve structural fidelity while training with fewer number of input frames. Additionally, future work should focus on further validating and applying the network to experimental datasets. A promising direction involves multi-modal DNA-PAINT experiments, which, as shown in [34], allow controlled switching between SMLM and SOFI imaging modes through calibrated blinking conditions. This capability makes it possible to generate SMLM-based ground truth images that correspond to SOFI acquisitions, offering a more effective framework for quantitatively evaluating the model's super-resolution reconstructions.

We benchmarked MISRGRU against widely used CNN architecture, U-Net. Recently, Chen et al. employed a U-Net architecture to accelerate SRRF imaging [22]. They trained the U-Net using a combination of experimental data and simulated noise, using 200 raw frames to reconstruct SRRF images as the target image. To optimize network performance, they tested various input frame counts, achieving a reconstruction speed of 10 frames per second with 5 input frames, improving the resolution of total internal reflection microscopy (TIRFM) images to 266 nm. However, their study did not extend to simulations for evaluating image quality under different blinking and SNR conditions, and as a result, potential artifacts introduced by the model and the SRRF method were not evaluated. In another study, U-Net was used to speed up SIM reconstruction by reducing the number of input frames under low SNR conditions [21]. While this approach improved the temporal resolution and image quality for live-cell applications, SIM is fundamentally limited to a two-fold resolution enhancement.

Comparative analysis across multiple evaluation metrics revealed that both U-Net and MISRGRU deliver similar performance under the tested conditions. However, U-Net's symmetric encoder-decoder structure requires resizing the input to match the output dimensions, which significantly increases both training and inference time. In contrast, MISRGRU eliminates the need for such pre-processing and supports variable numbers of input frames during both training and inference enabling both flexibility in dataset and computational efficiency. This adaptability is particularly advantageous for targeting higher-order SOFI reconstructions, which can potentially surpass the two-fold resolution limit. Moreover, MISRGRU's design makes it better suited for future extensions involving heterogeneous datasets and diverse biological structures, facilitating generalization across experimental conditions without extensive data pre-processing.

The original MISRGRU design also incorporates an additional encoder unit for processing

a reference image, enabling the network to co-register input frames within the encoded feature space and correct for pixel-level misalignments. This capability is especially beneficial in scenarios that require drift correction. In our current study, we did not utilize this feature, as drift was negligible during the 20-frame acquisition window. However, for future applications involving longer sequences or time-lapse imaging, this built-in co-registration mechanism could eliminate the need for separate image alignment during data preparation.

In summary, we present a recurrent neural network-based architecture tailored for real-time super-resolution imaging within the microscopy domain. The proposed model is capable of reconstructing high-quality super-resolution images from as few as eight input frames, while minimizing background artifacts and achieving a two-fold spatial resolution enhancement near the theoretical limit. In addition to image reconstruction, the model serves as a powerful pre-processing tool for downstream tasks such as image segmentation, classification, and rare event detection. By leveraging pre-trained weights from synthetic datasets and fine-tuning with fixed-cell microscopy data, the network is well-suited for live-cell imaging scenarios. Thanks to its low latency, the model holds strong potential for real-time sample screening and for enhancing spatio-temporal resolution during live experiments.

Methods

3.1 Simulations

The SOFI simulation tool was adapted to more accurately represent noise and background effects, following the methodology described in [37]. To account for electronic noise, a gain and standard deviation map of the sCMOS camera used in our experimental setup was generated and incorporated into the simulation [38, 39]. In addition, a non-stationary background model was implemented in which background signals arise from out-of-focus blinking emitters. Inspired by the approach used in SACD [12], the distribution of emitters in the focal plane was geometrically shifted and transformed to generate a corresponding background distribution, which was subsequently convolved with a broader point spread function (PSF) to simulate out-of-focus blur.

The simulation begins by generating a binary mask of the target structure, which may consist of either filamentous or annular patterns. These patterns were produced by modifying the pattern generation algorithms originally developed for the DBlink and SACD methods. Emitters were randomly distributed across the simulation area at a density of 5000 fluorophores/μm^2, and only those overlapping with the binary mask were retained. The spatial grid for emitter placement was set to 1280×1280 pixels with a resolution of 10 nm per pixel. Filamentous structures were simulated with thicknesses ranging from 10 nm to 40 nm, while annular structures were assigned a fixed thickness of 20 nm and radii between 10 nm and 40 nm.

The temporal blinking behavior of individual emitters was simulated using a continuous-time Markov process [37] , resulting in high-resolution video sequences. The simulations assumed an emitter on-time of 10 ms, with off-times set to 600 ms, 1200 ms, or 2400 ms to model different blinking densities. Signal-to-noise ratio (SNR) was further controlled by adjusting the Ion value, which represents the average number of photons emitted per frame by each fluorophore. Ion values of 100, 60, and 40 were used to reflect different SNR conditions, mimicking the photophysics of fluorescent proteins under low illumination.

Both in-focus and out-of-focus emitter planes were convolved with Gaussian PSFs, having full-width at half-maximum (FWHM) values of 220 nm and 2000 nm, respectively. The maximum background contribution per pixel was calculated based on the mean brightness of fluorophores in the focal plane, using a fixed ratio between in-focus and out-of-focus signals. The resulting noise-free image stack was then downsampled by a factor of 10. Finally, Poisson-distributed photon noise and sCMOS-specific electronic noise were added to each frame to produce the final simulated dataset.

3.2 Loss Function

Spatial domain loss functions such as mean absolute error is commonly used in super-resolution

reconstruction tasks. Fuoli et al. combined spatial loss with Fourier domain loss to improve the restoration of high-frequency components in super-resolution methods [40]. For simplicity of training, we compared only the spatial loss function (L_1) and the Fourier domain loss function ($L_{\mathcal{F}}$).

$$L_1 = \left|\hat{Y}_{u,v} - Y_{u,v}\right| \tag{1}$$

$$L_{\mathcal{F}} = L_{\mathcal{F}_A} + L_{\mathcal{F}_\angle} \tag{2}$$

$$L_{\mathcal{F}_\mathcal{A}} = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \left| |\hat{Y}|_{u,v} - |Y|_{u,v} \right| \tag{3a}$$

$$L_{\mathcal{F}_\angle} = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \left| \angle\hat{Y}_{u,v} - \angle Y_{u,v} \right| \tag{3b}$$

Here, both SR images $\hat{Y}$ and the target image Y are transformed into the Fourier space by applying fast Fourier transform (FFT), where the absolute amplitude difference $L_{\mathcal{F}_\mathcal{A}}$ and absolute phase difference $L_{\mathcal{F}_\angle}$ are calculated. Due to symmetry in the Fourier space (Hertimitian symmetry), only half of the spectral components is considered.

3.3 Training Scheme

We generated simulated datasets of 128×128 pixel images incorporating a range of SNR levels and blinking conditions, as detailed in the Simulations section. The dataset was split into training (2,000 samples), evaluation, and test sets (480 samples each). Model training was performed using the Adam optimizer with a learning rate (η) of 0.001 and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Xavier initialization was used for the model weights [41], as it is well-suited for networks utilizing sigmoid activation functions. To mitigate overfitting, early stopping was applied with a maximum of 400 training epochs and a patience threshold of 10. A batch size of 10 was selected to balance GPU memory usage and computational efficiency. All training and testing were conducted on an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory.

To extend the model to experimental data, we used the pre-trained weights from the simulated dataset as initialization for fine-tuning on DNA-PAINT measurements. Four independent recordings of the same cell type were used, and data augmentation techniques—such as random cropping and rotation—expanded the dataset to approximately 2,000 samples. This transfer learning approach facilitates the model's applicability to live-cell imaging, enhancing its practicality for real-world experimental scenarios.

To further reduce inference latency, we applied TensorRT acceleration to all models trained within the PyTorch framework [42]. This optimization included techniques such as layer fusion and precision calibration, including the use of lower-precision formats like fixed-point floating point, to maximize computational efficiency.

3.4 DNA-PAINT Imaging

COS-7 cells (DSMZ GmbH) were cultured in high-glucose Dulbecco's Modified Eagle Medium (DMEM; ThermoFisher), supplemented with 10% fetal bovine serum (FBS; Gibco, ThermoFisher), 1% sodium pyruvate (Gibco, ThermoFisher), 1% L-glutamine (Gibco, ThermoFisher), and 1% Penicillin-Streptomycin (Gibco, ThermoFisher). Cells were seeded in 24 mm high-precision cover glasses (Carl Roth) placed in a 6-well plate or in μ-Slide 8-well high glass bottom chambers (Ibidi). Following seeding, cells were incubated overnight at 37°C in a humidified atmosphere containing 5% CO.

Once the cultures reached moderate confluency with a clear presence of individual cells, fixation was initiated. Cell extraction was performed at room temperature for 90 seconds using a pre-warmed (37°C) extraction buffer containing 0.3% (v/v) Triton X-100 and 0.25% (w/v) glutaraldehyde in microtubule-stabilizing buffer (Kapitein). The extraction buffer was subsequently replaced with fixation buffer, and cells were incubated for 10 minutes at room temperature. After fixation, cells were washed with PBS and stored in PBS containing 50% (v/v) glycerol at 4°C for up to three days prior to quenching. The quenching protocol followed the method described in [34].

Fixed cells were imaged in an imaging buffer consisting of PBS with 500 mM NaCl and varying

concentrations of imager strands, at room temperature. In experiments assessing the effect of ethylene carbonate (EC), 5% (v/v) EC (Fisher Scientific) was added to the imaging buffer. Imaging was performed on a custom-built microscope equipped with total internal reflection fluorescence (TIRF) illumination using a 638 nm laser. A total of 10,000 frames were acquired with a 30 ms exposure time using a BSI Express sCMOS camera. Fluorescence from Atto 655 was filtered using a bandpass emission filter (ET706/95m, AHF Analysentechnik AG). Full specifications of the custom-built microscope are available on the Grussmayer Lab github page.

3.5 Performance evaluation

We evaluate the performance of SOFI reconstructions and AI-generated predictions using three complementary methods. First, we estimate image resolution through image decorrelation analysis, a technique that operates on a single image and relies on partial phase correlation to determine spatial resolution [28]. Second, we employ the resolution-scaled Pearson correlation (RSP) metric to quantify the similarity between the super-resolution (SR) image and a reference image [43]. For experimental data, the reference is the standard deviation projection of the image sequence, while for simulations, the ground truth is used. Prior to correlation computation, the SR image is adjusted to match the resolution of the reference image using an estimated resolution-scaling function (RSF), providing a more robust comparison than traditional metrics like the structural similarity index (SSIM).

Lastly, we use Li thresholding to binarize simulated images and assess pixel-level agreement between predictions and ground truth [44, 45]. From these binarized comparisons, we construct confusion matrices for each model, reporting true positive (TP), false positive (FP), true negative (TN), and false negative (FN) rates to quantify classification performance.

Acknowledgements

We thank the TU Delft AI lab initiative for the support of BIOlab.

Contributions

M.T, J.K and K.S.G conceived the idea. M.T and J.K designed the simulations and evaluation. J.K tested U-Net and MISGRU network designs and optimized them using SOFI simulations and DNA-paint experiments. M.T performed further analysis to test latency, fast moving simulations and generalization of the models. B.K.Z and R.H. designed and performed the DNA-paint experiments. M.T and J.K wrote the manuscript. K.S.G an N.T reviewed the manuscript. N.T supervised the deep network implementation and evaluation. K.S.G supervised the microscopy experiments, data analysis and network evaluation.

References

[1] Pylvänäinen, J. W., Gómez-de Mariscal, E., Henriques, R. & Jacquemet, G. Live-cell imaging in the deep learning era. *Current Opinion in Cell Biology* **85**, 102271 (2023). URL https://www.sciencedirect.com/science/article/pii/S0955067423001205.

[2] Hell, S. W. & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics Letters* **19**, 780 (1994). URL https://opg.optica.org/abstract.cfm?URI=ol-19-11-780.

[3] Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy: SHORT COMMUNICATION. *Journal of Microscopy* **198**, 82–87 (2000). URL https://onlinelibrary.wiley.com/doi/10.1046/j.1365-2818.2000.00710.x.

[4] Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science* **313**, 1642–1645 (2006). URL https://www.science.org/doi/10.1126/science.1127344.

[5] Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods* **3**, 793–796 (2006). URL https://www.nature.com/articles/nmeth929.

[6] Zheng, X. *et al.* Current challenges and solutions of super-resolution structured illumination microscopy. *APL Photonics* **6**, 020901 (2021). URL https://doi.org/10.1063/5.0038065.

[7] Jacquemet, G., Carisey, A. F., Hamidi, H., Henriques, R. & Leterrier, C. The cell biologist's guide to super-resolution microscopy. *Journal of Cell Science* **133**, jcs240713 (2020). URL https://doi.org/10.1242/jcs.240713.

[8] Dertinger, T., Colyer, R., Iyer, G., Weiss, S. & Enderlein, J. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proceedings of the National Academy of Sciences* **106**, 22287–22292 (2009). URL https://pnas.org/doi/full/10.1073/pnas.0907866106.

[9] Pawlowska, M., Tenne, R., Ghosh, B., Makowski, A. & Lapkiewicz, R. Embracing the uncertainty: the evolution of sofi into a diverse family of fluctuation-based super-resolution microscopy methods. *Journal of Physics: Photonics* **4**, 012002 (2021).

[10] Basak, S. *et al.* Super-resolution optical fluctuation imaging. *Nature Photonics* 1–9 (2025).

[11] Gustafsson, N. *et al.* Fast live-cell conventional fluorophore nanoscopy with ImageJ through super-resolution radial fluctuations. *Nature Communications* **7**, 12471 (2016). URL https://www.nature.com/articles/ncomms12471.

[12] Zhao, W. *et al.* Enhanced detection of fluorescence fluctuations for high-throughput super-resolution imaging. *Nature Photonics* **17**, 806–813 (2023).

[13] Agarwal, K. & Macháň, R. Multiple signal classification algorithm for super-resolution fluorescence microscopy. *Nature communications* **7**, 13752 (2016).

[14] Geissbuehler, S. *et al.* Mapping molecular statistics with balanced super-resolution optical fluctuation imaging (bSOFI). *Optical Nanoscopy* **1**, 4 (2012). URL http://optnano.springeropen.com/articles/10.1186/2192-2853-1-4.

[15] Laine, R. F. *et al.* High-fidelity 3D live-cell nanoscopy through data-driven enhanced super-resolution radial fluctuation. *Nature Methods* 1949–1956 (2023).

[16] Cunha, I., Latron, E., Bauer, S., Sage, D. & Griffié, J. Machine learning in microscopy–insights, opportunities and challenges. *Journal of cell science* **137** (2024).

[17] Ronneberger, O., Fischer, P. & Brox, T. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F. (eds) *U-net: Convolutional networks for biomedical image segmentation.* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241 (Springer International Publishing, Cham, 2015).

[18] Jin, L. *et al.* Deep learning enables structured illumination microscopy with low light levels and enhanced speed. *Nature Communications* **11**, 1934 (2020). URL https://www.nature.com/articles/s41467-020-15784-x.

[19] Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nature Biotechnology* **36**, 460–468 (2018). URL https://www.nature.com/articles/nbt.4106.

[20] Qu, L. *et al.* Self-inspired learning to denoise for live-cell super-resolution microscopy. URL http://biorxiv.org/lookup/doi/10.1101/2024.01.23.576521.

[21] Speiser, A. *et al.* Deep learning enables fast and dense single-molecule localization with high accuracy. *Nature Methods* **18**, 1082–1090 (2021). URL https://www.nature.com/articles/s41592-021-01236-x.

[22] Chen, J. *et al.* Deep-learning accelerated super-resolution radial fluctuations (SRRF) enables real-time live cell imaging. *Optics and Lasers in Engineering* **172**, 107840 (2024).

URL https://www.sciencedirect.com/science/article/pii/S014381662300369X.

[23] Nehme, E., Weiss, L. E., Michaeli, T. & Shechtman, Y. Deep-storm: super-resolution single-molecule microscopy by deep learning. *Optica* **5**, 458–464 (2018). URL https://opg.optica.org/optica/abstract.cfm?URI=optica-5-4-458.

[24] Li, J., Tong, G., Pan, Y. & Yu, Y. Spatial and temporal super-resolution for fluorescence microscopy by a recurrent neural network. *Opt. Express* **29**, 15747–15763 (2021). URL https://opg.optica.org/oe/abstract.cfm?URI=oe-29-10-15747.

[25] Saguy, A. *et al.* DBlink: dynamic localization microscopy in super spatiotemporal resolution via deep learning **20**, 1939–1948. URL https://www.nature.com/articles/s41592-023-01966-0.

[26] Rifat Arefin, M. *et al.* Multi-image super-resolution for remote sensing using deep recurrent networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 816–825 (2020).

[27] Deschout, H. *et al.* Complementarity of palm and sofi for super-resolution live-cell imaging of focal adhesions. *Nature communications* **7**, 13693 (2016).

[28] Descloux, A., Grußmayer, K. S. & Radenovic, A. Parameter-free image resolution estimation based on decorrelation analysis. *Nature Methods* **16**, 918–924 (2019). URL https://www.nature.com/articles/s41592-019-0515-7.

[29] Culley, S. *et al.* Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods* **15**, 263–266 (2018). URL https://www.nature.com/articles/nmeth.4605.

[30] Lepcha, D. C., Goyal, B., Dogra, A. & Goyal, V. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion* **91**, 230–260 (2023).

[31] Lu, Z. & Chen, Y. Single image super-resolution based on a modified u-net with mixed gradient loss. *signal, image and video processing* **16**, 1143–1151 (2022).

[32] Wang, L. & Brown, A. Rapid movement of microtubules in axons. *Current Biology* **12**, 1496–1501 (2002).

[33] Taba, T. *et al.* Direction and speed of microtubule movements driven by kinesin motors arranged on catchin thick filaments. *Cell Motility and the Cytoskeleton* **65**, 816–826 (2008).

[34] de Zwaan, K. *et al.* High-throughput single-molecule microscopy with adaptable spatial resolution using exchangeable oligonucleotide labels. *ACS nano* (2024).

[35] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org.

[36] Demircan-Tureyen, E., Akbulut, F. P. & Kamasak, M. E. Restoring fluorescence microscopy images by transfer learning from tailored data. *IEEE Access* **10**, 61016–61033 (2022).

[37] Girsault, A. *et al.* Sofi simulation tool: a software package for simulating and testing super-resolution optical fluctuation imaging. *PLoS One* **11**, e0161602 (2016).

[38] Zhang, Z., Wang, Y., Piestun, R. & Huang, Z.-L. Characterizing and correcting camera noise in back-illuminated scmos cameras. *Optics express* **29**, 6668–6690 (2021).

[39] Liu, S. *et al.* scmos noise-correction algorithm for microscopy images. *Nature methods* **14**, 760–761 (2017).

[40] Fuoli, D., Gool, L. V. & Timofte, R. Fourier space losses for efficient perceptual image super-resolution. *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 2340–2349 (2021).

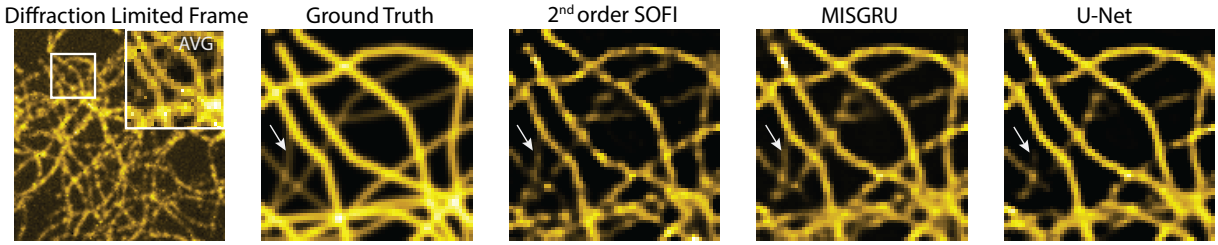
[41] Glorot, X. & Bengio, Y. Teh, Y. W. & Titterington, M. (eds) *Understanding the difficulty of training deep feedforward neural networks.* (eds Teh, Y. W. & Titterington, M.) *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Vol. 9 of *Proceedings of Machine Learning Research*, 249–256 (PMLR, 2010). URL https://proceedings.mlr.press/v9/glorot10a.html.

[42] Pytorch. Tensor rt. https://github.com/pytorch/TensorRT (2024).

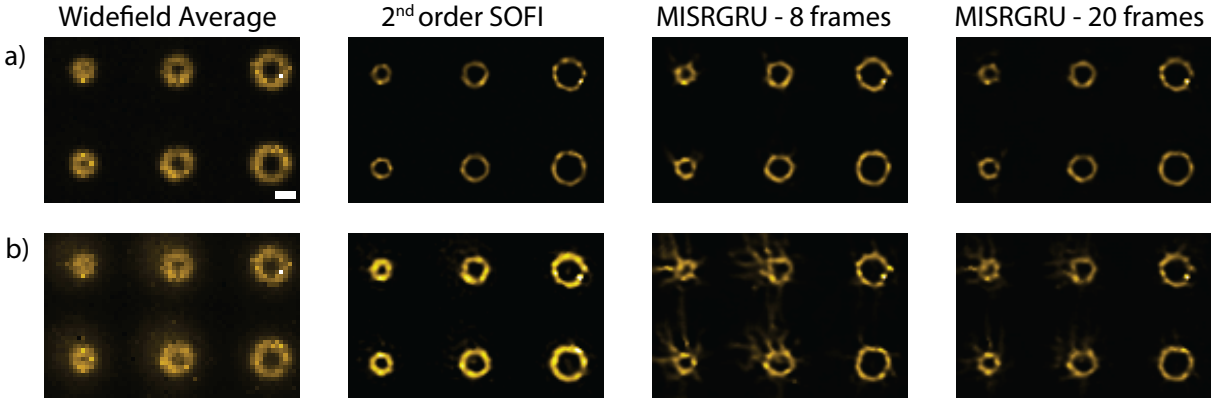
[43] Culley, S. *et al.* Author correction: Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods* **17**, 1167–1167 (2020).

[44] Li, C. & Lee, C. Minimum cross entropy thresholding **26**, 617–625. URL https://linkinghub.elsevier.com/retrieve/pii/003132039390115D.

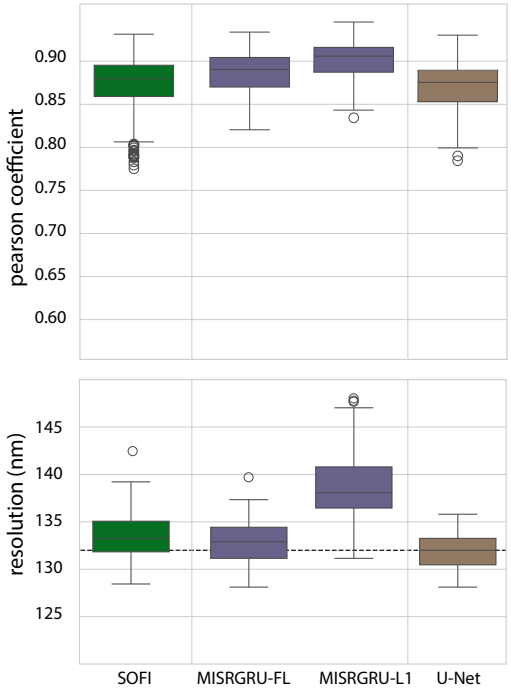
[45] Li, C. & Tam, P. An iterative algorithm for minimum cross entropy thresholding **19**, 771–776. URL https://linkinghub.elsevier.com/retrieve/pii/S0167865598000579.



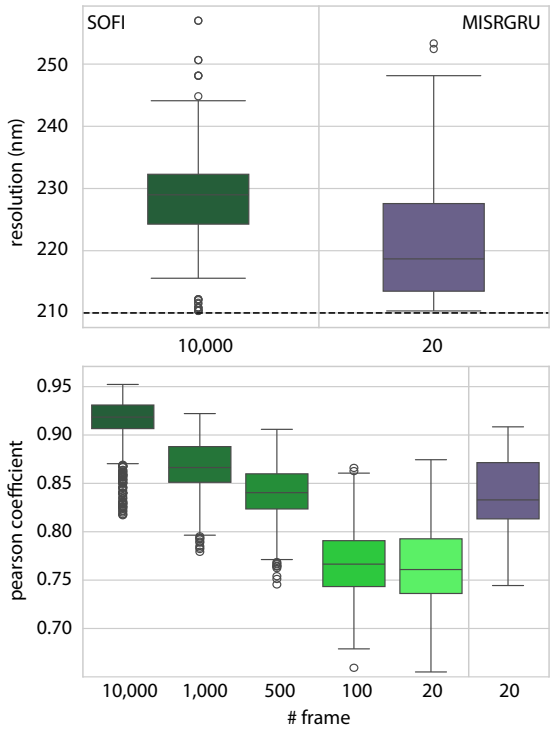
Extended Fig. 1 Comparison of super-resolution (SR) reconstructions of simulated microtubule structures using ground truth (GT), second-order SOFI using 100 frames, and the MISRGRU and U-Net models, based on 20 input frames.



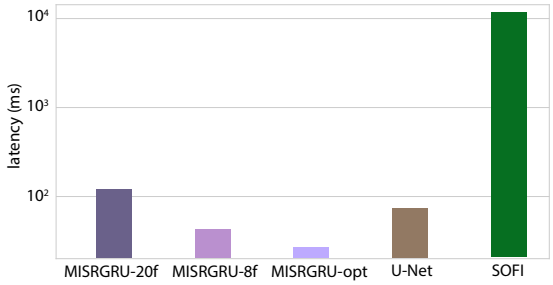
Extended Fig. 2 Testing the generalization of MISRGRU models based on 8 and 20 frames for simulated circular structures. For (a)low background and (b) high background conditions



Extended Fig. 3 Performance evaluation of MISRGRU using Fourier loss (FL), MISRGRU using L1 loss (L1) and U-Net using L1 loss in comparison to SOFI using resolution-scaled pearson correlation and image decorrelation analysis for resolution estimation



Extended Fig. 4 Performance evaluation of MISRGRU for DNA-PAINT experiments. In the top pannel, dashed line corresponds to the half-resolution of the averaged widefield image. Bottom panel compares resolution-scaled pearson coefficient scores of SOFI image reconstruction by using 10k, 1k, 500, 100, 20 frames and MISRGRU model trained with 20-frames input.



Extended Fig. 5 Comparison of computational latency for MISRGRU models, U-Net trained with 20 frames (73 ms) and SOFI reconstruction using 20 frames input (11 s). Different MISRGRU latency values correspond to the models trained using 20 frames (118 ms), 8 frames (41 ms), and optimized model by using with 8 hidden states instead of 24 (27 ms). Values corresponds to the mean of 500 repetition for a FOV of 512x512 pixels