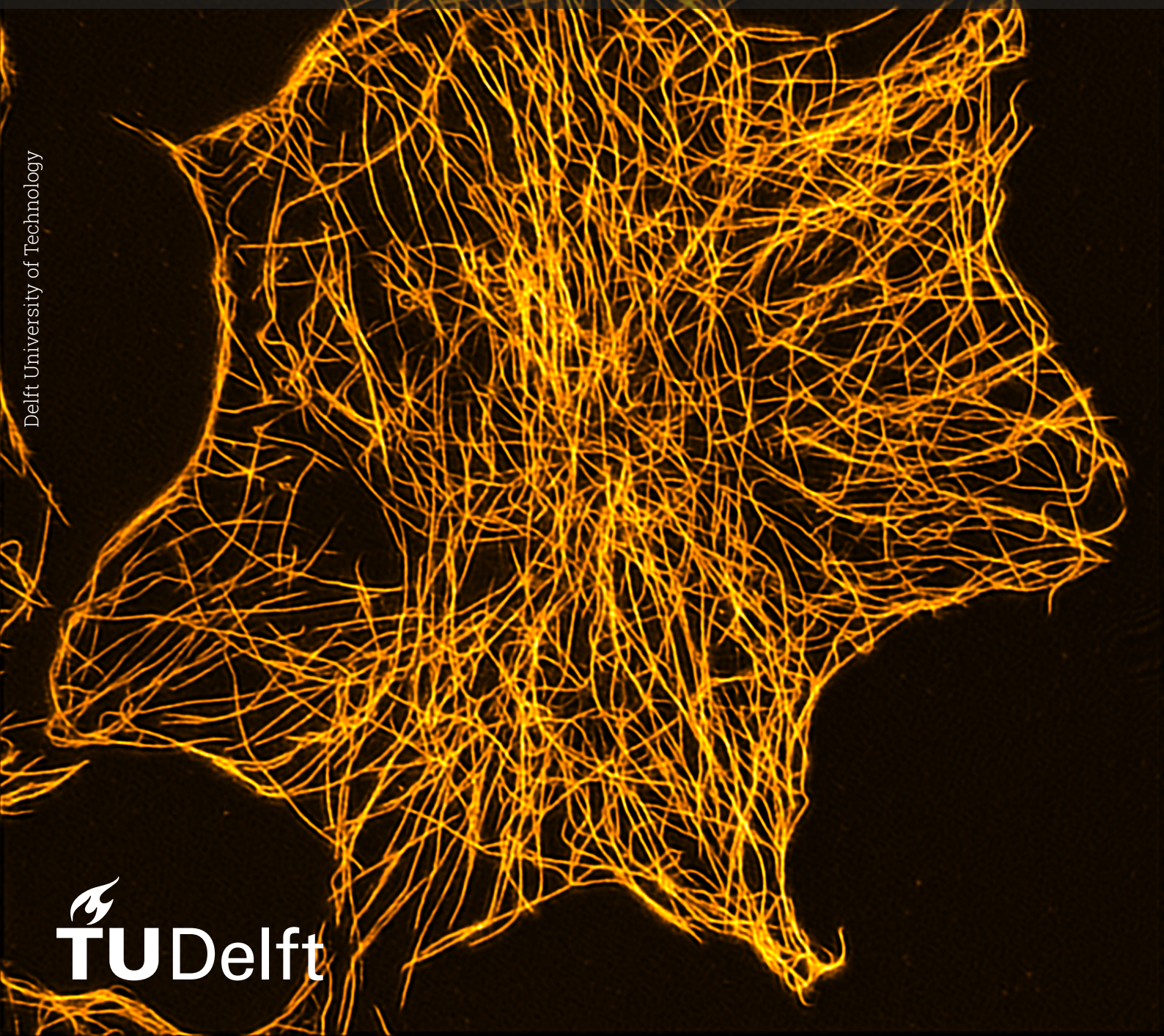
Accelerating SOFI with Deep Learning

Enabling Real-time Live-cell Imaging.

Pattern Recognition and Bioinformatics
Jelle Komen



Delft University of Technology

TU Delft

Accelerating SOFI with Deep Learning

Enabling Real-time Live-cell Imaging.

by

Jelle Komen

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday October 15, 2024 at 08:45 PM.

Student number:	5837995	
Project duration:	January 8, 2024 – October 15, 2024	
Thesis committee:	Dr. J. van Gemert,	TU Delft, Chair
	Dr. R. Marroquim,	TU Delft, Core Member
	Dr. N. Tömen,	TU Delft, Core Member
	M. Tekpınar ,	TU Delft, Member
	Dr. K. Grußmayer ,	TU Delft, External Member

TU Delft

Preface

This work, titled "*Accelerating SOFI with Deep Learning: Enabling Real-Time Live Cell Imaging*", represents my master's thesis for the Master of Science degree in the Embedded Systems program at TU Delft. I embarked on this journey in January 2024 and completed it in October 2024, dedicating this period to researching and writing.

My bachelor thesis was completed at LUMICKS, a company specializing in microscope systems for single-molecule analysis. One lesson from my supervisor that has stayed with me is: "The gold lies at the intersection of two or more domains; the more complex the domains, the more gold you will find." Determined to apply this insight, I explored an image denoising project organized by Ir. Miyase Tekpinar as part of the AI fundamentals course. Fascinated by microscope imaging, I was determined to undertake my master's project with the Grußmayer group, even though I initially lacked some of the necessary knowledge. Thankfully, Dr. Nergis Tömen believed in the potential of this work, and through her and Ir. Miyase Tekpinar guidance, I gained valuable expertise in microscope imaging, deep learning, and image processing. This effort culminated in a poster presentation at the SMLMS conference in Lisbon and a paper planned for publication in *Nature Methods*.

I cannot thank Ir. Miyase Tekpinar and Dr. Kristin Grußmayer enough for their patience, knowledge, and all the opportunities they provided me. I am especially grateful to Ir. Miyase Tekpinar, my co-supervisor, who always took the time to explain things to me and to think through challenges with me. I also want to thank Dr. Nergis Tömen, my supervisor, for believing in me that it could work out and ensuring the work stayed aligned through her critical insights. Additionally, I want to express my gratitude to Dr. Jan van Gemert for taking on the role of thesis advisor and chair, and for his leap of faith in me. Lastly, I would like to thank Dr. Ricardo Marroquim for his interest in the research and his willingness to be a core member of the committee.

Of course, I also want to thank my family, girlfriend, and friends for their unwavering support and for enduring the burden of my attempts to explain my work. This undoubtedly bored them and left them more confused than ever. As a result, the microtubule structures are now affectionately referred to as 'little worms.'

Jelle Komen
Anna Paulowna, August 2024

Abstract

Live-cell imaging captures dynamic cellular behaviors and aims to maximize both spatial and temporal resolution while minimizing sample damage, enabling advancements in fundamental cell biology. However, spatial resolution is limited by the diffraction barrier of optical lenses, which prevents the visualization of many subcellular structures. Single-molecule localization microscopy (SMLM) overcomes this barrier, achieving resolutions as fine as 10 nm, but it typically requires millions of frames and higher illumination, which reduces temporal resolution and can damage the sample. Super-resolution Optical Fluctuation Imaging (SOFI) operates at lower illumination levels and requires hundreds of frames, leveraging the statistical relationships of blinking fluorophores to achieve n-fold spatial resolution based on the nth-order SOFI calculation. Despite its benefits, SOFI still demands too many frames and involves extensive post-processing, making it impractical for real-time live-cell imaging. Without real-time imaging, researchers are unable to make immediate decisions, ultimately costing valuable time for the researchers.

To address this limitation, we introduce a supervised deep learning model designed to accelerate second-order SOFI. Our model reconstructs super-resolved second-order SOFI images using just 20 frames, compared to the hundreds typically required, while maintaining a 2-fold improvement in spatial resolution and showing minimal background artifacts. We demonstrate that after being trained on real fixed-cell (static) mitochondria data, the model is able to reconstruct super-resolved images in a dynamic environment by moving the microscope stage. The model achieves real-time temporal resolutions of up to 4.85 fps, unlocking new possibilities for real-time studies of live-cell dynamics.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ANN	Artificial Neural Network
bSOFI	Balanced SOFI
CARE	Content-aware Image Restoration
CNN	Convolution Neural Network
eSRRF	Enhanced SRRF
FFT	Fast Fourier Transform
FWHM	Full Width at Half Maximum
GRU	Gated Recurrent Unit
HF	High-frequency
HR	High-resolution
LR	Low-resolution
LSTM	Long Short-term Memory
MAE	Mean Absolute Error
MSE	Mean Squared Error
OTF	Optical Transfer Function
PALM	Photo-Activated Localization Microscopy
PSF	Point-spread-function
PSNR	Peak Signal to Noise Ratio
RIM	Random Illumination Microscopy
RNN	Recurrent Neural Network
ROI	Region Of Interest
RSP	Re-scaled Pearson Coefficient
SGD	Stochastic Gradient Descent
SIM	Structured Illumination Microscopy
SMLM	Single Molecule Localization Microscopy
SNR	Signal-to-noise ratio
SOFI	Super-resolution Optical Fluctuation Imaging
SR	Super-resolution
SRRF	Super-Resolution Radial Fluctuations
STED	Stimulated Emission Depletion
STD	Standard Deviation
STORM	Stochastic Optical Reconstruction Microscopy
SSIM	Structural Index Measure
TNR	True Negative Rate
TPR	True Positive Rate

Symbols Microscopy

Symbol	Definition	Unit
λ	Wavelength	[nm]
NA	Numerical aperture	[-]
$d_{x,y}$	Minimum resolvable distance in the lateral directions	[nm]
...		

Symbol	Definition	Unit
K	Number of independent blinking/fluctuating emitters	[-]
r_k	Position of the k-th emitter	[m]
N	Spatial dimensions	[-]
t	Time	[s]
s_k	Temporal intensity vector of the k-th emitter	[-]
ϵ_k	Constant molecular brightness of the k-th emitter	[-]
$x(r,t)$	True image intensity at position r and time t	[-]
$\delta(\cdot)$	Dirac delta function	[-]
$v(r)$	Point spread function (PSF)	[-]
$y(r,t)$	Diffraction-limited image at position r and time t	[-]
y_t	Acquired image at time t (discrete model)	[-]
Mq	Downsampling operator	[-]
U	Convolution operator representing PSF	[-]
x_t	True image at time t (discrete model)	[-]
L	Number of pixels in the finer grid	[-]
M	Number of pixels in the acquisition space	[-]
n_t	Noise component at time t	[-]
q	Factor of grid fineness	[-]
b	Background contribution	[-]
$\boldsymbol{P}(\mathsf{w})$	Realization of a multivariate Poisson variable with parameter w	[-]
$G_2(r,\tau)$	Second-order auto-correlation function of fluorescence intensity	[-]
$\delta y(t)$	Fluctuations in intensity at a given time	[-]
$w_{n,k}(\tau_1,...\tau_{n-1})$	Single-emitter cumulant of the n-th order	[-]
$k_2(r_1,r_2,\tau)$	Second-order cross-cumulant between positions r_1 and r_2	[-]
$\overrightarrow{r}$	Position vector of emitter	[-]
P	Total number of possible partitions in cross-cumulant calculation	[-]
$f_n(\rho_{\mathsf{on}})$	n-th order cumulant of a Bernoulli distribution with probability ρ_{on}	[-]
ρ_{on}	On-time ratio of the emitters	[-]
$\bar{g}$	Linearized cumulant image	[-]
$\hat{g}_n$	Deconvolved n-th order cumulant image	[-]

Symbols Deep Learning

Symbol	Definition	Unit
w	Weight vector	[-]
w_0	Bias term	[-]
$f(x)$	Decision boundary function	[-]
σ	Sigmoid activation function	[-]
$J(\mathsf{w})$	Perceptron cost function	[-]
δ_x	Class label indicator for perceptron cost	[-]
η	Learning rate	[-]
l_i	Low-Resolution image of a scene	[-]
Y	High-Resolution image	[-]
γ	Up-sampling factor	[-]
$\hat{Y}$	Predicted Super-Resolution image	[-]
N	Number of Low-Resolution images	[-]
...		

Symbol	Definition	Unit
θ	Parameters of the encoder	[-]
α	Parameters of the fusion stage	[-]
β	Parameters of the decoder stage	[-]
r_i	Latent representation of the input frame l_i	[-]
h_t	Hidden state within the ConvGRU	[-]
u_t	Update gate	[-]
r_t	Reset gate	[-]
c_t	Candidate hidden state	[-]
W	Weights of the convolution kernels	[-]
b	Biases of the convolution kernels	[-]
f_α	Fusion function in ConvGRU	[-]
h^{GRU}	Hidden representations in ConvGRU	[-]
h_{avg}	Averaged hidden representations	[-]
GAP	Global Average Pooling	[-]
$L_\mathcal{F}$	Fourier domain loss function	[-]
$L_{\mathcal{F}_\mathcal{A}}$	Absolute amplitude difference in Fourier domain	[-]
$L_{\mathcal{F}_\angle}$	Absolute phase difference in Fourier domain	rad/s
$\mathcal{F}x$	Fourier transform of image x	[-]
$\lvert X_{u,v}\rvert$	Amplitude in Fourier space	[-]
$\angle X_{u,v}$	Phase in Fourier space	rad/s
x	Image in spatial domain	[-]
X	Image in Fourier domain	[-]

1

Introduction

Live-cell 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. To visualize the structures, fluorescent molecules, called fluorophores, are first used to label the specimen. They bind to target proteins or structures and emit light upon excitation. A sequence of frames is then captured, with molecules randomly blinking across the frames. Fluorophores can blink sparsely for precise localization or densely. However, visualizing these cellular dynamics is constrained by the diffraction limit of optical lenses, which restricts spatial resolution to around 200-300 nm (half the wavelength of visible light).

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], which can achieve spatial resolution down to 10 nm. However, these methods either require complex microscope setups, higher signal-to-noise ratio (SNR) conditions (higher illumination), or even millions of frames to reconstruct an SR image, which can limit temporal resolution and can damage the sample[6, 7]. Other techniques, like Super-resolution Optical Fluctuation Imaging (SOFI)[8] and Super-Resolution Radial Fluctuations (SRRF) microscopy[9] do not require a complex microscope setups and typically use hundreds of frames to reconstruct an SR image as they can handle very dense blinking fluorophores.

Only SOFI generates SR images based on statistical theory by leveraging the correlation of blinking from a fluorophore over time and space (cross-cumulant). The resolution improvement arises from the independence of different fluorophores [10, 11]. Because noise does not correlate over time, SOFI can operate under low SNR conditions, making it suitable for live-cell imaging. For n-order SOFI, an n-fold improvement in spatial resolution is achieved; however, it still requires hundreds of frames, limiting its ability to capture dynamic cellular processes and making real-time imaging impractical. Without real-time capabilities, researchers cannot make immediate decisions, wasting valuable time. Deep learning could help reduce the number of frames needed for SR image reconstruction while maintaining spatial resolution improvements.

We propose an end-to-end deep learning model that accelerates cross-cumulant SOFI by reconstructing a second-order SR image from just 20 frames, compared to the hundreds typically required by SOFI, while still achieving a 2-fold spatial resolution improvement. Unlike U-Net methods that use only spatial information, our model leverages spatiotemporal data to extract correlated blinking fluorophores. It operates in three stages: encoding input frames into feature maps, fusing them with a recurrent structure to capture blinking correlations, and upsampling the result into a second-order SR image. The model is trained in a supervised manner, using the pre-trained weights from synthetic data, we can train the model on real fixed-cell microscopy data using at least four different measurements of the same cell type. By applying random cropping and rotations, the dataset expands to around 2000 samples. This approach enables the model to be used in live-cell experiments, making it a practical method for real-world applications.

1.1. Outline

In the following chapters, the theory behind fluorescent microscopy and deep learning is presented in chapter 2, with an in-depth explanation of SOFI in chapter 3. Chapter 4 presents our proposed architecture, followed by chapter 5, which explains how the datasets used in this research were created. Finally, the main findings of this research are provided in a manuscript in chapter 6.

2

Background

2.1. Fluorescence Microscopy

In the early 20th century, the fluorescence microscope was invented to better understand underlying intricacy in cellular biology. It made it possible capture the spatial and temporal details from both intrinsically fluorescent objects and those labeled with extrinsic fluorescent molecules. These observations extend to entities that are so minuscule they elude detection by the unaided eye[12].

One of the most commonly used fluorescence microscopes is the widefield microscope (see figure 2.1), which illuminates the entire field of view simultaneously. Widefield microscopy[12] is not only a simple and fast imaging modality but is also crucial for single-molecule detection. The ability to see individual molecules relies on a fundamental property of fluorescence: the emission of light at a longer wavelength (lower energy) than the excitation light. Dichroic filters play a key role by reflecting the excitation light and transmitting the emitted fluorescence. This separation allows the emitted fluorescence from single molecules to be captured on the camera, enabling detailed imaging at the molecular level.

We further discuss the theoretical foundations of this research, discussing the diffraction limit, point-spread-function (PSF), super-resolution techniques, and live cell imaging, aiming to provide a clear and detailed explanation.

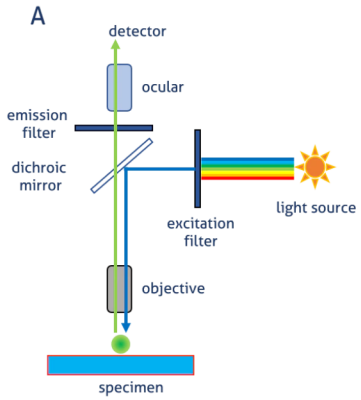


Figure 2.1: Widefield microscope.

2.1.1. Diffraction Limit

The spatial resolution of the fluorescence microscope is intrinsically constrained by its optics, as fundamental physical laws govern the upper limit achievable in fluorescence microscopy. These inherent barriers, dictated by diffraction limitations, pose challenges that cannot be surpassed through physical

means. Consequently, the optical instrument faces difficulties in distinguishing closely positioned objects.

Lord Rayleigh, established a standard formula to describe the spatial resolution of an optical device. According to this theory, the resolution limit is defined by the minimum distance between two distinguishable point sources. These two sources are considered just resolved when the highest point of one diffraction pattern aligns with the first minimum of the the other, determining the achievable resolution [12], see figure 2.2. The standard formula is given by:

$$d_{x,y} = 0.61 \frac{\lambda}{NA} \tag{2.1}$$

where λ is the wavelength of light and NA the numerical aperture of the objective which is given by $NA = n \sin \theta$ and is dependent on the refractive index n of the objective immersion medium and the half-angle θ of the cone of light collected by the lens.

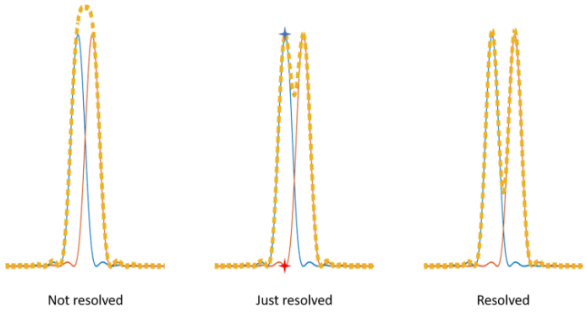


Figure 2.2: The Rayleigh criterion: Two points are deemed to be resolved when the peak of one diffraction pattern aligns with the initial minimum point of the other.

2.1.2. Point Spread Function

The microscope's ability to capture minuscule objects, like a single fluorescent protein, is explained by the Point Spread Function (PSF). In an ideal scenario, this function forms an Airy disk pattern in the focal plane, see figure 2.3. The size and shape of the Airy disk depend on factors like the numerical aperture of the lens and the wavelength of light. The PSF, fundamentally, describes how a point source of light is spread out or blurred in an image. It represents the response of an optical system to a single point source:

$$Y = H \otimes X \tag{2.2}$$

In the given equation, H represents the PSF of the system, X denotes the true image, and Y represents the blurred image resulting from the convolution with the PSF.

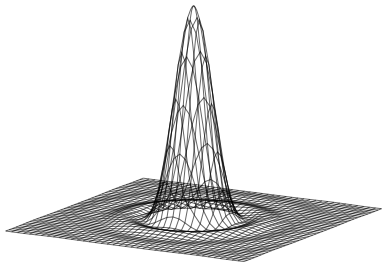


Figure 2.3: Airy disk, the most ideal scenario as a PSF.

2.1.3. The Inherent Deconvolution Problem

To see the smaller details, one needs to obtain an image with a spatial resolution that goes beyond the diffraction limit. To obtain such images, the images need to be deconvolved with the estimated PSF of the optical system. While this may sound straightforward, in reality, it is more complicated as it involves an ill-posed inverse problem. To enhance our understanding, in this subsection, we model the acquisition process of the fluorescence microscopy to formulate the complexity of the deconvolution[11].

Let us assume a continuous framework to represent the fluorescence imaging model. Where in the sample there exist K independent blinking/fluctuating emitters, where $K > 1$. Each emitter is positioned at $r_k \in \mathbb{R}^N$, where N represents the spatial dimensions. The emitter exhibits intensity over time, denoted as $t = 1, ..., T$, and this temporal information is collected in the vector $s_k \in \mathbb{R}^T$, where ϵ_k is the constant molecular brightness. The true image intensity at a given position r and time t is expressed as follows:

$$x(r,t) = \sum_{k=1}^{K} \delta(r - r_k) \cdot \epsilon_k \cdot s_k(t) \tag{2.3}$$

where $\delta(\cdot)$ is the Dirac delta.
Assuming that the microscope has a spatial-and time-invariant PSF denoted as $v(r)$, the diffraction-limited image $y(r,t)$ at position r and time t is obtained through the convolution of the system's PSF with $x(r,t)$:

$$y(r,t) = x(r,t) * v(r) = \sum_{k=1}^{K} v(r - r_k) \cdot \epsilon_k \cdot s_k(t) \tag{2.4}$$

To discretize the model from equation 2.4, the location of the emitters must be localized with great accuracy. Therefore, a finer grid is used in the discrete model. Additionally, a Gaussian noise component is added to the equation to represent the inherent noise of the system. The model for all $t = 1, ..., T$ with $T > 0$, is expressed as follows:

$$\mathsf{y}_t = M_q \left(U \left(\mathsf{x}_t \right) \right) + \mathsf{n}_\mathsf{t} \tag{2.5}$$

This model establishes a connection between the acquired image $\mathsf{y}_t \in \mathbb{R}^M$ and the true image $\mathsf{x}_t \in \mathbb{R}^L$, which cannot be directly observed. Instead, the true image x_t is accessed only through the model and lies on a grid L that is q^N-times finer than the acquisition space M. This relationship holds when $L = q^N M$. Additionally, the model consist of a convolution operator $U : \mathbb{R}^L \mapsto \mathbb{R}^L$, which represents the PSF of the system, and a downsampling operator $M_q : \mathbb{R}^L \mapsto \mathbb{R}^M$ which returns the sum of every q^N non-overlapping sequential pixel blocks. Finally, the noise component n_t describes the inherent noise of signal-independent measure noise with the additional model errors. This is represented as vector of identical distributed (i.i.d.) Gaussian random variables with zero mean and a constant variance.

Going one step further, in the context of a real microscopy setting, we introduce a background represented by the vector $\mathsf{b} \in \mathbb{R}^M$ to the model. This background represents the contributions from out-of-focus (and ambient) fluorescent molecules and coexists with signal-dependent photon noise. Thus, a more accurate model is given by:

$$\mathsf{y}_t = \boldsymbol{P} \left(M_q \left(U \left(\mathsf{x}_t \right) \right) + \mathsf{b}_\mathsf{t} \right) + \mathsf{n}_\mathsf{t} \tag{2.6}$$

where, $\mathsf{w} \in \mathbb{R}^M$, $\boldsymbol{P}(\mathsf{w})$ represents the realization of a multivariate Poisson variable with parameter w.

2.1.4. Super-Resolution Techniques

The spatial resolution of fluorescence microscopes is inherently limited by diffraction. Overcoming this barrier is crucial since many biological structures are smaller than the diffraction limit (see figure 2.4). Super-resolution (SR) techniques have been developed to surpass this limit and achieve higher spatial resolutions, each with its own strengths and weaknesses. In this thesis, we classify these techniques into two categories:

1. *Illumination- and Single-Molecule-based Methods:* require specific microscope configurations.
2. *Fluctuation-based Methods:* work with standard microscope setups.

Specific configurations refer to setups needing complex hardware or specialized fluorophores (single-molecule techniques typically use a widefield microscope). In contrast, standard setups involve simple hardware and standard fluorophores. In this subsection, we will explore these SR methods in more detail.

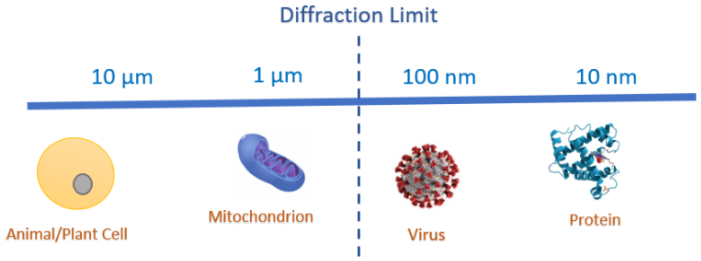


Figure 2.4: Beyond the diffraction limit, numerous important biological entities exist, representing the intricacies of life.

Illumination- and Single-Molecule-based Super-Resolution Methods

The first SR technique, Stimulated Emission Depletion (STED) microscopy[2], was introduced in the late 20th century, overcoming the diffraction limit by using a depletion light beam to shrink the PSF. This minimizes the illumination area, improving the spatial resolution to less than 50 nm laterally. However, STED techniques have drawbacks, including a slow acquisition process and the need for an expensive setup and special fluorophores.

Following STED, Structured Illumination Microscopy (SIM) emerged[3], utilizing patterned illumination for high temporal resolution with fast acquisitions. Despite sacrificing spatial resolution and requiring a specific illumination setup, SIM proved valuable. Random Illumination Microscopy (RIM)[13] is proposed as a robust alternative to SIM but provides a similar SR gain[14].

Single Molecule Localization Microscopy (SMLM) techniques, including Photo-Activated Localization Microscopy (PALM) [4] and Stochastic Optical Reconstruction Microscopy (STORM) [5], provide nanometric resolution, achieving spatial resolution up to 10 nm using a widefield microscope. These methods generate SR images by sequentially activating and accurately localizing individual molecules over typically millions of frames. While advanced software enables precise detection and localization, these techniques require specialized fluorophores, higher laser power, and rely on capturing a vast number of frames. This leads to limitations in temporal resolution and increases the risk of sample damage [15]

Fluctuation-based Super-Resolution Methods

Fluctuation-based super-resolution methods utilize the independent fluctuations of fluorophores (see figure 2.6) to generate a super-resolution (SR) image, typically using hundreds of frames. These methods can also operate in lower illumination environments with reduced laser power, thereby minimizing the risk of sample damage.

One notable technique is Super-resolution Optical Fluctuation Imaging (SOFI) [8]. This method reduces the size of the point spread function (PSF), thereby improving spatial resolution almost proportionally to the order of cumulants used, with an improvement factor of $\sqrt{n}$ [16]. Additionally, through deconvolution or Fourier reweighting, resolution can be improved up to n-fold. However, using higher-order

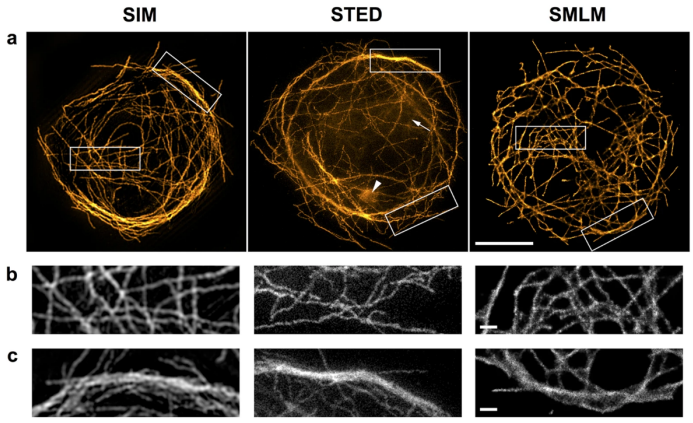


Figure 2.5: (a) Highlighting variations in the outcomes of SIM, STED, and SMLM microscopy through illustrated boxes delineating the areas of interest, accompanied by a 5μm scale bar. (b) Close-up view of a specific sample region, indicated by a scale bar of 0.5μm. (c) Additional region of interest, presented with the same scale bar.

cumulants in SOFI is limited because they can amplify variations in molecular brightness and blinking behavior, which can adversely affect resolution. The theoretical model underlying SOFI is based on the statistical properties of a fluorophore that blinks both temporally and spatially in a correlated manner, which we discuss in detail in Section 3. An extension of SOFI, known as balanced SOFI (bSOFI) [17], combines multiple cumulant orders to achieve better resolution compared to SOFI. While bSOFI has demonstrated a 4.6-fold resolution improvement (64 nm), it still falls short of the levels achieved by SMLM methods such as PALM and STORM.

Another method, Super-Resolution Radial Fluctuations (SRRF) microscopy[9], achieves SR by assessing local symmetry in each frame of a temporal stack of diffraction-limited frames. Using the symmetry of the microscope's PSF, SRRF calculates the degree of local gradient convergence, termed 'radiality', on a sub-pixel basis across the entire frame. Close to a fluorescent molecule, this results in high radiality, while a displaced sub-pixel exhibits low convergence due to the absence of a nearby molecule. Generating a single SR image involves temporal analysis through estimating a time average or higher-order statistical analysis, similar to SOFI. It can achieve a 5-fold spatial resolution improvement (60 nm) with a temporal rate of 1 second. An extension of SRRF is enhanced SRRF (eSRRF)[18], which uses automated data-driven parameter optimization, including indicating how many frames one needs for optimal reconstruction. This enables an ease of use over a wide range of microscope techniques. However, SRRF often introduces reconstruction artifacts and is not based on any theoretical statistics.[11].

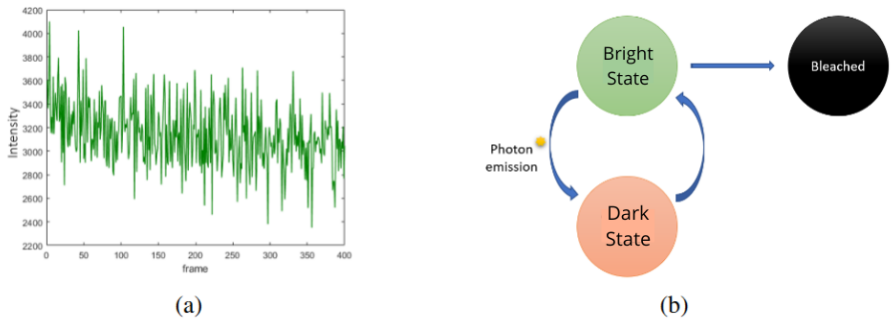


Figure 2.6: (a) The temporal characteristics of a pixel expressed in fluorescence intensity, acquired at a frequency of 40 frames per second (fps). (b) The various states of a fluorescent molecule.

2.1.5. Live-cell Imaging

Live-cell imaging within the realm of SR techniques aims to capture detailed spatio-temporal information while minimizing sample damage. This imaging can be conducted through post-processing methods, such as SOFI or SRRF, or in real-time [1]. SOFI requires simple hardware, such as a widefield microscope, and has the potential to achieve more than a 2-fold resolution improvement when using higher cumulant orders. Additionally, users can switch between SOFI and single molecule localization microscopy (SMLM), allowing for enhanced resolution in both live and fixed samples. Real-time imaging significantly shortens the feedback loop, facilitating quicker adjustments and seamless transitions between modalities. Figure 2.7 illustrates the balance between sample health, temporal resolution, and spatial resolution in live-cell imaging.

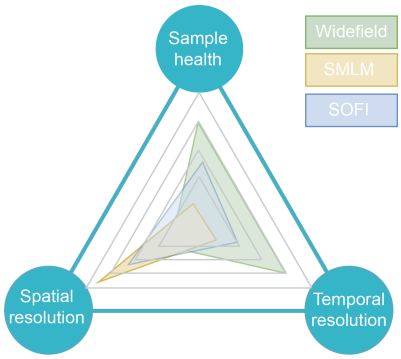


Figure 2.7: Live-cell imaging in fluorescent microscopy poses distinct challenges that demand a delicate balance between sample health, temporal resolution, and spatial resolution to accurately measure the biological sample.

2.2. Deep Learning

In 1957, the perceptron was developed by Frank Rosenblatt, and it became the fundamental building block in the field of machine learning and artificial neural networks today. It is a discriminative model, meaning that it directly models the decision boundary between the different classes in the input space[19]. As an example, assume we have two linear separable classes w_1 and w_2 such that there exist a decision boundary $f(x) = 0$ that separates these classes, as given by:

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + w_0 > 0 \quad & \text{if } f(x) = +1 \\ \mathbf{w}^T\mathbf{x} + w_0 < 0 \quad & \text{if } f(x) = -1 \end{aligned} \tag{2.7}$$

The perceptron itself is given by the equation:

$$f(x) = \sigma(\mathbf{w}^T\mathbf{x} + w_0) \tag{2.8}$$

Here, σ is the sigmoid activation function, mapping real numbers between 0 and 1. This non-linear activation function plays a crucial role in the perception as it enables the algorithm to learn complex non-convex relations in the data. To learn these patterns, the weights w need to be updated according to the data, and therefore, a cost function is established, namely the perceptron cost, defined as:

$$J(\mathbf{w}) = \sum_{x \in Y} \delta_x(\mathbf{w}^T\mathbf{x} + w_0) \tag{2.9}$$

where Y is the subset of training vectors that are misclassified by the perceptron defined by the weights. The variable δ_x is chosen so that $\delta_x = -1$ if $\mathbf{x} \in w_1$ and $\delta_x = +1$ if $\mathbf{x} \in w_2$. The goal is to minimize the cost function $J(\mathbf{w})$ to achieve a good fit with our training data; in machine learning terms, this is equivalent to achieving generalization. To update our weights, gradient descent is used, see equation 2.10, where η is the learning parameter that determines the step size in weight updates. This procedure is done iteratively, where the gradient goes down the slope until convergence is achieved, leading to the effective separation of the two classes.

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \tag{2.10}$$

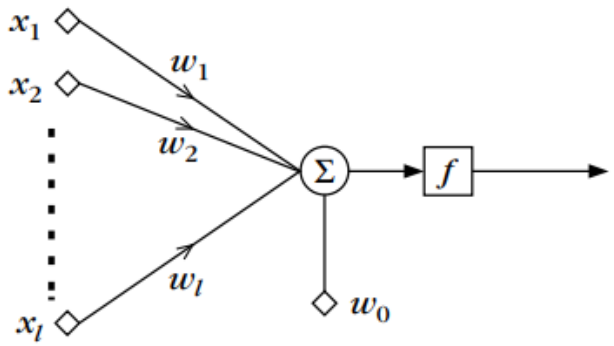


Figure 2.8: The basic perceptron model: a linear combination of input features and their respected weights, summed up with the bias, followed by the activation function sigma.

However, a single perceptron is limited in its learning capabilities. For example, it cannot learn the XOR problem, but if multiple connected perceptrons are used, a so-called multi-layer perceptron, it is possible to learn the XOR problem[20]. Here, we briefly introduced the fundamental building block of neural networks, namely the perceptron. In this section, we will further discuss the artificial neural networks (ANN), convolution neural networks (CNN), and recurrent neural networks (RNN), which lay the foundation of deep learning in the context of this thesis.

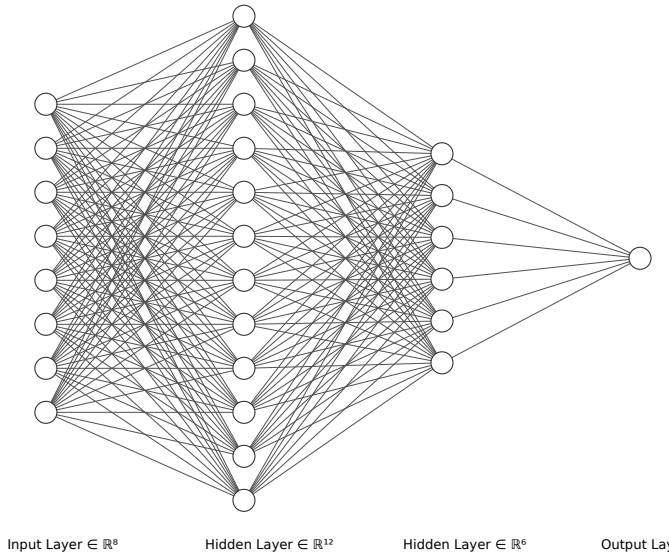


Figure 2.9: Example of a feedforward neural network with 2 hidden layers and its corresponding input and output layers.

2.2.1. Artificial Neural Networks

Solving complex non-convex problems using just a single perceptron is not feasible. However, with fully connected perceptrons, it becomes possible. These networks are known as multilayer perceptron, or, in modern-day terminology, feedforward neural networks[20]. They are universal learning algorithms where the goal is to learn $f(x)$ from $f^*(x)$. In practice, the algorithm learns an approximation, $\hat{f}(x)$, of the true function $f^*(x)$. By introducing deeper connections in the network, known as hidden layers, the algorithm can achieve a better approximation of $f^*(x)$. Yet, this improvement comes with the risk of overfitting, especially when there is insufficient training data. To train the network, the training data is parsed through the network in process called the forward step, where the loss of the network is calculated. Based on this loss function, the weights and biases are updated to minimize it, a procedure known as back-propagation. Back-propagation involves applying the chain rule from calculus to compute the gradients in all the neurons. For example, assume that $x \in \mathbb{R}^M$, $y \in \mathbb{R}^N$, g maps from $\mathbb{R}^M$ to $\mathbb{R}^N$, and f maps from $\mathbb{R}^N$ to $\mathbb{R}$. Given that $y = g(x)$ and $z = f(y)$, the procedure becomes

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_i} \tag{2.11}$$

It is assumed here that the whole training set can fit in the memory of the system, but more often than not, this is not the case. Therefore, minibatches are used, which are drawn i.i.d. from the dataset. Then, again, the forward pass and backpropagation are applied to this minibatch and the weights and biases are updated according to the batchgradient. This procedure is called stochastic gradient descent (SGD). However, due to the i.i.d. drawn mini-batches, the learning process is not very stable; namely, it has a more stochastic character. SGD has a fixed learning rate, which might lead to suboptimal convergence or slow learning in certain parts of the parameter space. Therefore, other techniques exist, such as RMSProp or Adam, which attempt to overcome this issue by dynamically changing the learning rate and applying momentum to the gradient, thus enabling a faster and more stable learning process. However, it does not guarantee you have better generalization performances[21].

2.2.2. Convolutional Neural Networks

Convolutional neural networks (CNNs) are a special kind of neural networks used to process images, making use of the mathematical operation convolution, which blends two functions together. In CNNs, this involves convolving the first argument, referred to as the input, with a kernel where the weights are learned, and the output of this convolution is referred to as the feature map[20]. CNNs typically exhibit sparse connectivity because the kernel is smaller than the input. This implies that for processing images, it requires less memory and fewer computational operations compared to a feedforward network, where every pixel is considered a dimension. The procedure of convolution involves sliding a

kernel over the input tensor, performing element-wise multiplication with kernel weights and summing the results to procure the output feature map, figure 2.10 illustrates this principle.

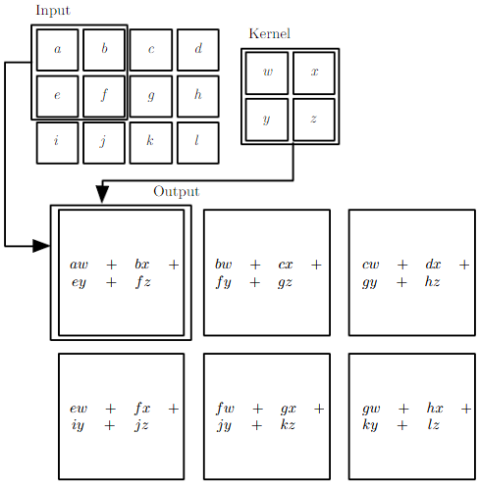


Figure 2.10: An example of a 2D convolution showcasing the transformation of the input tensor through the kernel.

A CNN usually consists of multiple layers, where each layer typically having three stages. The fist stage is the convolution stage. The second stage is the activation stage, where the feature maps are run trough a nonlinear activation function, such as the rectified linear activation function. The third stage is the pooling layer, which serves as a form of downsampling. For example, in max pooling, it partitions the input into a set of non-overlapping rectangles and, for each such sub-region, outputs a maximum value. Thereby, reducing the spatial dimensions, hence downsampling. This helps in reducing computational complexity and also provides a degree of translation invariance, meaning small translations of the input have no effect on most of the pooled outputs.

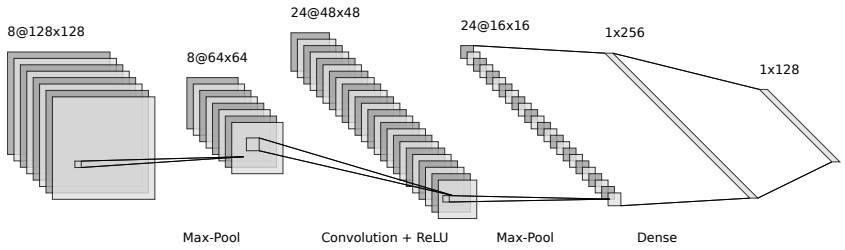


Figure 2.11: Here's an example of a CNN: the first layer is the max pooling layer, which reduces the spatial dimensions by half. The second layer is the convolution layer, where 8 convolution filters with a kernel size of 16x16 are used with an additional non-linearity function of type ReLU. This is followed by another max pooling layer to decrease the spatial dimensions again by half. The output can then be used in a feedforward network for tasks such as image recognition.

2.2.3. Recurrent Neural Networks

For some machine learning problems, there is a need to process sequential information to understand the underlying patterns. For example, this can involve sequences of words, where the model attempts to predict the next word, or weather predictions, where the model processes a sequence of weather data to forecast whether it will rain in the next hour. A recurrent neural network (RNN) is a type of neural network specialized in processing sequences of values $x^1, ..., x^\tau$ [20]. In RNNs, there are primarily three design patterns: one-to-many, many-to-one, and many-to-many, as depicted in Figure 2.12. These networks are challenging to train not only because they need to process sequences sequentially but also due to the issues of exploding or vanishing gradients during backpropagation. These problems can make the training process unstable and may result in the network being unable to effectively learn long-term dependencies. To mitigate these issues, different architectures have been proposed, such as LSTM and GRU, although they are not completely immune to these problems. They are similar to each other, but the GRU uses fewer parameters and only two gates: the update gate u_t and the reset

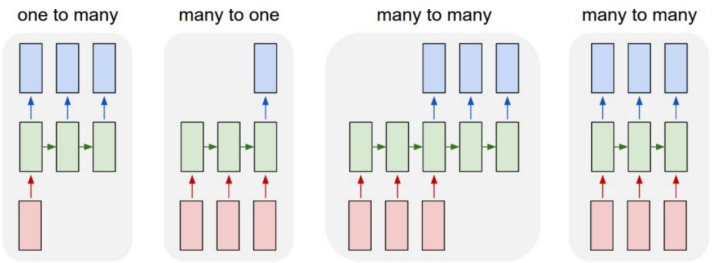


Figure 2.12: The rectangles represent the vectors, and arrows depict a linear transformation, i.e., matrix multiplication. The corresponding colors for the vectors are as follows: the input vector is red, the output vectors are blue, and the green vectors represent the RNN's state.

gate r_t. The gate u_t determines the update speed of the hidden state, while the gate r_t decides how much information to forget by resetting parts of the memory. In contrast, the LSTM includes three gates: the forget gate f_t, the input gate i_t, and the output gate o_t. The gate f_t determines how much of the previous information to forget, the gate i_t determines how much information to write into memory, and the gate o_t decides the output based on the current information [22].

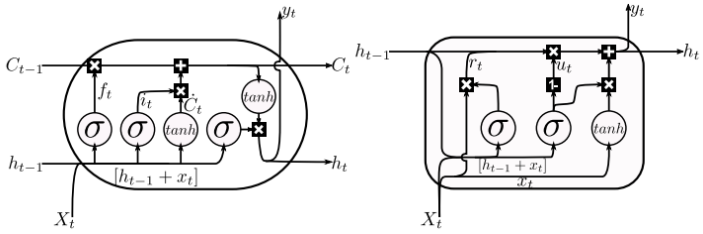


Figure 2.13: Architecture of the LSTM (left) and GRU (right).

2.2.4. Autoencoders

The goal of an autoencoder is to learn how to copy the message from the input to the output. It consists of two main parts: an encoder that maps the message from the input to the hidden layer, h, which represents a *code*, which is a representation of the input, and a decoder that maps the *code* to the output, attempting to reconstruct the original message [20]. In formal terms, the encoder function $b = f(x)$ and a decoder that produces $r = g(h)$. The architecture is visualized in figure 2.15.

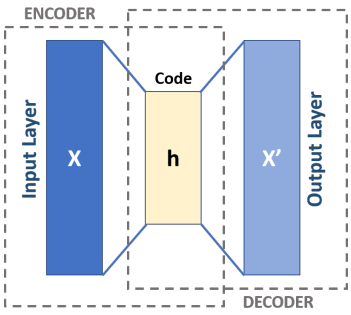


Figure 2.14: Architecture of an autoencoder, representing the two main parts of encoder and a decoder.

The goal of an autoencoder is not simply to learn $g(f(x)) = x$ everywhere, but rather to approximate it. This is achieved by reducing the dimensionality of the hidden layer h relative to the input and output layers, a process referred to as the bottleneck. This constraint forces the network to learn only the most useful representations of the data, discarding noise or irrelevant features. Autoencoders come in various types, such as denoising autoencoders, which are designed to reconstruct clean input from noisy ones, and variational autoencoders, which introduce a probabilistic framework. These architectures are widely used in applications like dimensionality reduction, anomaly detection, and image denoising.

2.2.5. U-Net

The U-Net architecture [23] was developed to meet the demand for efficient and accurate image segmentation in biomedical applications, especially when dealing with small datasets, such as those found in cell tracking and tissue analysis. This architecture is composed of a contracting path on the left and an expansive path on the right. The contracting path consists of a series of convolutional operations that gradually reduce the size of the feature space. It employs the repeated use of two unpadded 3×3 convolutions, each followed by a rectified linear unit (ReLU), and includes a 2×2 max pooling operation with a stride of 2 for downsampling. At each step of downsampling, the number of feature channels is doubled, and skip connections pass features from the contracting path to the corresponding layer in the expansive path, preserving crucial spatial information.

Conversely, the expansive path upsamples the feature maps at each step using a 2×2 transposed convolution (upsampling) that reduces the number of feature channels by half, followed by a concatenation with the corresponding cropped feature map from the contracting path. This process is followed by two 3×3 convolutions, each accompanied by a ReLU activation. Cropping is necessary to account for the loss of border pixels that occurs during the convolution operations. In the final layer, a 1×1 convolution is applied to map each 64-dimensional feature vector to the specified number of classes. In total, the network comprises 23 convolutional layers.

Originally, the U-Net architecture was designed for biomedical tasks. Due to its simplicity and flexibility, it has been extensively used in various domains, including super-resolution microscopy [24, 25].

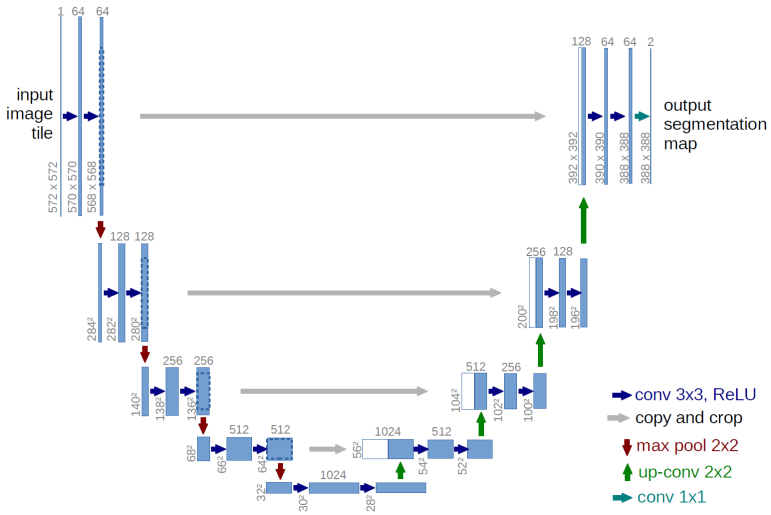


Figure 2.15: The U-Net architecture, illustrated for a 32×32 pixel resolution at the the lowest level. Each blue box in the diagram represents a feature map with multiple channels, which is given above the each box. The dimensions (width and height) of the feature maps are displayed at the bottom left corner of the box. The white boxes indicate feature maps that have been copied from the previous layers. The arrows in the diagram illustrate the various operations, such as convolutions, pooling, and upsampling.

3

Understanding Super-resolution Optical Fluctuation Imaging (SOFI)

SOFI (Stochastic Optical Fluctuation Imaging) is a super-resolution technique that leverages the statistical relationship between blinking emitters over space and time. This technique effectively reduces the convolved PSF in the image by a factor of $\sqrt{n}$ for a given order of SOFI, thereby enhancing the spatial resolution of the image[26].

One might wonder why we do not simply deconvolve the PSF directly from the image. While this approach might seem straightforward, it is actually quite complex due to the ill-posed nature of the inverse problem, as discussed in Chapter 2.

Other techniques, such as Lucy–Richardson deconvolution[27], use iterative procedures to recover the underlying structure. However, determining the optimal number of iterations needed is challenging, and there is no guarantee that the resulting image accurately represents the true structure.

This chapter delves into the theoretical framework of SOFI, exploring the functioning of auto-cumulant, their extension to cross-cumulants, and the processes of flattening and linearizing the cumulants.

3.1. Auto-cumulant

Let us assume a continuous two-dimensional framework with a sample that consists of N independent emitters whose brightness is time-dependent (figure 3.1(b)). The resulting time-dependent fluorescence image at position r, which is convolved with the system's PSF, is given by equation 2.4. It is important to note that it is assumed here that the positions of the emitters remain static throughout the entire measurement, and any temporal changes are only caused by blinking (figure 3.1(c)).

Given that a measurement is taken by a widefield microscope, it involves a time average over the $y(r,t)$ image. This image is the sum of averaged contributions from each emitter, convolved with the PSF. On the other hand, SOFI derives the image from correlations of time-dependent signal (figure 3.1(d)). The second order of SOFI performs auto-correlation on the fluorescence intensity and is given by:

$$G_2(r,\tau) = \langle \delta y(r,t+\tau) \cdot \delta y(r,t) \rangle = \sum_{k=1}^{N} U^2(r-r_k) \cdot \epsilon_k^2 \cdot \langle \delta s_k(t) \cdot \delta s_k(t+\tau) \rangle \tag{3.1}$$

Where r_k is the emitter location, ϵ_k is the constant molecular brightness, $s_k(t)$ is the time-dependent component with values between 0 and 1. Furthermore, $\langle \cdot \rangle$ denotes time averaging, and $\delta y(t) = y(t) - \langle y \rangle$ describes the fluctuations, i.e., the difference with respect to the average intensity at a given time. The assumption here is that the emitters fluctuate independently from each other, meaning that the emissions of the emitters are uncorrelated in time. As a result, cross terms proportional to the product

of two independent emitters intensities $\langle \delta s_i(t) \cdot \delta s_j(t+\tau) \rangle$, for $i \neq j$, average to zero, and equation 3.1 contains only a single sum over emitters.

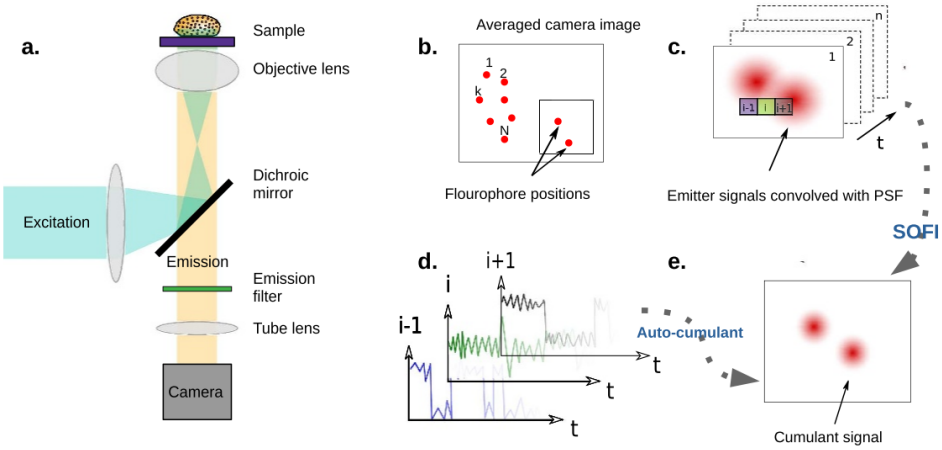


Figure 3.1: (a) Sample measurement with a widefield fluorescence microscope. (b-e) Principle of SOFI. (b) Emitter distribution in the object plane, which blink independently over time from each other. (c) A time laps of images, where each emitter in the image is convolved with the PSF of the system. (d) Each pixel has an intensity time trace of the sum of individual emitter signal whose PSFs reach that pixel. Auto correlation is then performed (second-order) from the fluctuations at each pixel. (e) At last, the SOFI intensity value is assigned for each pixel, which is given by the integral over the second-order correlation function. The resulting SOFI image has resolution improvement of a $\sqrt{2}$.

In equation 3.1, it can already be observed that a squared PSF results in a $\sqrt{2}$ improvement in spatial-resolution making the appearance of sharper features in the image (figure 3.1(e)). Hence, we have super-resolution.

In more generalize terms, for the n-th order SOFI is given by:

$$k_n(r, \tau_1, ..., \tau_{n-1}) = U^n(r - r_k) \cdot \epsilon_k^n \cdot w_{n,k}(\tau_1, ...\tau_{n-1}) \tag{3.2}$$

where $w_{n,k}(\tau_1, ...\tau_{n-1})$ is the single-emitter cumulant of the n-th order. Ultimately, achieving a spatial resolution improvement of a $\sqrt{n}$.

3.2. Cross-cumulant

We can also extend to cross-cumulant-based resolution enhancement obtained from spatio-temporal correlations for SOFI. Consider a second-order cross-cumulant between the positions r_1 and r_2 for a Gaussian-shaped PSF[26]:

$$k_2(r_1, r_2, \tau) = \sum_{k=1}^{N} U(r_1 - r_k) \cdot U(r_2 - r_k) \cdot w_k(\tau) = U\left(\frac{r_1 - r_2}{\sqrt{2}}\right) \sum_{k=1}^{N} U^2\left(\frac{r_1 + r_2}{2} - r_k\right) \cdot w_{2,k}(\tau) \tag{3.3}$$

where $w_{2,k}(\tau)$ is the single-emitter cumulant of the second-order.

What is interesting about the expression above is that it shows a cross-cumulant between r_1 and r_2, resulting in a signal at the geometric center of the points $\left(\frac{r_1 - r_2}{2}\right)$. Secondly, the expression is weighted by a distance factor $U\left(\frac{r_1 - r_2}{\sqrt{2}}\right)$. This is because emitters fluctuate independently, and only points within the same emission PSF can contribute to the same emitter, resulting in non-zero correlation. The significance of this geometric centered signal becomes apparent when considering a camera sensor with a set of a finite-sized pixels. Performing cross-cumulant on neighboring pixels creates additional "virtual pixels", thereby increasing the pixel density[10], as can be seen in figure 3.4. Unlike simple interpolation, these pixels carry additional information. However, if auto-cumulants are performed in increasing orders, the resolution improvements will eventually be limited by the pixel size. This limitation is why cross-cumulants are preferred for reconstructing the super-resolved SOFI image, avoiding the

constraints imposed by auto-cumulants.

In general, calculating cross-cumulants involves a computationally intensive formula, requiring sums over partitions. This complexity arises from combining various pixel to derive a new value, rendering fast recursive methods impractical. The equation utilized for cross-cumulant calculation is given by[10]:

$$K_n = \left(\overrightarrow{r} = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{r_i} \right) = \sum_{P} (-1)^{|P|-1} (|P|-1)! \prod_{p \in P} \left\langle \prod_{i \in p} F(\overrightarrow{r})_i \right\rangle_t \tag{3.4}$$

Where P represents the total number of possible partitions, while p signifies the individual parts within each partition. Furthermore, i serves as the index for various pixel positions considered during the computation, where F represents the stack of images comprising the contributing pixels. The cross-cumulant methods enables the creation of virtual pixels based on the cumulant order. These virtual pixels can be generated in a specific pattern from the original pixels, either be combinations with repetitions or without repetitions[28]. In figure 3.2, the key differences between using repetitions and not using them are discussed. Additionally, figure 3.5 outlines the necessary steps to calculate the n-th order SOFI cross-cumulant, providing insight into how equation 3.4 is performed step by step.

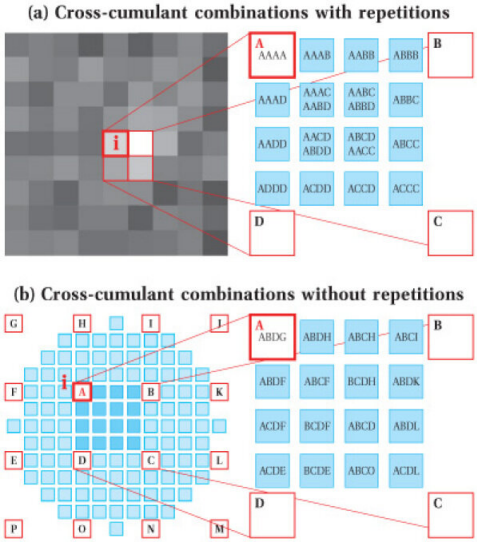


Figure 3.2: Fourth-order cross-cumulant combinations for a pixel i with or without repetitions. Various combinations within the pixel's neighborhood matrix can generate 15 inter-pixels between the original pixel matrix (ABCD). Combinations leading to the same inter-pixel are averaged. (a) For a 2×2 neighborhood, starting with A and allowing repetitions, all combinations are computed and easily expandable to any order (n). However, shot noise suppression is limited without non-zero time lags due to autocumulants. (b) In a 4×4 neighborhood of pixel i, combinations can generate inter-pixels in a circular arrangement, leading to better shot noise suppression by excluding repetitions (autocumulants). Only combinations with the shortest sum of distances to their corresponding inter-pixels are considered for computational efficiency.

3.3. Flattening and Linearization

Up until now, the single-emitter cumulant of the n-th order, $w_{n,k}$, has not been discussed in greater detail. To understand it better, consider a sample composed of M independently emitter fluctuations and assuming a simple two-state blinking model with τ_{on} and τ_{off}. The cumulant order n without time-lags can be interpreted as follows[17]:

$$\begin{aligned} k_n(\vec{r}) &\propto \sum_{k=1}^{M} \epsilon_k^n U^n(\vec{r} - \vec{r_k}) f_n(\rho_{on},\, k) \\ &\approx \epsilon^n f_n(\rho_{on}) \sum_{k=1}^{M} U^n(\vec{r} - \vec{r_k}) \end{aligned} \tag{3.5}$$

where ϵ^n represents spatial distribution of the molecular brightness and $\rho = \frac{\tau_{on}}{\tau_{on}+\tau_{off}}$ is the on-time ratio. $U(\vec{r})$ is the system's PSF and $f_n(\rho_{on})$ is the n-th order cumulant of a Bernoulli distribution with probability ρ_{on}. Note that in equation 3.3, $w_{n,k}$ has become $f_n(\rho_{on})$. The n-th order cumulant of a Bernoulli distribution can be written as:

$$\begin{aligned} f_1(\rho_{on}) &= \rho_{on} \\ f_2(\rho_{on}) &= \rho_{on}(1-\rho_{on}) \\ &\vdots \\ f_n(\rho_{on}) &= \rho_{on}(1-\rho_{on})\frac{\partial f_{n-1}}{\partial \rho_{on}} \end{aligned} \tag{3.6}$$

It can be observed that for higher-order cumulants, there is a non-linear response to the molecular brightness levels, as shown in figure 3.3.

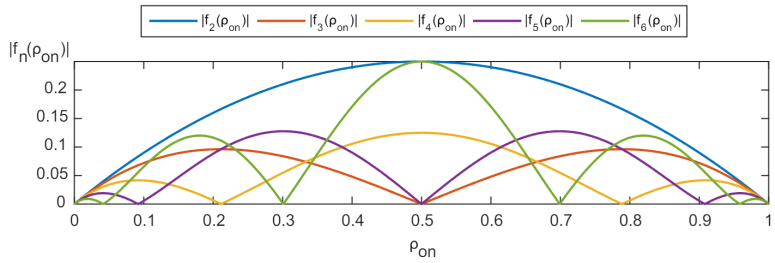


Figure 3.3: Second to sixth order polynomials of the on-time ratio as a function of the on-time ratio.

These amplified brightness levels can be corrected without compromising resolution[17]. To perform this correction, the cumulants must be deconvolved, allowing high-frequency components to be recovered. This is typically done using Lucy–Richardson deconvolution or Fourier reweighting, which enables an n-fold resolution improvement when the PSF is raised to the n-th power. In this case, Lucy–Richardson deconvolution is used, as it provides the most likely object representation given an image with a known PSF and assuming Poisson-distributed noise, typically requiring 10-100 iterations.

The standard way to linearize the brightness response[29] is to take the n-th root of the deconvolved n-th order cumulant image $\hat{g_n}$.

$$\bar{g} = \hat{g}^{\frac{1}{n}} \tag{3.7}$$

where $\bar{g}$ is the linearized cumulant image.

An other method to linearize is based on the on-time ratio[17], where the correction factor for deconvolved n-th order cumulant image $\hat{g}_n$ is $1/f_n(\rho_{on})$, which can be written as:

$$\frac{\hat{g}_n}{f_n(\rho_{on})} = \hat{g}_n^{\frac{log_{10}(\hat{g}_n/f_n(\rho_{on}))}{log_{10}(\hat{g}_n)}} \tag{3.8}$$

Here, instead of taking the n-th root, the adaptively linearized cumulant image is:

$$\bar{g}_n = \hat{g}_n^{\frac{1}{n}\frac{log_{10}(\hat{g}_n/f_n(\rho_{on}))}{log_{10}(\hat{g}_n)}} \tag{3.9}$$

In both methods, to reduce amplified noise and the creation of deconvolution artifacts, small values (typically 1-5% of the maximum value) are truncated and the image is reconvolved with $U(n\vec{r})$, where $U(\vec{r})$ is the systems PSF and n the cumulant order. This results in an SR image with an approximate n-fold improvement in spatial resolution compared to the diffraction-limited image.

Adaptive linearization requires a good estimation of molecular parameters, thus only making sense if a reliable fourth-order SOFI is available.

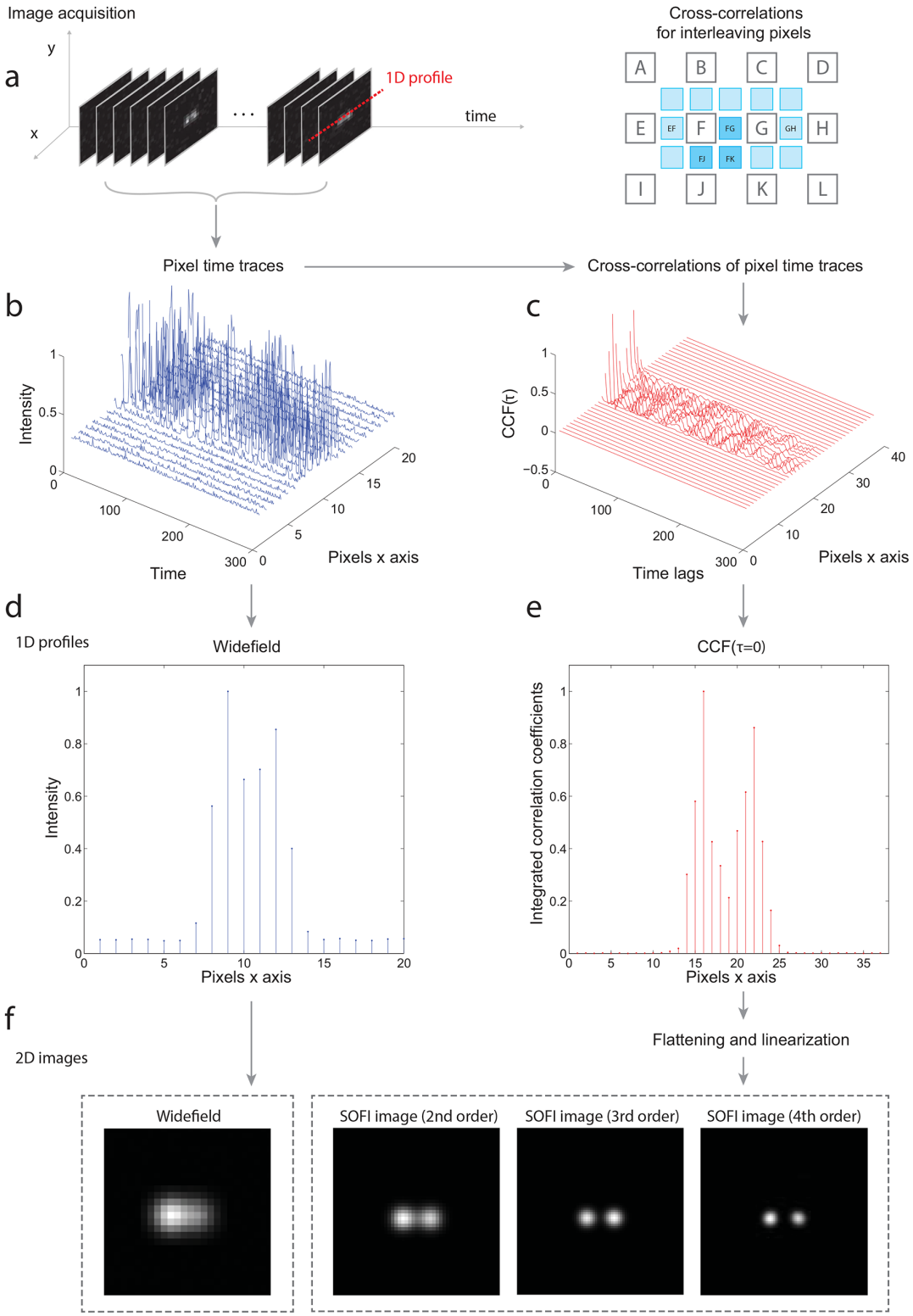


Figure 3.4: The principle of SOFI with cross-cumulants in a one-dimensional example. (a) One-dimensional profile extracted from the a series of images featuring two blinking emitters. (b) Corresponding one-dimensional intensity time trace. (c) Second order cross-cumulant are computed from these traces, primarily focusing on zero-time lag ($\tau = 0$). Interleaving pixels are also determined through this process. (d) The resulting widefield image shows the temporal average of intensity traces. (e) The 2nd order cross-cumulants for $\tau = 0$. (f) the resulting 2D SOFI images up to the 4th order cumulant order, following flattening and linearization.

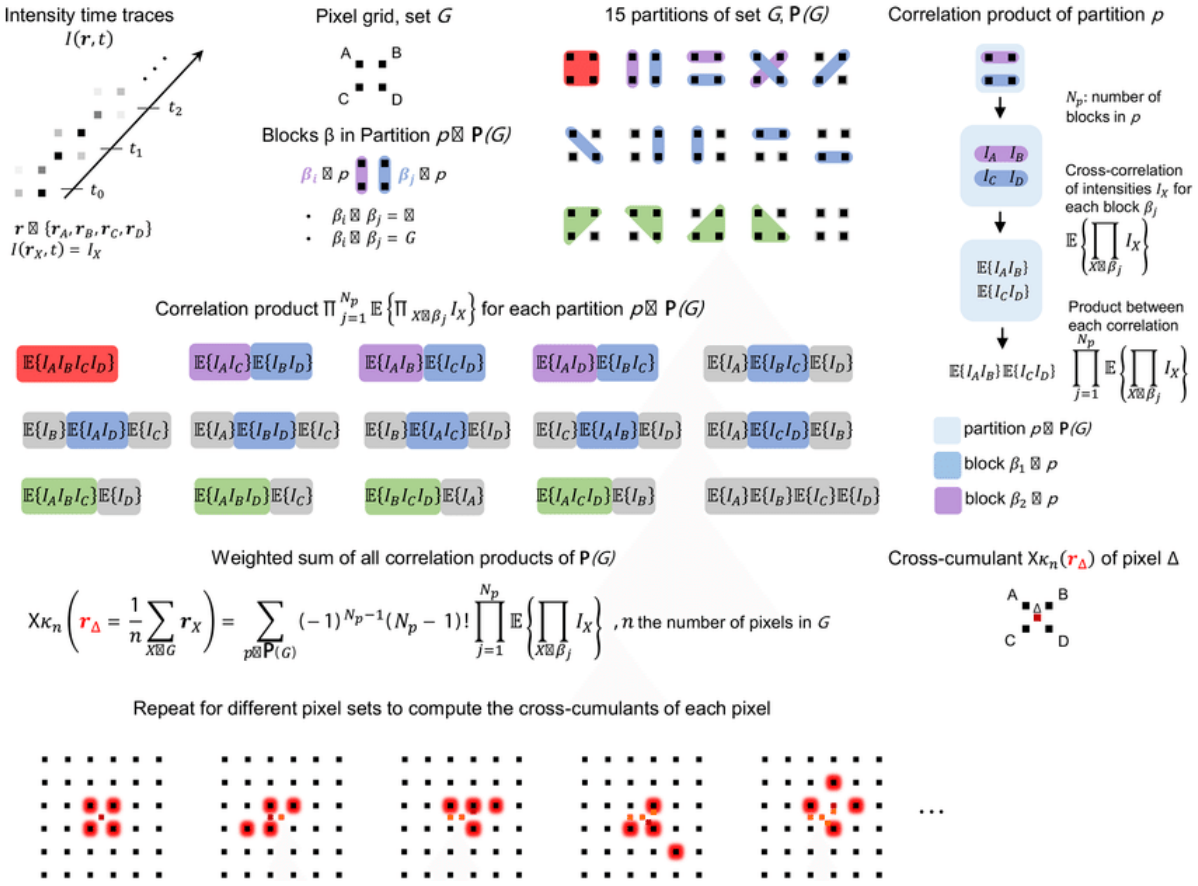


Figure 3.5: The process of SOFI for cross-cumulant calculation. The n-th order cross-cumulant K_n for pixel i is determined by summing weighted contributions from all possible combinations of n pixels within a set G. The location of pixel i is determined by the geometric mean of the n pixels in set G. This method allows the calculation of the n-th order cross-cumulant for any large grid of pixels by varying the sets of n pixels used.

4

SOFI Architecture

The aim is to reconstruct a SOFI-based SR image while using the least amount of LR images to achieve similar spatial resolutions and gain temporal resolution. The network we use is mostly inspired from the work of [30] due to its simple design, which has fewer trainable parameters compared to the U-Net architecture[23] and incorporates both spatial and temporal information. We modified this network and used it to accelerate cross-cumulant version of SOFI. This section introduces our SOFI model architecture and the used loss function. Different parts of the architecture, namely the encoder, fusion, and decoder stage are presented in a 3D representation.

4.1. Architecture

To learn a mapping from low-resolution (LR) images to a single SR image is called multi-image super-resolution (MISR). We define the LR image of a scene $l_i \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent depth, height, and width, respectively. The high-resolution (HR) image is defined as $Y \in \mathbb{R}^{C\times \gamma H\times \gamma W}$, where γ represents the up-sampling factor, approximately 1.984 for 2nd order. The channel depth C is 1 within the context of this research, as it represents gray scale images. The model is formulated as $\hat{Y} = f^{\gamma}_{\theta,\alpha,\beta}(\{l_1, ..., l_N\})$, where $\hat{Y}$ is the predicted SR image, N represents the number of LR images, and θ, α, and β represents the parameters of the encoder, fusion, and decoder stages within the architecture. An overview of the architecture can be seen in figure 4.1, which depicts the different stages. The main reasoning behind them is as follows:

1. *Encoder*: Encodes relevant latent representations from the LR images.
2. *Fusion*: Extracts correlated blinking information across the latent representations and averages the feature maps.
3. *Decoder*: Reconstructs the HR image.

In the following sub-sections, the stages are explained in detail.

4.1.1. Encoder

The encoder consist of two convolutional layers and two residual blocks[31]. Each block combines the two convolution layers, similar to [32], with Parametric ReLU activation functions[33]. Furthermore, 3×3 kernels are used for the convolution layers with 24 filters producing 24 feature maps for each frame at the input. The output of the encoder is given by $r_i = Encoder(l_i)$, where $r_i \in \mathbb{R}^{24\times H\times W}$ represents the latent representation of the input frame l_i. With a total of N LR frames to process, the encoder stage becomes:

$$(r_i)_{i=1}^{N} = Encoder\left((l_i)_{i=1}^{N}\right). \tag{4.1}$$

4.1.2. Fusion

The LR frames contain blinking emitters, and leveraging these blinking statistics requires information flow between these frames within the model. This process resembles how RNNs process sequential

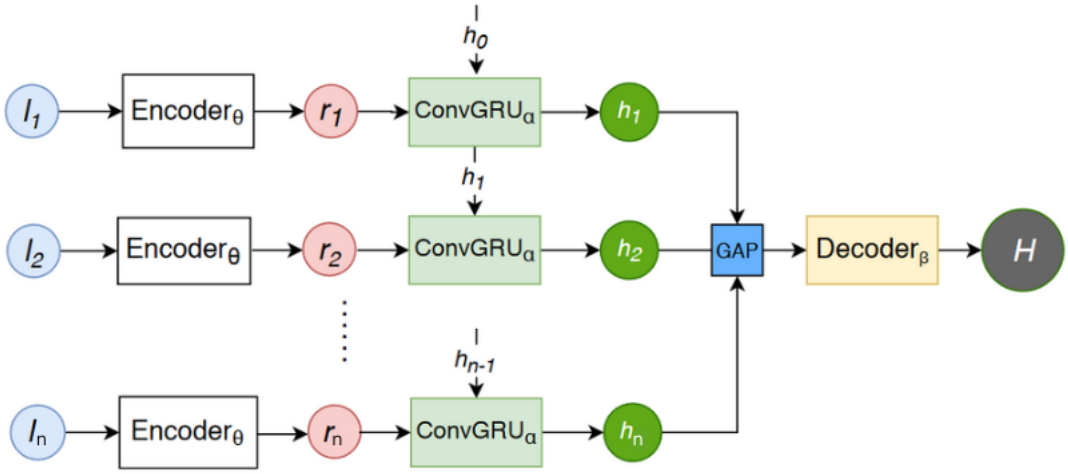


Figure 4.1: Overview of the architecture which depicts the encoder, fusion, and decoder stages.

data, but in the context of CNNs. An example of such a model designed for processing these frames is the ConvGRU model[34], which incorporates a GRU architecture[20]. In the ConvGRU, the fully-connected layers are replaced with convolutional layers in both the input-to-state and state-to-state connections. The hidden state, h_t, within the ConvGRU is recurrently connected to its adjacent sequential states. Updating the hidden state involves a convolutional operation with the input feature map, x_t, and the previous hidden state, h_{t-1}, according to the following procedure:

$$u_t = \sigma\left(W_i * [x_t, h_{t-1} + b_i]\right) \tag{4.2a}$$

$$r_t = \sigma\left(W_j * [x_t, h_{t-1} + b_j]\right) \tag{4.2b}$$

$$c_t = tanh\left(W_h * [x_t, r_t \odot h_{t-1}] + b_h\right) \tag{4.2c}$$

$$h_t = (1 - u_t) \odot h_{t-1} + u_t \odot c_t \tag{4.2d}$$

Where W and b represents the weights and biases of the convolutions kernels, respectively. The symbols $*$ and $\odot$ denote the mathematical operations of convolution and element-wise multiplication, respectively.

ConvGRU can be described as $f_\alpha : \mathbb{R}^{N\times 24\times H\times W} \mapsto \mathbb{R}^{N\times C_{h^{GRU}}\times H\times W}$, which converts the N input representation $(r_i)_{i=1}^N$ to N hidden representations $(h_i^{GRU})_{i=1}^N$. After which, global average pooling (AVG) is applied on the first dimension to return $h_{avg} \in \mathbb{R}^{C_{h^{GRU}}\times H\times W}$.

To extract more complex features across larger areas in the frames, the fusion stage f_α can be stacked, denoted as h^{GRU^l} with $l = 1, ..., L$, where L is the number of layers. The fusion stage is defined as follows:

$$(h_i)_{i=1}^N = \left(h_i^{GRU^L}\right)_{i=1}^N = f_\alpha\left(r_1, ..., r_N\right) \tag{4.3a}$$

$$h_{avg} = GAP(h_1, ..., h_N) \tag{4.3b}$$

In the context of this research, L is 2 and $C_{h^{GRU}}$ equals 24.

4.1.3. Decoder

In this stage, the combined representations h_{avg} are upsampled and averaged into a tensor $\hat{Y}$ of the same shape as the SOFI image Y. The decoder stage is defined as:

$$\hat{Y} = decoder_\beta^\gamma\left(H_{avg}\right) \in \mathbb{R}^{c\times\gamma H\times\gamma W} \tag{4.4}$$

It consists of a deconvolution layer [35], followed by two convolution layers with Parametric ReLU activation functions for minor adjustments. Subsequently, a 1×1 convolution layer projects the output

feature map $\mathbb{R}^{24\times\gamma H\times\gamma W}$ into the SOFI image space of dimension $\mathbb{R}^{1\times\gamma H\times\gamma W}$. The two convolution layers use 3×3 kernels. The deconvolution layer is set according to the SOFI order experiment.

4.2. Loss Function

The Mean Absolute Error (MAE) and the Mean Squared Error (MSE) are popular loss functions in image SR reconstruction[36], which primarily focuses on the spatial domain, promoting pixel-wise estimates during training. However, SR is closely associated with the frequency domain; in our case, HF content above the Nyquist-frequency η_c must be recovered from a set of LR frames $l_i \in \mathbb{R}^{1\times H\times W}$ to reconstruct the HR SOFI image $\hat{Y} \in \mathbb{R}^{1\times\gamma H\times\gamma W}$. Unlike the spatial domain, where these missing frequency cannot be fully separated, they can be in the Fourier domain. Therefore, we adopted the loss function proposed in [37], with the exception that we do not normalize the frequency components. Frequency normalization could downweight HF components relative to LF ones, which would be detrimental to preserving fine details in the SR task. Our loss function is defined as follows:

$$L_{\mathcal{F}} = L_{\mathcal{F}_A} + L_{\mathcal{F}_\angle} \tag{4.5a}$$

$$L_{\mathcal{F}_\mathcal{A}} = \frac{2}{UV}\sum_{u=0}^{U/2-1}\sum_{v=0}^{V-1}\left||\hat{Y}|_{u,v} - |Y|_{u,v}\right| \tag{4.5b}$$

$$L_{\mathcal{F}_\angle} = \frac{2}{UV}\sum_{u=0}^{U/2-1}\sum_{v=0}^{V-1}\left|\angle\hat{Y}_{u,v} - \angle Y_{u,v}\right| \tag{4.5c}$$

Here, both SR images $\hat{Y}$ and 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. The amplitude and phase component in the transformed image $X_{u,v}$ can be determined as follows:

$$|\mathcal{F}\{x\}_{u,v}| = |X_{u,v}| = \sqrt{\mathbb{R}\{X_{u,v}\}^2 + \mathbb{I}\{X_{u,v}\}^2} \tag{4.6a}$$

$$\angle\mathcal{F}\{x\}_{u,v} = \angle X_{u,v} = arctan\left(\frac{\mathbb{I}\{X_{u,v}\}}{\mathbb{R}\{X_{u,v}\}}\right) \tag{4.6b}$$

The discrete Fourier transform of the image is calculated as follows. Note that we do not normalize the frequency components:

$$\mathcal{F}\{x\}_{u,v} = X_{u,v} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x_{h,w}\cdot e^{-i2\pi\left(u\frac{h}{H}+v\frac{w}{W}\right)} \tag{4.7}$$

Where the image $x \in \mathbb{R}^{C\times H\times W}$ is transformed into the Fourier space $X \in \mathbb{C}^{C\times U\times V}$.

5 Datasets

In this chapter, we discuss how the datasets were created, as there are no publicly available datasets for SOFI. First, we explain how the synthetic microtubules dataset was generated. Finally, we describe how we created a dataset from real microscope images.

5.1. Synthetic Data

To create the synthetic dataset, the SOFI simulation tool[1] and the SOFI package[2] are used to generate the low-resolution (LR) frames and the corresponding high-resolution (HR) SOFI images in Matlab. The pipeline used to generate this dataset is shown in figure 5.1. Note that from the set of 100 frames, only the first 25, 20, etc., are used as input for our model and the SOFI image as target.

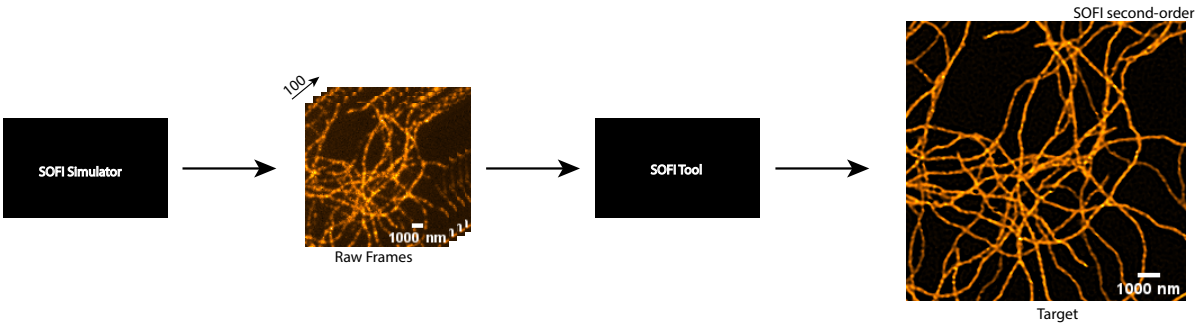


Figure 5.1: Pipeline on how the synthetic dataset was generated using the SOFI simulation tool and SOFI package.

5.1.1. Simulation

The uniqueness of the SOFI simulation tool lies in its use of the real physical properties of an sCMOS camera [38], combined with a fluorescent microscope model. With this tool, one can generate movies that are more realistic compared to other simulations, which simply add Gaussian or Poisson noise. A real sCMOS-based noise map can be established from measurements [39], consisting of an offset map, variance map, and photon response map. These maps introduce noise to the frames generated by our fluorescent microscope model, in accordance with the physical characteristics of the sCMOS camera. By combining these with a fluorescence emitter model and an optics model, we can create more realistic simulations.

The camera model we use is called the *Willy FPS_100*, and the following general simulation parameters are applied:

[1] `https://github.com/GrussmayerLab/Sofi_Simulations.git`
[2] `https://github.com/GrussmayerLab/sofipackage.git`

Table 5.1: General parameters of the SOFI simulation tool.

fps	rectime [s]	fwhm_psf [nm]	sbr
100	1	220	[15, 15, 15, 15]

Here, *fps* stands for frames per second, *rectime* is measurement duration, *fwhm_psf* represents the system's point spread function (PSF) in nanometers, and *sbr* is the signal-to-background ratio.

For the fluorescence emitter model, the following parameters are used:

Table 5.2: Parameters settings of the fluorescence emitter model of the SOFI simulation tool.

dens [emitter/μm^2]	ontime [ms]	offtime [ms]	avbleach [s]	Ion [signal/emitter/frame]
5000	[10, 10, 10]	[600 1200 2400]	500	[100, 80, 60, 40]

Where *dens* represents the density of the emitters, *ontime* signifies the on-time duration of the emitters, *offtime* refers to the off-time duration of the emitters, *avbleach* denotes the average bleaching, and Ion represents ionization, contributing to the average signal per emitter per frame.

With these settings, the microtubules dataset was simulated based on the physical microtubule model from [40]. The model simulates the random growth and distribution of microtubules within a 2D binary grid, where ones represent the microtubule structure and zeros represent the background. It generates a random number of microtubules, each starting from a random position and extending in random directions while following physical constraints such as bending stiffness and length. The growth continues step-by-step until the microtubule reaches a predetermined length or hits the edge of the field of view. The output is a collection of paths representing the trajectories of individual microtubules in the 2D binary grid.

For the 2D binary grid, a grid size of 2560 by 2560 pixels with a pixel size of 10 nm was used, and the number of microtubules was randomly set between 20 and 100. Emitters were then randomly distributed over the grid, but only where the grid indicated a microtubule structure (a one in the grid). These emitters exhibit the blinking characteristics outlined in table 5.2. A one-second movie with a frame rate of 100 frames per second was captured using these blinking emitters. The movie was then convolved with the PSF, 10$\times$ downsampled, and further corrupted by the sCMOS noise map, resulting in a simulated microtubule structure of 100 frames.

For the dataset, we simulated a range of emitter densities and signal-to-noise ratios (SNR) to mimic real-world conditions. To achieve different emitter densities, the off time was adjusted: longer off times resulted in sparser densities, while shorter off times resulted in denser distributions. Three different emitter densities were generated, each with varying Ion values to produce different SNR levels. An Ion value of 40 results in the worst SNR, while an Ion value of 100 provides the best SNR level. In total, 12 different sets of 100 frames with a size of 256$\times$256 pixels were generated for one dataset, out of a total of 212 datasets[3], which includes a training set and test set.

5.1.2. SOFI

With the frames in place, the SOFI tool can be used to generate the second order SOFI SR images. The following main settings were used in the SOFI tool:

Table 5.3: Parameters settings of SOFI tool.

blcor	dcor	orders	subseqlength [frames]	fwhm [μm]	iter	psfmodel	pxy[nm]
0	0	[1:3]	100	1.1	10	'gaussian'	100

[3]More details of the script and settings can be found at `https://github.com/GrussmayerLab/Sofi_Simulations.git`

Here, *blcor* refers to bleach correction, which accounts for photobleaching, while *dcor* refers to drift correction. Orders represents the SOFI orders, and *fwhm* stands for full width at half maximum, which defines the PSF (point spread function) and is halved after second-order SOFI. *Iter* denotes the number of iterations used during the flattening and linearization steps, where Richardson-Lucy deconvolution is applied to deconvolve the PSF. *Psfmodel* specifies the type of PSF model. Finally, *pxy* indicates the pixel size.

In the end, the training set consisted of 2000 samples, while the test and evaluation sets each contained 500 samples. The frames were cropped to a size of 128×128 pixels, and the target images to 249×249 pixels. All images were normalized by dividing by the maximum value of the 16-bit range. An example of simulated microtubules is provided in figure 5.2.

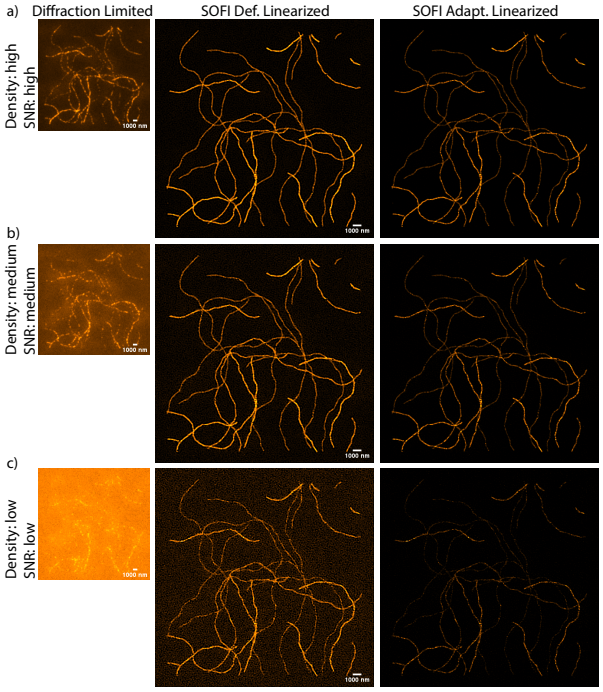


Figure 5.2: Left to right: single diffraction-limited frame of a simulated microtubule at different emitter densities and SNR levels. Here, the emitter density is 5000 emitters/μm^2 with on times of 10 ms and off times of 600 ms, 1200 ms, and 2400 ms, respectively, to simulate different densities. The SNR level is additionally set by the Ion value, which in this case is 100, 60, and 40, respectively. Each row represents different emitter densities and SNR levels of a microtubule structure; default linearized SOFI SR image based on 100 frames; adaptive linearized SOFI SR image based on 100 frames. Scale bar: 1000 nm.

5.2. Microscopy Data

As real SOFI data is scarce, we applied random cropping of 128×128 patches and 90° or 270° rotations to match the dataset size to that of the synthetic data. The dataset we used had the following minimum requirements:

1. At least 4 fixed-cell datasets of the same cell type.
2. 3,000 or more frames.
3. High-density fluctuation data.
4. A field of view (FOV) of approximately 350×350 pixels or larger.

After which the SOFI tool is used to create the corresponding target images. Again, all images were normalized by dividing by the maximum value of the 16-bit range. The following settings were used for the microtubules and mitochondria. Additionally, background subtraction by *Tekpinar et al.* [41], was applied for mitochondria data to reduce background artifacts in the SR reconstruction for SOFI.

Table 5.4: Parameters settings of SOFI tool microtubules.

blcor	dcor	orders	subseqlength [frames]	fwhm [μm]	iter	psfmodel	pxy[nm]
0	1	[1:3]	1000	4.2	10	'gaussian'	105

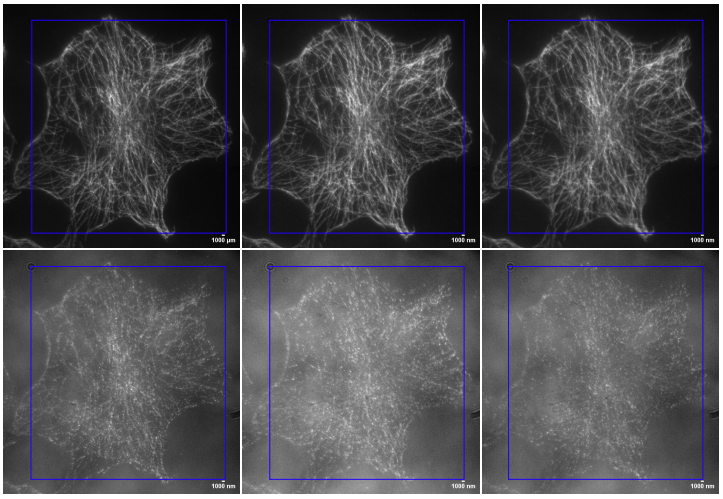


Figure 5.3: This dataset consists of fixed-cell microtubules in six sets. Three of the sets have a relatively high signal-to-noise ratio (SNR) and dense fluorophores, while the other three have a relatively lower SNR and sparser fluorophores. The blue box represents the area for random cropping, generating 128×128 fields of view (FOV). The upper three sets contain 10,000 frames, and the lower three sets contain 3,387 frames. All images have a resolution of 782×804 pixels. Scale bar: 1000 nm.

Table 5.5: Parameters settings of SOFI tool mitochondria.

blcor	dcor	orders	subseqlength [frames]	fwhm [μm]	iter	psfmodel	pxy[nm]
1	1	[1:3]	5000	3.8	10	'gaussian'	105

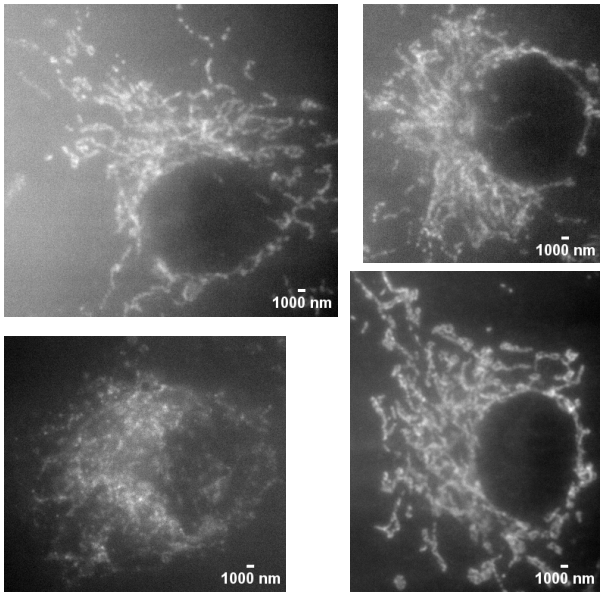


Figure 5.4: This dataset consists of four sets of fixed-cell mitochondria, characterized by a relatively lower signal-to-noise ratio (SNR) compared to the microtubules dataset, but with very dense fluorophores. Each set contains 10,000 frames. The image resolutions, from top left to bottom right, are 441×413, 313×342, 372×338, and 330×424 pixels, respectively. Random cropping is applied across the entire image. Scale bar: 1000 nm.

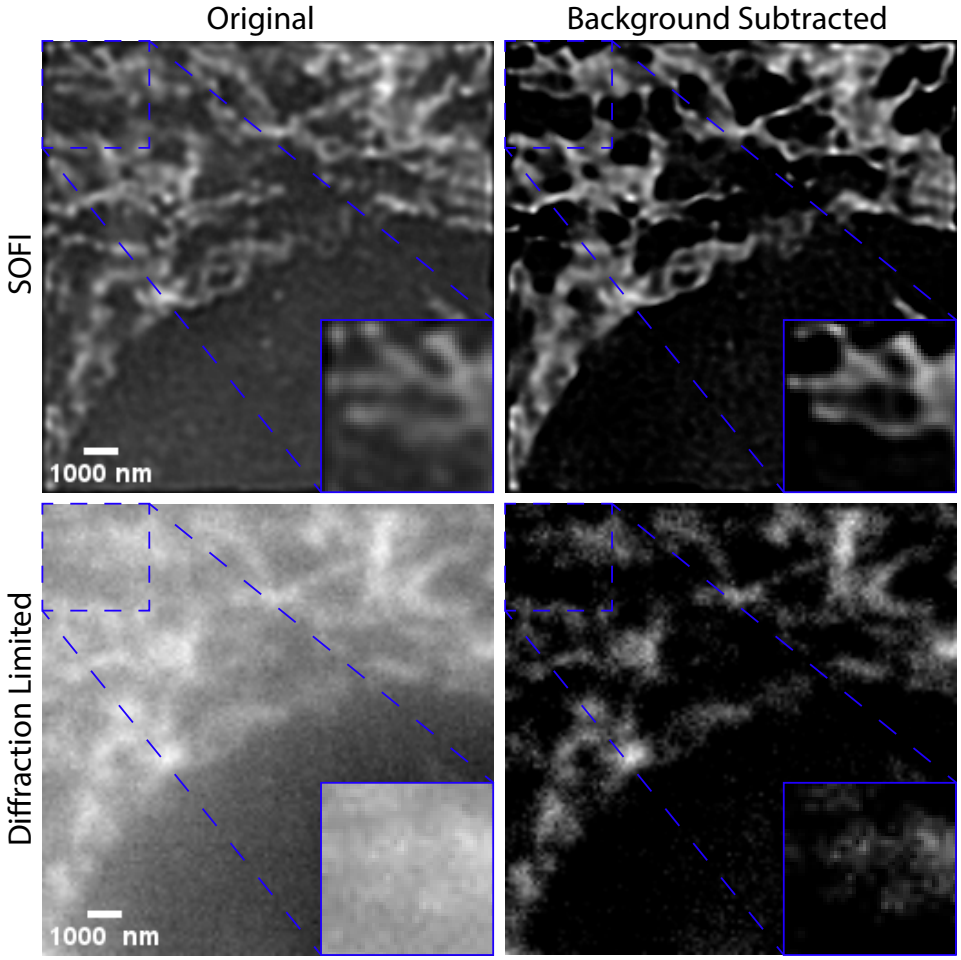


Figure 5.5: The background subtraction algorithm significantly reduces the number of background artifacts. However, finer mitochondrial structures are lost, making the remaining structure appear somewhat averaged compared to the original diffraction-limited frame. Diffraction-limited frame bilinearly interpolated to match the resolution of SOFI. Scale bar: 1000 nm.

6

Accelerating SOFI for Live-cell Imaging

The main results are presented in a manuscript following the guidelines of *Nature Methods*.

Accelerating SOFI with Deep Learning: Enabling Real-Time live-cell Imaging.

Jelle Komen[1,2*], Miyase Tekpinar[1], Kristin Grußmayer[1], Nergis Tömen[2]

[1*]Deparment of Bionanoscience and Kavli Institute of Nanoscience Delft, Delft University of Technology, Delft, Netherlands.
[2]Computer Vision Lab, Delft University of Technology, Delft, Netherlands.

*Corresponding author(s). E-mail(s): j.j.m.komen@student.tudelft.nl;
Contributing authors: m.tekpinar@tudelft.nl; k.s.grussmayer@tudelft.nl;
n.tomen@tudelft.nl;

Abstract

Live-cell imaging captures dynamic cellular behaviors, but many structures are beyond the diffraction limit. Super-resolution Optical Fluctuation Imaging (SOFI) overcomes this by using the statistical relationship of the blinking fluorophores, achieving n-fold spatial resolution for nth order SOFI cumulant calculations. However, SOFI requires hundreds of frames and extensive post-processing, making it unsuitable for real-time live-cell imaging of fast processes in live-cells. We introduce a deep learning model to accelerate SOFI, enhancing temporal resolution while maintaining spatial improvements. The model exchanges temporal information to extract correlated blinking information across latent representations. Using synthetic and real fixed-cell microtubule data, our model generates super-resolved SOFI images from just 20 diffraction-limited frames, eliminating background artifacts and achieving a 2-fold spatial resolution. After training on static mitochondria data, it can reconstruct super-resolved images in dynamic environments, enabling real-time live-cell studies up to 4.85 fps.

Keywords: Fluorescence microscopy, Deep Learning, Super Resolution, Live-cell imaging

1 Introduction

Live-cell 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. To visualize the structures, fluorescent molecules, called fluorophores, are first used to label the specimen. They bind to target proteins or structures and emit light upon excitation. A sequence of frames is then captured, with molecules randomly blinking across the frames. Fluorophores can blink sparsely for precise localization or densely. However, visualizing these cellular dynamics is constrained by the diffraction limit of optical lenses, which restricts spatial resolution to around 200-300 nm (half the wavelength of visible light).

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], which can achieve spatial resolution down to 10 nm. However, these methods either require

complex microscope setups, higher signal-to-noise ratio (SNR) conditions (higher illumination), or even millions of frames to reconstruct an SR image, which can limit temporal resolution and can damage the sample[6, 7]. Other techniques, like Super-resolution Optical Fluctuation Imaging (SOFI)[8] and Super-Resolution Radial Fluctuations (SRRF) microscopy[9] do not require a complex microscope setups and typically use hundreds of frames to reconstruct an SR image as they can handle very dense blinking fluorophores.

Only SOFI generates SR images based on statistical theory by leveraging the correlation of blinking from a fluorophore over time and space (cross-cumulant). The resolution improvement arises from the independence of different fluorophores [10, 11]. Because noise does not correlate over time, SOFI can operate under low SNR conditions, making it suitable for live-cell imaging. For n-order SOFI, an n-fold improvement in spatial resolution is achieved; however, it still requires hundreds of frames, limiting its ability to capture dynamic cellular processes and making real-time imaging impractical. Without real-time capabilities, researchers cannot make immediate decisions, wasting valuable time. Deep learning could help reduce the number of frames needed for SR image reconstruction while maintaining spatial resolution improvements.

Deep learning for SR methods has been extensively explored like SIM[12–14] and SMLM [15–18]. These networks aim to enhance both spatial and temporal resolution for live-cell imaging. Most methods employ the U-Net architecture[12, 14, 15, 19], while others utilize an encoder-decoder network[20] or uniquely leverage spatial and temporal information[13, 16, 17]. A notable example is the study named DBlink by *Saguy et al.*[17], achieving 15 ms temporal resolution with 30 nm spatial resolution. However, it post-process based, meaning it still requires millions of frames, thus compromising the health of the sample.

For SOFI, a study by a study by *Qu et al.*[19] introduced a self-supervised denoising model to address artifacts in second-order auto-cumulant SOFI images caused by using fewer frames [21]. This approach allows for the use of 20 frames while achieving a 130 nm spatial resolution improvement compared to the 140 nm by SOFI, enabling live-cell imaging. However, its multiple processing stages render it unsuitable for real-time applications.

We propose an end-to-end deep learning model that accelerates cross-cumulant SOFI by reconstructing a second-order SR image from just 20 frames, compared to the hundreds typically required by SOFI, while still achieving a 2-fold spatial resolution improvement. Unlike U-Net methods that use only spatial information, our model leverages spatiotemporal data to extract correlated blinking fluorophores. It operates in three stages: encoding input frames into feature maps, fusing them with a recurrent structure to capture blinking correlations, and upsampling the result into a second-order SR image. The model is trained in a supervised manner, using the pre-trained weights from synthetic data, we can train the model on real fixed-cell microscopy data using at least four different measurements of the same cell type. By applying random cropping and rotations, the dataset expands to around 2000 samples. This approach enables the model to be used in live-cell experiments, making it a practical method for real-world applications. My contributions are as follows:

1. Propose a SOFI model, tailored from the framework of [22] (see Chapter 4).
2. Our model reconstructs second-order SR images from 20 frames of real microscopy data, enabling capturing dynamic movements in live-cell experiments, while SOFI requires thousands of frames.
3. The model effectively works with motion-controlled real mitochondria data, validating its potential for live-cell imaging, which for SOFI is not possible in these temporal resolutions.
4. Optimized the model's latency and benchmarked it against SOFI and the U-Net architecture, achieving real-time temporal resolution of up to 4.85 fps.

Additional contribution:

1. Integrated a microtubule physics model from [23] into the SOFI simulation tools [10] by *Tekpinar et al.* to create synthetic datasets which includes more realistic noise and background models. This can be used for future research for higher-order SOFI datasets.

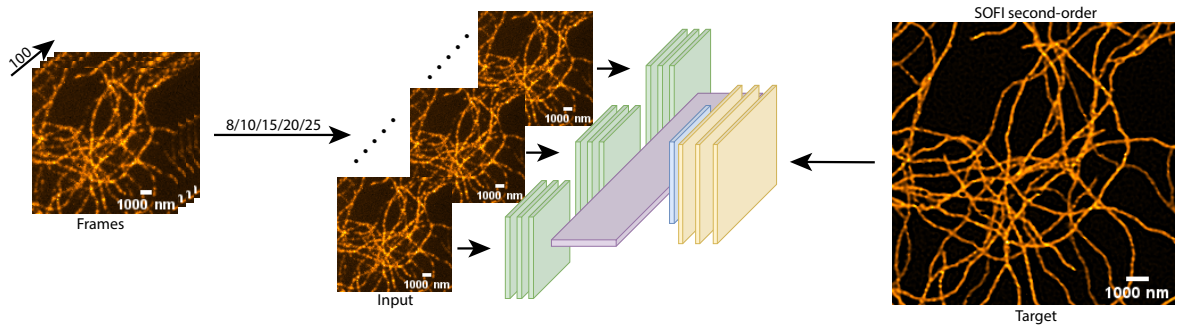


Fig. 1 Experimental setup for accelerating second-order SOFI, using 100 frames. The model is trained with input sizes of 8, 10, 15, 20, and 25 frames. Green represents the encoders, purple the fusion layer, blue the global average pooling, and yellow the decoder, which upsamples the combined representation into a second-order SR image. For details, see chapter 4 and supplementary material, chapter A. The scale bar for frames, input, and target is 1000 nm.

2 Results

2.1 SOFI Acceleration

In this section, we use synthetic fixed-cell (static scene) microtubules, including ground-truth (GT) representing the ideal SOFI SR reconstruction, as a static scene (see chapter 5) to determine the minimal numbers of frames required for our SOFI model to reconstruct an SR image while maintaining spatial resolution improvement. Additionally, we asses the preservation of the microtubule structure, as spatial resolution alone does not guarantee accurate structural integrity. Specifically, this involves:

1. Assessing the spatial resolution of the predicted SR image using decorrelation analysis[24], which is commonly used in SR microscopy.
2. Evaluating the correlation between the predicted image and the GT image using Pearson correlation, a standard measure of similarity in bioimaging and SR microscopy.
3. Assessing the True Positive Rate (TPR), True Negative Rate (TNR), and confusion matrix of the SR reconstruction by binarizing the images (see supplementary material, chapter B). These are commonly used metrics in machine learning.

The dataset includes various SNR levels and fluorophore densities simulating real-world conditions with a diffraction-limit resolution of 220 nm. It is divided into training (2,000 samples), evaluation, and test sets (480 samples each). The model is trained on pairs of adaptively linearized second-order SOFI images from 100 frames. Although adaptively linearized SOFI images show lower brightness and more structural loss, they avoid artifacts in SR reconstruction, leading to better generalization than using default linearization (see supplementary material, chapter E). In supplementary material, chapter F, a comparison between models trained on adaptively linearized and default SOFI images is provided, showing almost no background artifacts (see TNR in E.2(d) and F.2(d)). This section focuses on adaptive linearization for conciseness.

We trained models using 8, 10, 15, 20, and 25 frames to reconstruct SR images, as in [25]. Figure 1 illustrates the experimental setup.

In figure F.2(a), the models with 25 and 20 frames come closest to the theoretical spatial resolution of 110 nm, deviating by about 5 nm, while SOFI also approaches this limit. In contrast, the Pearson correlation in figure F.2(b) shows that all model sizes score higher than SOFI, likely due to the fact that adaptively linearized SOFI is not typically used for second-order SOFI (see chapter 3.3). Lower brightness levels and structural loss in adaptively linearized SOFI are reflected not only in figure F.2(b) but also in the TPR (figure F.2(e)) and the confusion matrices (figure B.3).

Despite these issues, the models generalized well, successfully reconstructing filaments as SOFI target images lost structure at lower SNR levels and fluorophore densities (see figure 2). As this only affected a subset of the dataset, the model was able to generalize during training, resulting

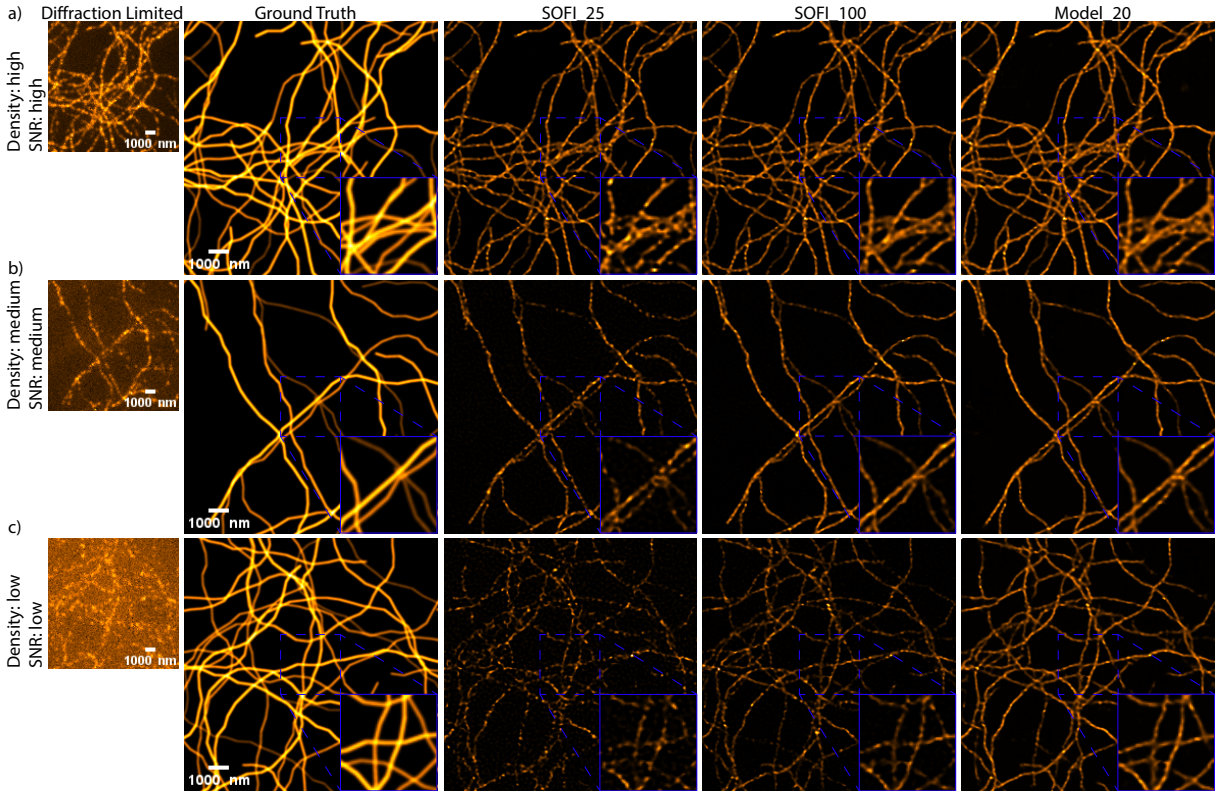


Fig. 2 Comparison of SR reconstructions between the GT, adaptive linearized SOFI, and model. Left to right: single diffraction-limited frame of a simulated microtubule at different fluorophores densities and SNR levels. Here, the fluorophores density is 5000 fluorophores/μm^2 with on times of 10 ms and off times of 600 ms, 1200 ms, and 2400 ms, respectively, to simulate different densities. The SNR level is additionally set by the Ion value, representing illumination intensity, which in this case is 100, 60, and 40, respectively. Each row represents different fluorophores densities and SNR levels of a microtubule structure; SR ground truth; SOFI SR image based on 25 frames; SOFI SR image based on 100 frames; model-based SR image based on 20 frames. Region of interest (ROI) marked by a blue dashed line, showing better filament reconstructions for the model based SR reconstructions. Scale bar: 1000 nm.

in better filament reconstruction (figure 2c). The ability to operate in lower SNR conditions allows for reduced laser power, beneficial for samples, or longer measurements. The model's use of temporal information further enhances filament reconstruction, as demonstrated in figure F.3, where closely located but disconnected filaments are connected if the trajectory is correct. In contrast, the U-Net architecture, which lacks temporal information, struggled with filament reconstruction, as evidenced in figure D.3. This limitation is also reflected in the TPR (figure D.1(d)) and confusion matrix (figure D.2), despite U-Net being a larger model (see supplementary material, chapter D).

Finally, a more detailed assessment of the Pearson correlation in figure F.2b reveals a significant downward trend starting at model size 15. This pattern is echoed in the TPR (figure F.2(e)), where models with 10 and 8 frames score lower. Confusion matrices in figures B.3(e) and B.3(f) show that models using 25 and 20 frames perform similarly, with only a 0.05% difference in true positives (TP) and false negatives (FN). Ultimately, the model trained on 20 frames strikes the best balance between spatial resolution, structural preservation, and temporal resolution.

2.2 Fixed-Cell Microtubules

The model based on 20 frames, which has demonstrated to work with synthetic data, is further evaluated using real fixed-cell microtubules obtained from a microscope (see methods) to validate its performance on real data and assess whether the results align with those from the first

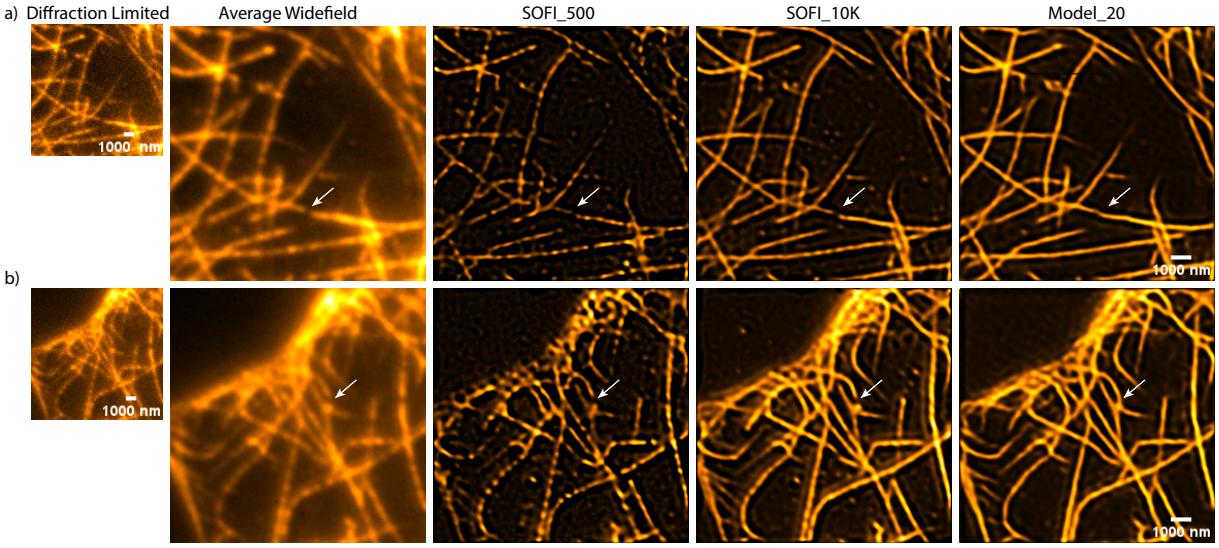


Fig. 3 (a-b) From left to right: single diffraction-limited frame of a microtubule; average widefield image based on 20 frames which is up-sampled using bilinear interpolation to match the SOFI and model resolutions. The white arrows depict the location where the model connect the disconnected filaments while there is a gap between the emitters in the average widefield image; SR reconstruction of SOFI based on 500 frames; SR reconstruction of SOFI based on 10k frames; SR reconstruction of the model based on 20 frames. Scale bar: 1000 nm.

experiment. Since no GT image is available in this scenario, we utilize the SQUIRREL algorithm [26] to assess structural integrity. SQUIRREL enables the comparison of the SR prediction with the standard deviation (STD) widefield image, which is computed from the total number of frames. It achieves this by degrading the SR image using an estimated point-spread function (PSF), a mathematical representation of how a microscope blurs a point of light and limits spatial resolution, derived from the STD widefield image until the convolved SR image matches the STD widefield image. This process facilitates a direct comparison of both images and allows us to calculate the re-scaled Pearson correlation (RSP) coefficient to evaluate the accuracy of the SR reconstruction for both synthetic and real data. Again, decorrelation analysis [24] is used to asses the spatial resolution.

Using pre-trained weights from synthetic data allows us to leverage learned features and patterns, potentially speeding up the training process and improving generalization, especially when constrained by limited microscope data [27] [17, 28]. 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. The first three sets consist of 10k frames each, while the second set contains 3,387 frames, with a resolution of 782×804 pixels.

By randomly cropping to 128×128 pixels in a field of view (FOV) (see chapter 5) and applying data augmentation techniques—including 90-degree and 180-degree rotations—the dataset is expanded to include 2,250 samples. The evaluation and test sets each contain 480 samples. The second-order default linearization SOFI target images are based on either all 10k or 3,387 frames. The dataset features an asymmetrical PSF with an approximate diffraction limit of 420 nm.

The results are presented in supplementary material, chapter G, figure G.1, and figure 3. Similar properties are observed: the model-based SR reconstructions show almost no background artifacts and effectively approximate the filaments again. However, the model is more aggressive in connecting the filaments compared to the first experiment (see figure F.3), where larger gaps between fluorophores were filled in, given the correct trajectory (see figure 3(a-b)). This behavior can be attributed to the fact that half of the dataset consists of sparse fluorophore distribution,

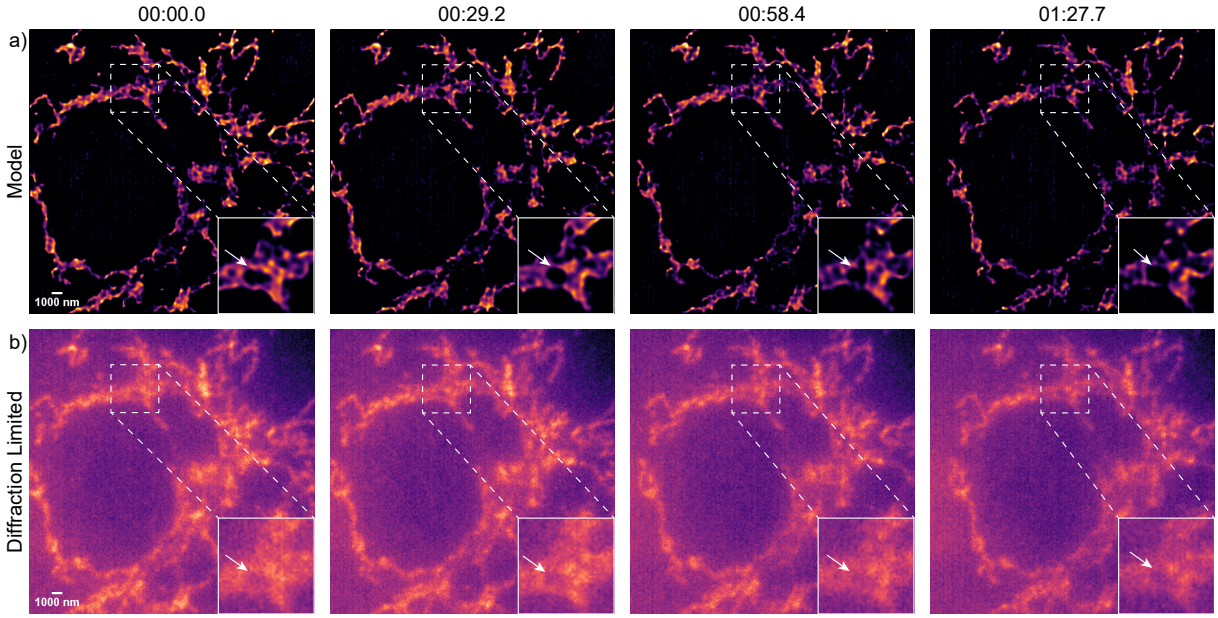


Fig. 4 Reconstruction of a 1.2-minute video of mitochondria in a motion-controlled environment. Using a rolling window, we can output SR images equal to the camera's frame rate. (a) Reconstruction of the model, showing almost no background artifacts. (b) Diffraction-limited frame bilinearly interpolated to match the resolution of the model. The ROI, marked by a white dashed line, illustrates the model's ability to detect structural changes over time. The white arrow highlights the model's capability to capture the forming of a "bubble". Scale bar: 1000 nm.

while the other half consists of very dense fluorophore distribution. This mix encourages the model to make more aggressive approximations during training to minimize loss, particularly when using only 20 frames, where some fluorophores are ultimately lost. As a result, there are more filament approximations, even though there are larger gaps between disconnected filaments (see average widefield in figure 3(a-b)).

From the decorrelation analysis in figure G.1(d), the model achieves spatial resolutions closer to the theoretical limit, deviating by around 9 nm compared to SOFI's 20 nm. This is likely because decorrelation analysis is sensitive to background artifacts[24], which the model exhibits minimally. Comparing the models trained on real and synthetic data, they achieve a 1.90-fold and 1.91-fold improvement in spatial resolution, respectively, demonstrating similar levels of spatial resolution enhancement.

Figure G.1(a) shows the RSP distributions of SR reconstructions compared to the STD widefield images from either 10k or 3,387 frames, as evaluated using SQUIRREL. For reference, various SOFI-based images illustrate the increasing number of frames required to match the RSP performance of the model based on 20 frames. Notably, SOFI with 500 frames achieves similar RSP performance. In figure G.1(b-c), RSP distributions are compared between the model and SOFI for both real and synthetic data. Both models demonstrate similar RSP scores, with 0.85 on real data and 0.84 on synthetic data. Ultimately, this demonstrates that the results are consistent with those from the first experiment. However, due to half of the dataset consisting of sparser fluorophore distributions, the model becomes more aggressive in approximating filaments. To minimize this behavior, future datasets should focus on very dense fluorophore distributions including a range of SNR conditions, so the model can be used in lower SNR conditions. This is advantageous, as it allows the use of fewer frames with lower illumination (lower laser power) to visualize the sample, thereby reducing sample damage—a benefit not achievable with other SR methods[2–4].

2.3 Motion-controlled Mitochondria

To validate the model's ability to capture dynamic changes and its potential for live-cell imaging, it was tested on synthetic (see supplementary material, chapter J) and motion-controlled, fixed-cell mitochondria by moving the microscope stage in the x-direction. This setup simulates real-time cellular movement, bringing the experiment closer to live-cell conditions. The model was trained on four datasets of fixed-cell mitochondria with a diffraction limit of 380 nm, following the same pipeline as before (see Chapter 5), using 2,250 samples for training and 500 for evaluation and testing, derived from either 10k or 8,198 frames. Due to the microscope setup penetrating less deeply into the sample (see methods), the SOFI images had more background artifacts compared to the fixed-cell experiment. To address this, a pre-processing step was applied to subtract the background[29], which significantly reduced artifacts (see chapter 5, figure 5.5). As observed in the first experiment (see supplementary material, chapter E), the model is sensitive to background artifacts, and without background subtraction, it resulted in no SR improvement(see supplementary material, chapter H, figure H.3). However, this subtraction also resulted in the loss of finer structures, making the remaining features appear somewhat averaged compared to the original diffraction-limited frame.

Nevertheless, the model was able to capture dynamic structural changes (see figure 4), where as SOFI could not (see figure H.2). By using a rolling window, we can produce SR images at the camera's frame rate, N. However, the temporal resolution remains limited by a window of $\frac{20}{N}$. The white arrows highlight the model's ability to capture the formation of a "bubble," which evolves over time. This bubble formation can also be observed in the diffraction-limited frame, where it forms and shifts over time. Through decorrelation analysis, we measured an average spatial resolution of 221.41 nm, and using the SQUIRREL algorithm, we obtained an average RSP of 0.66, closely matching the test set average results of 214.9 nm in spatial resolution and a RSP value of 0.63 (see table H.1).

Ultimately, this demonstrates that the model is capable of reconstructing SR images in a dynamic environment. Conducting a live-cell experiment, as in [17, 18], would further validate the model's capabilities. Comparing the structural changes observed in a live-cell experiment to previously published work, such as studies on mitochondrial dynamics, including fusion and fission events [30], would provide valuable insights.

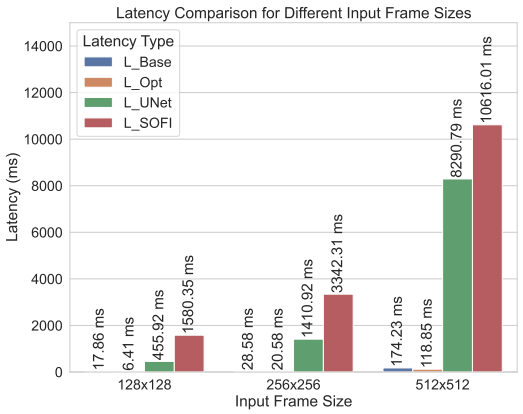


Fig. 5 Average latencies of the baseline model, optimized model, SOFI of 20 frames as input size. The avarage latencies are based on 500 repetitions.

2.4 Latency

For real-time imaging, the latency to calculate the SR image is crucial, as lower latency means higher temporal resolution. Initially, for our proposed model, the number of hidden layers was set to 24. Generally, more hidden layers enable the model to learn more complex features [27], but this comes at the cost of increased latency. Hence, for the model based on 20 frames, the first experiment was redone with a reduced number of hidden layers, ranging from 8 to 22 in steps of 2. The results can be found in the supplementary material, chapter H. Ultimately, 20 hidden layers were found to be the most optimal, as it came closest to the theoretical spatial resolution of 110 nm, even retaining more structure compared to 24 hidden layers and improving the latency. Therefore, 20 hidden layers were chosen as the optimal balance between latency and spatial performance.

For further optimizations, PyTorch TensorRT [31] was used to perform model optimization, such as layer fusion and precision calibration (e.g., using fixed-point floating points). The results can be found in figure 5, where the average latency is provided for different image sizes based on 500

repetitions. Additionally, the latency of the U-Net model (see supplementary material, chapter: E), as well as SOFI based on 20 frames, are presented for reference. Ultimately, the optimized model achieves real-time temporal resolutions of 4.85 fps, 4.53 fps, and 3.14 fps for frame sizes of 128×128, 256×256, and 512×512, respectively, assuming N is 100 fps. Despite the relatively low latencies, the model's input size of 20 frames remains a bottleneck for achieving higher real-time temporal resolutions. This limitation can be addressed by transitioning the architecture from a many-to-one configuration to a many-to-many configuration, like in [17], which would allow temporal resolutions of N.

3 Discussion

Live-cell imaging aims to maximize both spatial and temporal resolution while minimizing sample damage. While SMLM is widely used in fixed-cell imaging for its spatial resolutions up to 10 nm, it sacrifices temporal resolution and sample health. SOFI offers a more suitable alternative for live-cell imaging, as it manages denser fluorophores, operates under lower SNR conditions, and requires hundreds frames instead of millions like SMLM, though it still compromises temporal resolution. To address this, we developed a model that uses just 20 frames to reconstruct second-order SOFI images, minimizing background artifacts while approaching theoretical spatial resolutions.

Our model leverages temporal information to capture correlated fluorophore blinking across latent representations. Using synthetic microtubules, we found that 20 frames strike an optimal balance between spatial resolution, structural preservation, and temporal resolution, outperforming U-Net-based SR reconstructions. It also delivers superior SR performance in low SNR conditions compared to SOFI, reducing sample damage, or enables longer measurements. The model, pre-trained and further trained on six sets of real fixed-cell microtubule data, showed similar performance in Pearson correlation, spatial resolution, and minimal background artifacts as with synthetic data. However, sparser fluorophore distributions in half of the dataset led to filament over-approximation. To improve this, future datasets should focus on dense fluorophore distributions across varying SNR conditions, allowing for lower illumination and reduced sample damage.

When trained on real fixed-cell mitochondria data and tested in a motion-controlled environment, the model successfully generated SR images in dynamic settings, which SOFI was not able to do. Using a rolling window, it produced SR images for each frame, capturing dynamic mitochondrial changes. Optimization with PyTorch TensorRT enabled real-time temporal resolutions of up to 4.85 fps, demonstrating potential for real-time live-cell imaging.

A drawback emerged: the model appears sensitive to background artifacts. In the first experiment, SOFI target images with more background artifacts led to poorer generalization compared to those without artifacts. This phenomenon was also observed in the mitochondria experiment, where, without background subtraction, SOFI images contained more artifacts, resulting in no SR improvement. This is problematic because clean SOFI images are not always easily obtainable, and the background subtraction pre-processing step was less than ideal, producing in what appears averaged structures.

Future work should involve live-cell experiments to validate structural changes against published results and using SMLM as ground truth to confirm the model's SR reconstruction. Exploring transfer learning effects and revisiting the initial experiment with only dense fluorophores could improve temporal resolution. Switching the network to a many-to-many configuration would enable temporal resolution to a single frame. Exploring alternative SOFI linearization methods that produce fewer background artifacts is also important. Additionally, using higher-order SOFI could enhance spatial resolution.

Our proposed method demonstrates that, with just 20 frames, we can reconstruct SR images with minimal background artifacts, achieving a two-fold improvement in near-theoretical spatial resolution. By leveraging pre-trained weights from synthetic data, the model can be trained on real fixed-cell microscopy data using four distinct measurements of the same cell type, using very dense fluorophores, for it to be used in a live-cell experiment. It makes a practical method for real-time live-cell imaging with temporal resolutions of up to 4.85 fps—surpassing state-of-the-art single-molecule-based SR methods.

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] 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.

[10] Girsault, A. *et al.* SOFI simulation tool: A software package for simulating and testing super-resolution optical fluctuation imaging. *PLOS ONE* **11**, e0161602 (2016). URL https://dx.plos.org/10.1371/journal.pone.0161602.

[11] Stergiopoulou, V. Learning and optimization for 3d super-resolution in fluorescence microscopy (2023). URL https://theses.hal.science/tel-04089027.

[12] 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.

[13] Christensen, C. N., Lu, M., Ward, E. N., Lio, P. & Kaminski, C. F. Spatio-temporal vision transformer for super-resolution microscopy. URL http://arxiv.org/abs/2203.00030. 2203.00030[physics].

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Chen, R. *et al.* Single-frame deep-learning super-resolution microscopy for intracellular dynamics imaging **14**, 2854. URL https://www.nature.com/articles/s41467-023-38452-2.

[19] 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.

[20] 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.

[21] 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.

[22] 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).

[23] Shariff, A., Murphy, R. F. & Rohde, G. K. A generative model of microtubule distributions, and indirect estimation of its parameters from fluorescence microscopy images. *Cytometry Part A* **77A**, 457–466 (2010). URL https://onlinelibrary.wiley.com/doi/10.1002/cyto.a.20854.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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).

[29] Miytek, M. Preprocess_sofi. https://github.com/GrussmayerLab/PreProcess_SOFI (2024).

[30] Lefebvre, A. E. Y. T., Ma, D., Kessenbrock, K., Lawson, D. A. & Digman, M. A. Automated segmentation and tracking of mitochondria in live-cell time-lapse images **18**, 1091–1102. URL https://www.nature.com/articles/s41592-021-01234-z.

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

[32] 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.

[33] 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).

Methods

3.1 Training Scheme

The Adam optimizer is employed with a learning rate (η) of 0.001 and beta values (β_1, β_2) set to 0.9 and 0.999, respectively. Model weights are initialized with Xavier Initialization[32] because of the sigmoid activation functions in our model. Additionally, an early stopping procedure with 400 epochs and a patience of 10 is implemented to prevent overfitting. A batch size of 10 is utilized based on memory constraints and computational efficiency. Training and testing is conducted on an NVIDA GeForce RTX 3090 of 24 GB of memory.

3.2 Loss Function

The high frequency (HF) content above the Nyquist-frequency η_c must be recovered from a set of low-resolution frames $l_i \in \mathbb{R}^{1\times H\times W}$ to reconstruct the high-resolution SOFI image $\hat{Y} \in \mathbb{R}^{1\times \gamma H\times \gamma W}$. Unlike the spatial domain, where these missing frequency cannot be fully separated, they can be in the Fourier domain. Therefore, we opted to the loss function proposed in the works of [33], which is defined as follows:

$$L_{\mathcal{F}} = L_{\mathcal{F}_A} + L_{\mathcal{F}_\angle} \tag{1a}$$

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

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

Here, both SR images $\hat{Y}$ and Y are transformed into the Fourier space by applying fast Fourier transform (FFT), where the absolute amplitude difference $L_{\mathcal{F}_A}$ 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 Microtubules

Obtained with Total Internal Reflection Fluorescence (TIRF) microscopy and DNA-PAINT (Points Accumulation for Imaging in Nanoscale Topography) which are advanced imaging techniques. TIRF selectively illuminates molecules near the glass surface, reducing background noise, while DNA-PAINT uses transient binding of dye-labeled DNA probes for high-resolution imaging. Together, they enhance signal-to-noise ratio (SNR), providing sharper and more accurate visualization at the nanoscale.

3.4 Mitochondria

Obtained with widefield microscopy of COS-7 cells, using DNA-PAINT probes targeting TOMM20, reveals detailed mitochondrial structures. COS-7 cells, derived from monkey kidney tissue, are widely used in research, while TOMM20 is a key protein in mitochondrial protein transport. Although DNA-PAINT enhances resolution, widefield imaging results in lower SNR compared to TIRF but reduces phototoxicity, minimizing sample damage.

3.5 Real-time temporal resolution

To compute the real-time temporal resolution depends on the both the camera frame rate and the latency of the model. The formula is:

$$R_{real-time} = \frac{1}{\frac{W}{N} + T_{latency}} \tag{2}$$

Where W is the number of input frames used by the model, N is the frame rate of the camera, and $T_{latency}$ is the models latency to compute a SR image.

Code availability

Code is available online at https://github.com/GrussmayerLab/SOFI-MISRGRU

Acknowledgements

M. Tekpinar provided guidance and assistance for the simulations, fixed-cell microtubule data were acquired by K. Zwaan, while both static and motion-controlled fixed-cell mitochondria data were acquired from R. Huo. The samples were kindly prepared by N. van Vliet (Deparment of Bionanoscience and Kavli Institute of Nanoscience Delft, Delft University of Technology).

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. ISSN: 0955-0674. https://www.sciencedirect.com/science/article/pii/S0955067423001205 (2023).
2. Hell, S. W. & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. en. *Optics Letters* **19,** 780. ISSN: 0146-9592, 1539-4794. https://opg.optica.org/abstract.cfm?URI=ol-19-11-780 (2024) (June 1994).
3. Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy: SHORT COMMUNICATION. en. *Journal of Microscopy* **198,** 82–87. ISSN: 0022-2720, 1365-2818. https://onlinelibrary.wiley.com/doi/10.1046/j.1365-2818.2000.00710.x (2024) (May 2000).
4. Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. en. *Science* **313,** 1642–1645. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.1127344 (2024) (Sept. 2006).
5. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). en. *Nature Methods* **3,** 793–796. ISSN: 1548-7091, 1548-7105. https://www.nature.com/articles/nmeth929 (2024) (Oct. 2006).
6. Zheng, X. *et al.* Current challenges and solutions of super-resolution structured illumination microscopy. *APL Photonics* **6,** 020901. ISSN: 2378-0967. eprint: https://pubs.aip.org/aip/app/article-pdf/doi/10.1063/5.0038065/20020143/020901_1_5.0038065.pdf. https://doi.org/10.1063/5.0038065 (Feb. 2021).
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. ISSN: 0021-9533. eprint: https://journals.biologists.com/jcs/article-pdf/133/11/jcs240713/3509892/jcs240713.pdf. https://doi.org/10.1242/jcs.240713 (June 2020).
8. Dertinger, T., Colyer, R., Iyer, G., Weiss, S. & Enderlein, J. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). en. *Proceedings of the National Academy of Sciences* **106,** 22287–22292. ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.0907866106 (2024) (Dec. 2009).
9. Gustafsson, N. *et al.* Fast live-cell conventional fluorophore nanoscopy with ImageJ through super-resolution radial fluctuations. en. *Nature Communications* **7,** 12471. ISSN: 2041-1723. https://www.nature.com/articles/ncomms12471 (2024) (Aug. 2016).
10. Girsault, A. *et al.* SOFI Simulation Tool: A Software Package for Simulating and Testing Super-Resolution Optical Fluctuation Imaging. *PLOS ONE* **11** (ed Degtyar, V. E.) e0161602. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0161602 (2024) (Sept. 1, 2016).
11. Stergiopoulou, V. *Learning and optimization for 3D super-resolution in fluorescence microscopy* May 2023. https://theses.hal.science/tel-04089027.
12. Herman, B. *Fluorescence microscopy* (BIOS Scientific Publ., 1998).
13. Mudry, E. *et al.* Structured illumination microscopy using unknown speckle patterns. en. *Nature Photonics* **6,** 312–315. ISSN: 1749-4885, 1749-4893. https://www.nature.com/articles/nphoton.2012.83 (2024) (May 2012).
14. Idier, J. *et al.* On the Superresolution Capacity of Imagers Using Unknown Speckle Illuminations. *IEEE Transactions on Computational Imaging* **4,** 87–98. ISSN: 2333-9403, 2334-0118. http://ieeexplore.ieee.org/document/8100885/ (2024) (Mar. 2018).
15. Li, H. & Vaughan, J. C. Switchable Fluorophores for Single-Molecule Localization Microscopy. en. *Chemical Reviews* **118,** 9412–9454. ISSN: 0009-2665, 1520-6890. https://pubs.acs.org/doi/10.1021/acs.chemrev.7b00767 (2024) (Sept. 2018).

16. Dertinger, T., Colyer, R., Vogel, R., Enderlein, J. & Weiss, S. Achieving increased resolution and more pixels with Superresolution Optical Fluctuation Imaging (SOFI). en. *Optics Express* **18,** 18875. ISSN: 1094-4087. https://opg.optica.org/oe/abstract.cfm?uri=oe-18-18-18875 (2024) (Aug. 2010).

17. Geissbuehler, S. *et al.* Mapping molecular statistics with balanced super-resolution optical fluctuation imaging (bSOFI). *Optical Nanoscopy* **1,** 4. ISSN: 2192-2853. http://optnano.springeropen.com/articles/10.1186/2192-2853-1-4 (2024) (2012).

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

19. Theodoridis, S. & Koutroumbas, K. *Pattern Recognition 4th Edition* (Academic Press, 2009).

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

21. Zhou, P. *et al. Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning* 2021. arXiv: 2010.05627 [cs.LG].

22. Cahuantzi, R., Chen, X. & Güttel, S. in *Intelligent Computing* 771–785 (Springer Nature Switzerland, 2023). ISBN: 9783031379635. http://dx.doi.org/10.1007/978-3-031-37963-5_53.

23. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Publisher: [object Object] Version Number: 1. https://arxiv.org/abs/1505.04597 (2024) (2015).

24. 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. ISSN: 0143-8166. https://www.sciencedirect.com/science/article/pii/S014381662300369X (2023) (Jan. 2024).

25. Qu, L. *et al. Self-inspired learning to denoise for live-cell super-resolution microscopy* Jan. 23, 2024. http://biorxiv.org/lookup/doi/10.1101/2024.01.23.576521 (2024).

26. 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. ISSN: 2515-7647. https://iopscience.iop.org/article/10.1088/2515-7647/ac3838 (2024) (Jan. 2022).

27. Lucy, L. B. An iterative technique for the rectification of observed distributions. *The Astronomical Journal* **79,** 745. ISSN: 00046256. http://adsabs.harvard.edu/cgi-bin/bib_query?1974AJ....79..745L (2024) (June 1974).

28. Geissbuehler, S., Dellagiacoma, C. & Lasser, T. Comparison between SOFI and STORM. *Biomedical Optics Express* **2,** 408. ISSN: 2156-7085. https://opg.optica.org/boe/abstract.cfm?uri=boe-2-3-408 (2024) (Mar. 1, 2011).

29. Geissbuehler, M. & Lasser, T. How to display data by color schemes compatible with red-green color perception deficiencies. *Optics Express* **21,** 9862. ISSN: 1094-4087. https://opg.optica.org/oe/abstract.cfm?uri=oe-21-8-9862 (2024) (Apr. 22, 2013).

30. Rifat Arefin, M. *et al. Multi-Image Super-Resolution for Remote Sensing using Deep Recurrent Networks* in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (IEEE, Seattle, WA, USA, June 2020), 816–825. ISBN: 978-1-72819-360-1. https://ieeexplore.ieee.org/document/9150720/ (2024).

31. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. Publisher: [object Object] Version Number: 1. https://arxiv.org/abs/1512.03385 (2024) (2015).

32. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. Publisher: [object Object] Version Number: 1. https://arxiv.org/abs/1603.08155 (2024) (2016).

33. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Publisher: [object Object] Version Number: 1. https://arxiv.org/abs/1502.01852 (2024) (2015).

34. Ballas, N., Yao, L., Pal, C. & Courville, A. Delving Deeper into Convolutional Networks for Learning Video Representations. Publisher: [object Object] Version Number: 4. https://arxiv.org/abs/1511.06432 (2024) (2015).
35. Zeiler, M. D., Krishnan, D., Taylor, G. W. & Fergus, R. *Deconvolutional networks* in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, San Francisco, CA, USA, June 2010), 2528–2535. ISBN: 978-1-4244-6984-0. http://ieeexplore.ieee.org/document/5539957/ (2024).
36. Wang, Z., Chen, J. & Hoi, S. C. H. Deep Learning for Image Super-Resolution: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43,** 3365–3387. ISSN: 0162-8828, 2160-9292, 1939-3539. https://ieeexplore.ieee.org/document/9044873/ (2024) (Oct. 1, 2021).
37. Fuoli, D., Gool, L. V. & Timofte, R. *Fourier Space Losses for Efficient Perceptual Image Super-Resolution* 2021. arXiv: 2106.00783 [eess.IV]. https://arxiv.org/abs/2106.00783.
38. Zhang, Z., Wang, Y., Piestun, R. & Huang, Z.-l. Characterizing and correcting camera noise in back-illuminated sCMOS cameras. *Opt. Express* **29,** 6668–6690. https://opg.optica.org/oe/abstract.cfm?URI=oe-29-5-6668 (Mar. 2021).
39. Diekmann, R. *et al.* Photon-free (s)CMOS camera characterization for artifact reduction in high- and super-resolution microscopy. *Nature Communications* **13,** 3362. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-30907-2 (2024) (June 11, 2022).
40. Saguy, A. *et al.* DBlink: dynamic localization microscopy in super spatiotemporal resolution via deep learning. *Nature Methods* **20,** 1939–1948. ISSN: 1548-7091, 1548-7105. https://www.nature.com/articles/s41592-023-01966-0 (2024) (Dec. 2023).
41. Miytek, M. *PreProcess_SOFI* https://github.com/GrussmayerLab/PreProcess_SOFI. 2024.
42. Christensen, C. N., Lu, M., Ward, E. N., Lio, P. & Kaminski, C. F. *Spatio-temporal Vision Transformer for Super-resolution Microscopy* Feb. 28, 2022. arXiv: 2203.00030[physics]. http://arxiv.org/abs/2203.00030 (2024).
43. Li, J., Tong, G., Pan, Y. & Yu, Y. Spatial and temporal super-resolution for fluorescence microscopy by a recurrent neural network. *Optics Express* **29,** 15747. ISSN: 1094-4087. https://opg.optica.org/abstract.cfm?URI=oe-29-10-15747 (2024) (May 10, 2021).
44. Li, C. & Lee, C. Minimum cross entropy thresholding. *Pattern Recognition* **26,** 617–625. ISSN: 00313203. https://linkinghub.elsevier.com/retrieve/pii/003132039390115D (2024) (Apr. 1993).
45. Li, C. & Tam, P. An iterative algorithm for minimum cross entropy thresholding. *Pattern Recognition Letters* **19,** 771–776. ISSN: 01678655. https://linkinghub.elsevier.com/retrieve/pii/S0167865598000579 (2024) (June 1998).
46. Culley, S. *et al.* Author correction: Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods* **17,** 1167–1167 (2020).
47. Descloux, A., Grußmayer, K. S. & Radenovic, A. Parameter-free image resolution estimation based on decorrelation analysis. *Nature Methods* **16,** 918–924. ISSN: 1548-7091, 1548-7105. https://www.nature.com/articles/s41592-019-0515-7 (2024) (Sept. 2019).
48. Shariff, A., Murphy, R. F. & Rohde, G. K. A generative model of microtubule distributions, and indirect estimation of its parameters from fluorescence microscopy images. *Cytometry Part A* **77A,** 457–466. ISSN: 1552-4922, 1552-4930. https://onlinelibrary.wiley.com/doi/10.1002/cyto.a.20854 (2024) (May 2010).
49. Glorot, X. & Bengio, Y. *Understanding the difficulty of training deep feedforward neural networks* in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (eds Teh, Y. W. & Titterington, M.) **9** (PMLR, Chia Laguna Resort, Sardinia, Italy, May 13, 2010), 249–256. https://proceedings.mlr.press/v9/glorot10a.html.
50. Jin, L. *et al.* Deep learning enables structured illumination microscopy with low light levels and enhanced speed. *Nature Communications* **11,** 1934. ISSN: 2041-1723. https://www.nature.com/articles/s41467-020-15784-x (2024) (Apr. 22, 2020).

51. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nature Biotechnology* **36,** 460–468. ISSN: 1087-0156, 1546-1696. https://www.nature.com/articles/nbt.4106 (2024) (May 2018).

52. Speiser, A. *et al.* Deep learning enables fast and dense single-molecule localization with high accuracy. *Nature Methods* **18,** 1082–1090. ISSN: 1548-7091, 1548-7105. https://www.nature.com/articles/s41592-021-01236-x (2024) (Sept. 2021).

A

SOFI Architecture Parts

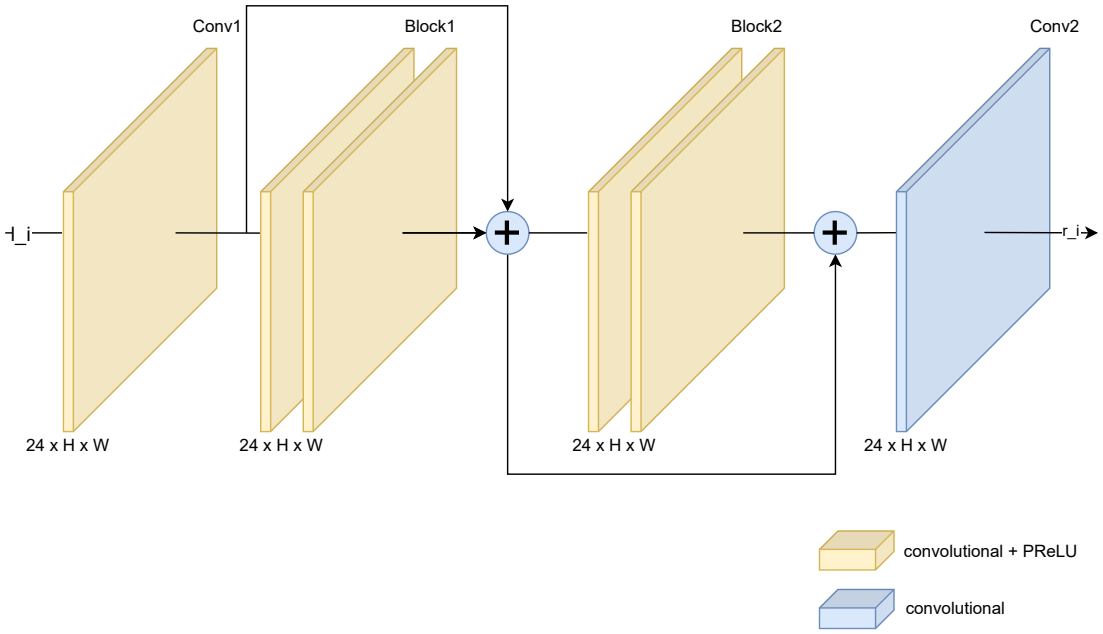


Figure A.1: The encoder stage, where the number of encoders depends on the number of frames used in the architecture, resulting in N encoders. Furthermore, each convolution layer consists of a 3×3 kernel with a stride of 1 with 24 filters producing 24 feature maps for one input frame l_i.

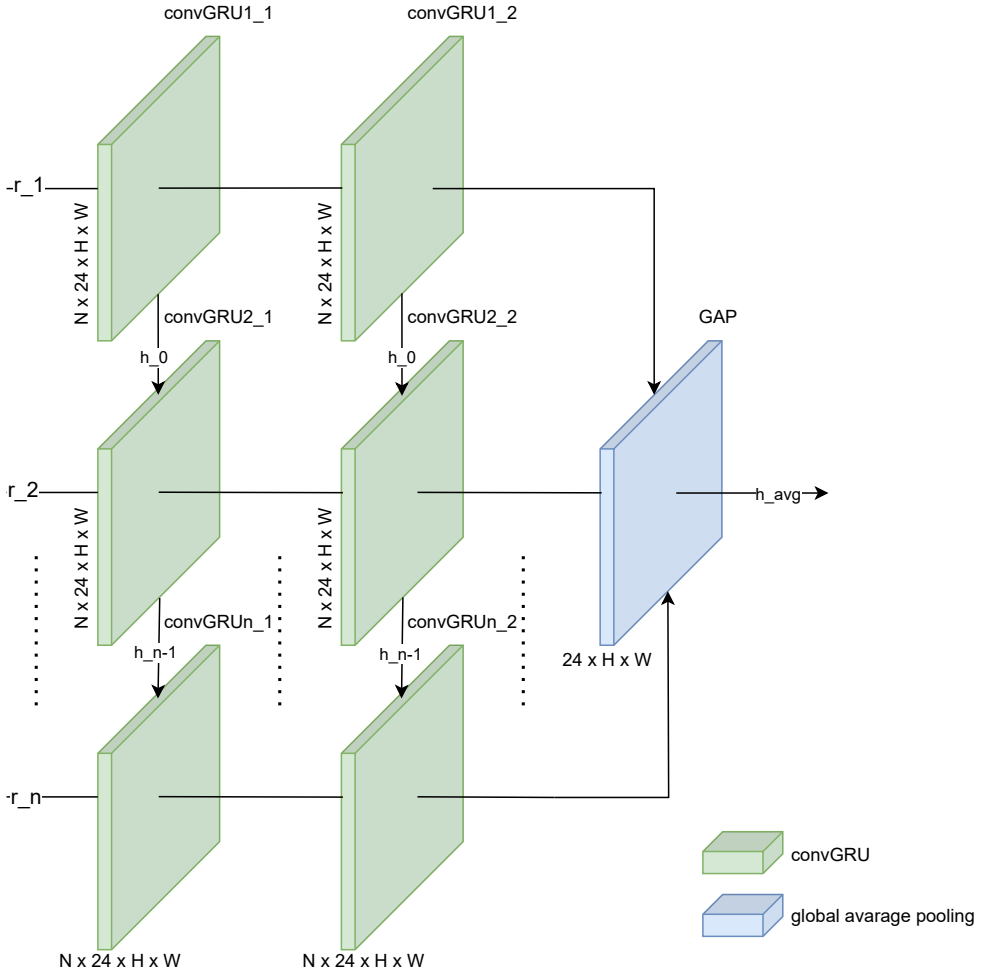


Figure A.2: The fusion stage consisting of the convGRU architecture stacked upon each other to process the sequential information between the latent representations r_i. After which, global average pooling is used on the first dimension to return $h_{avg} \in \mathrm{R}^{C_{GRU} \times H \times W}$, which in our experiment C_{GRU} is set to 24.

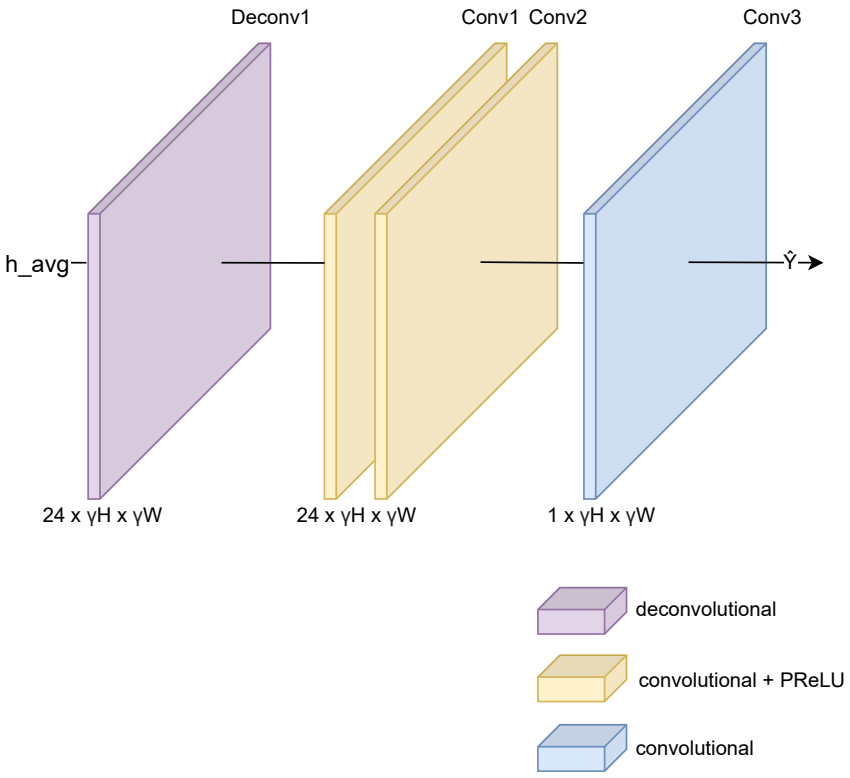


Figure A.3: In the decoder stage, the h_{avg} feature map is up-sampled to the size of the SOFI image using a deconvolutional layer. Following this, it undergoes processing by two convolutional layers with PReLU activation functions for minor adjustments. Finally, the feature maps are projected into the predicted SOFI image $\hat{Y} \in \mathrm{R}^{1\times H\times W}$.

B

Reconstruction Quality Quantification

Spatial resolution alone does not necessarily indicate a high-quality super-resolution (SR) reconstruction. Several metrics are used to assess the quality of SR reconstructions, such as Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM), and Peak Signal to Noise Ratio (PSNR)[42] [43]. However, as noted by [40], these metrics don't effectively describe the reconstruction in this context. Some FOVs have more the background pixels, often values close to zero, compared to the signal, leading to an overly optimistic assessment when evaluating matrices dominated by near-zero values. Additionally, these metrics don't fully communicate which parts of the prediction image are incorrect. Ideally, one wants to measure if the model is missing structures or if there are any artifacts in the SR reconstruction. In [40], they tried to resolve this by binarizing the predication with a threshold. From there, one can calculate the True Positive Rate (TPR) and the True Negative Rate (TNR). However, using the same threshold value across different SR images is less than ideal in our case as different signal-to-noise (SNR) conditions yield different intensity rates in the super-resolved images. If set too high, structures are considered background; if too low, backgrounds with a relative higher intensity value become a signal. This can even be problematic for ground truth images, as the background can have relatively higher intensity values if it has denser structures.

Ideally, an algorithm should find the most optimal threshold value to separate the background from the signal, independent of the intensity rates in the super-resolved images. One such algorithm is minimum cross-entropy thresholding [44], also known as Li thresholding. Specifically, the iterative algorithm proposed in [45] is used, which has been shown to effectively binarize ground truth, SOFI-based, and prediction-based SR images. It minimizes the cross-entropy between the foreground and its mean, as well as between the background and its mean, to obtain the optimal threshold value. This is provided there are no more than two peaks in the histogram, which can cause the iterative procedure to get stuck in a local optimum. From there, we can calculate the TPR, TNR, and the confusion matrices, where the TPR and TNR are calculated as:

$$TPR = \frac{TP}{TP + FN} \tag{B.1a}$$

$$TNR = \frac{TN}{TN + FP} \tag{B.1b}$$

Additionally, we use Pearson correlation to evaluate the relationship between our predicted image and the ground truth (GT). When no GT is available, the SQUIRREL algorithm [46] is employed to measure the correlation between a standard deviation (STD) widefield image and the predicted image. Finally, decorrelation analysis [47] is used to assess spatial resolution. First, we will discuss how we binarized the images with the results, followed by an explanation of Pearson correlation, the SQUIRREL algorithm, and decorrelation analysis.

B.1. Theory Li Thresholding

Li thresholding minimizes the cross-entropy between the foreground and its mean, as well as between the background and its mean. These means are defined by the zeroth and first moments of the foreground and background portions of the thresholded histogram, where the histogram h is defined on the gray level range $[1, L]$. The moments of the foreground and background are defined as follows:

$$\begin{aligned} m_{0_\mathsf{a}}(t) &= \sum_{i=1}^{t-1} h(i), \qquad & m_{0_\mathsf{b}}(t) &= \sum_{i=t}^{L} h(i), \\ m_{1_\mathsf{a}}(t) &= \sum_{i=1}^{t-1} i \cdot h(i), \qquad & m_{1_\mathsf{b}}(t) &= \sum_{i=t}^{L} i \cdot h(i) \end{aligned} \tag{B.2}$$

The means are defined as:

$$\mu_\mathsf{a}(t) = \frac{m_{1_\mathsf{a}}(t)}{m_{0_\mathsf{a}}(t)}, \qquad \mu_\mathsf{b}(t) = \frac{m_{1_\mathsf{b}}(t)}{m_{0_\mathsf{b}}(t)} \tag{B.3}$$

with a and b being the background and foreground, respectively. The minimum cross-entropy is then defined as:

$$\eta(t) = -m_{1_\mathsf{a}} \cdot \log\left(\mu_\mathsf{a}(t)\right) - m_{1_\mathsf{b}} \cdot \log\left(\mu_\mathsf{b}(t)\right) \tag{B.4}$$

and the optimal threshold t_op is given by

$$t_\mathsf{op} = \arg\min_t \ \eta(t) \tag{B.5}$$

Originally, this involved calculating all possible threshold values of t [44], which has been replaced by the numerical method introduced in[45] by taking the derivative of $\eta(t)$ and setting it to zero. After simplification and assuming $h(t) \neq 0$, we get:

$$t = \frac{\mu_\mathsf{b}(t) - \mu_\mathsf{a}(t)}{\log\left(\mu_\mathsf{b}(t)\right) - \log\left(\mu_\mathsf{a}(t)\right)} \tag{B.6}$$

After applying the one-point iteration method to equation B.6, we obtain the following procedure to find the optimal threshold:

$$\begin{aligned} &t_{n+1} = \text{round}\left\{\frac{\mu_\mathsf{b}(t) - \mu_\mathsf{a}(t)}{\log\left(\mu_\mathsf{b}(t)\right) - \log\left(\mu_\mathsf{a}(t)\right)}\right\}, \\ &n \geqslant 0 \end{aligned} \tag{B.7}$$

where t_0 is initialized with the mean value of the image in our case. This iterative procedure continues until convergence, which occurs when $t_{n+1} = t_n$. The round(x) rounds x to the nearest integer. Figure B.1 illustrates an example of the minimization of the cross-entropy to obtain the optimal threshold of a second-order SOFI image.

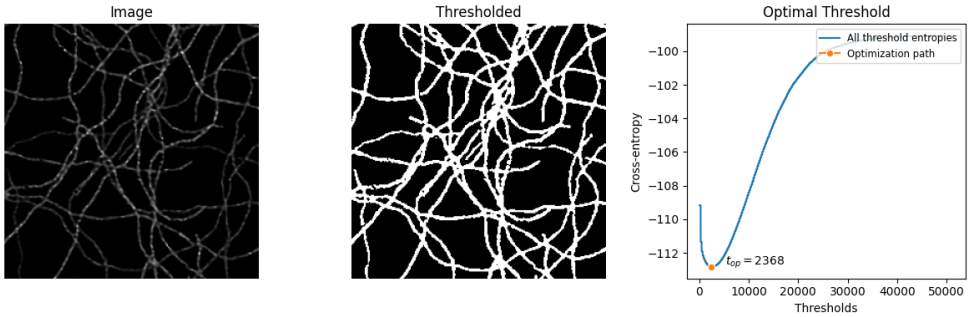


Figure B.1: Example of the minimization of the cross-entropy to obtain the optimal threshold of a 16 bit second-order SOFI image. The first column represents the second-order SOFI image, the second column shows the binarized image, and the last column illustrates the cross-entropy loss.

B.1.1. Results Binarization

Figure B.2 shows the binarization results using Li thresholding for ground truth, SOFI-based (100 frames), and prediction-based (20 frames) SR images under different SNR conditions. The SR images are effectively binarized for all SNR conditions. However, not all structures are considered as signal; for instance, SOFI binarization optimistically excludes most background artifacts. Binarization is a complex task, and no method is perfect. While Li thresholding is effective in this instance, comparisons between the ground truth and super-resolved bit masks should not expect exact matches due to imperfections in binarization. Rather, the focus should be on comparing the relative performance of SOFI-based and model-based SR reconstruction methods. In figure B.3 the confusion matrices can be found.

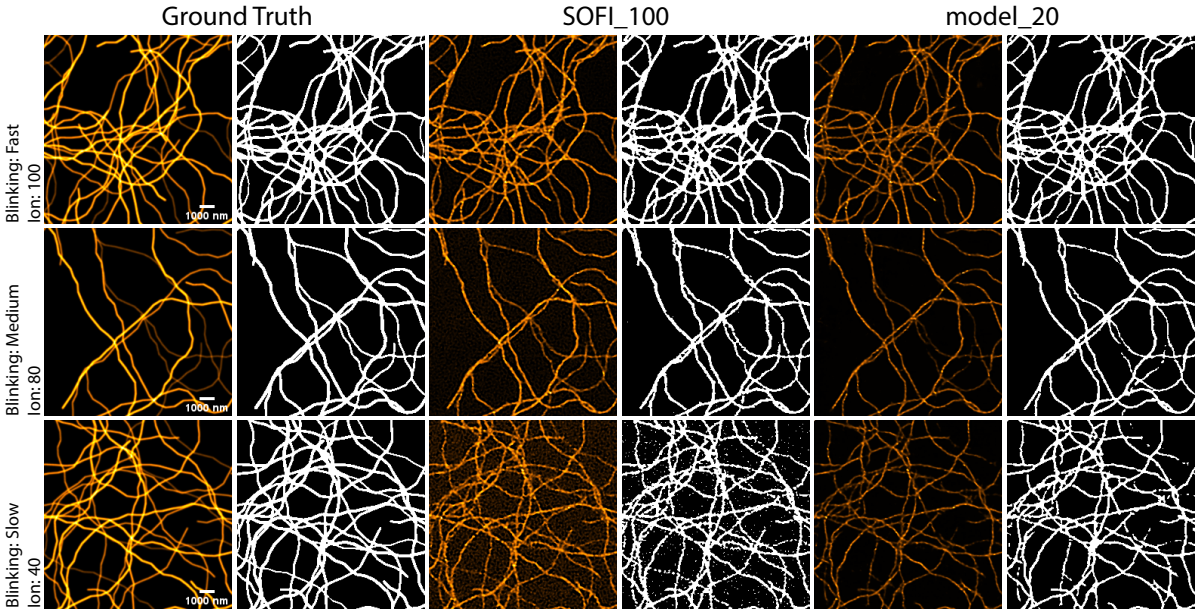


Figure B.2: Binarization of the SR image using Li thresholding of the ground truth, SOFI based on 100 frames, and the model predictions based on 20 frames, given different SNR conditions.

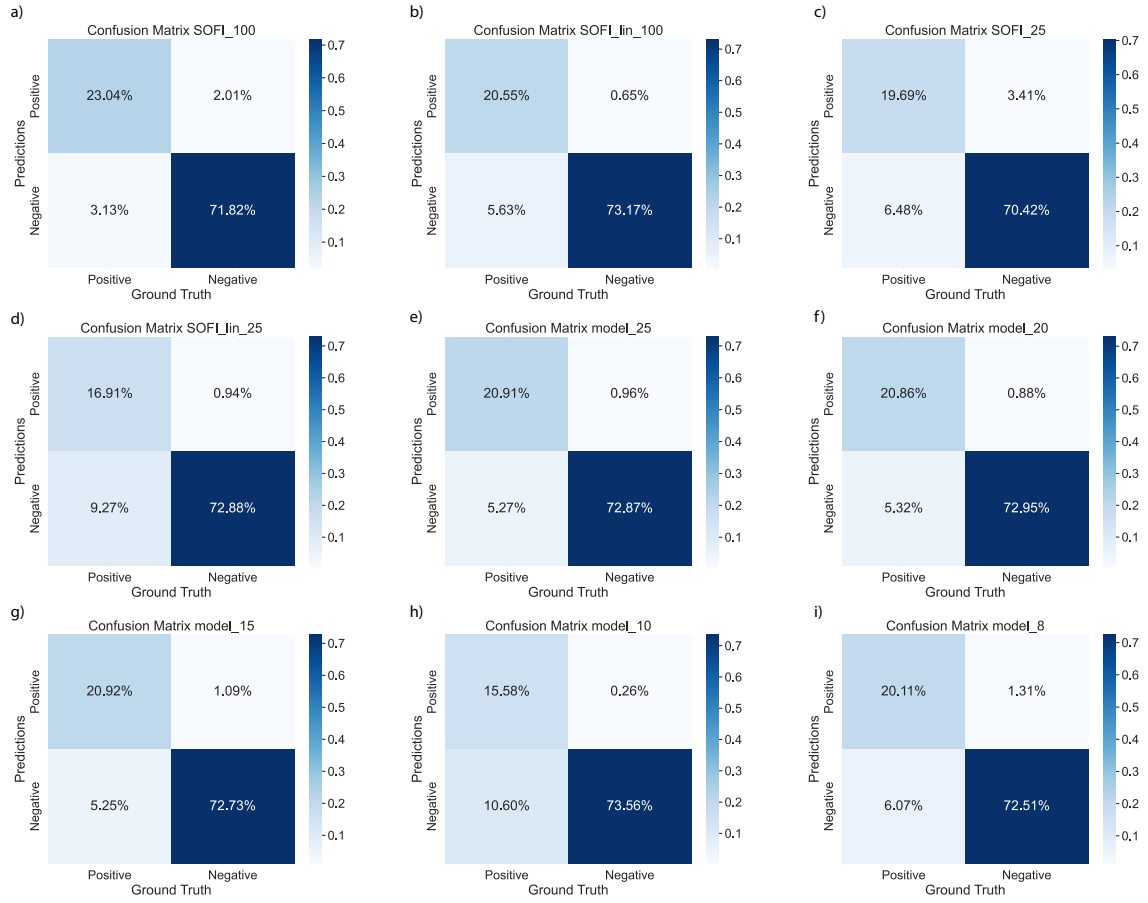


Figure B.3: (a-i) Respective confusion matrices of SOFI and model-based SR reconstructions.

B.2. Pearson Correlation

The Pearson correlation is the ratio of how much two variables change together (their covariance) to the product of their individual variability (their standard deviations). This normalization ensures that the correlation value always falls between −1 and 1, providing a standardized measure of their relationship. The 2D Pearson correlation is given as:

$$r = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(X_{ij}-\bar{X})(Y_{ij}-\bar{Y})}{\sqrt{\sum_{j=1}^{m}\sum_{i=1}^{n}(X_{ij}-\bar{X})^2}\sqrt{\sum_{j=1}^{m}\sum_{i=1}^{n}(Y_{ij}-\bar{Y})^2}} \tag{B.8}$$

Where X_i and Y_i are individual data points from variables X and Y, $\bar{X}$ and $\bar{Y}$ are the means of X and Y, n is the number of data points.

B.3. SQUIRREL

The SQUIRREL algorithm [46] uses two images: a super-resolution (SR) image and a reference image. The reference image is the raw image obtained from a microscope, which can either be the average widefield image or the standard deviation (STD) widefield image. The principle of the algorithm can be seen in figure B.4. Essentially, it estimates a point spread function (PSF) and convolves it with the SR image, assuming the PSF is uniform across all pixels. The algorithm iterates until the convolved SR image matches the reference image. This process allows us to further use Pearson correlation to assess the correlation between the SR image and the reference image.

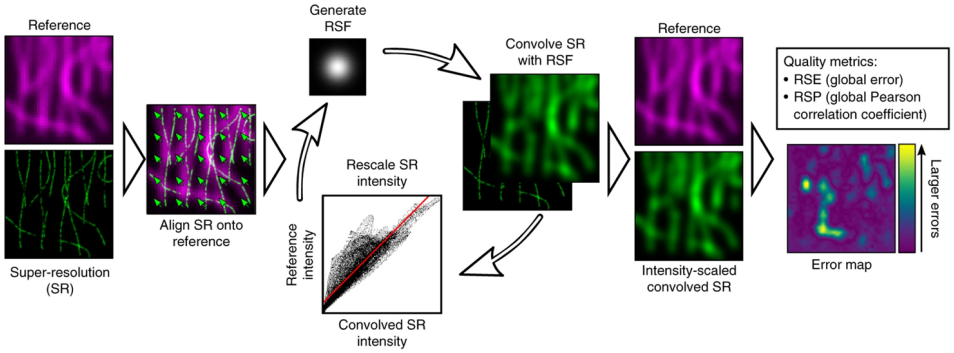


Figure B.4: Workflow of the SQUIRREL algorithm: iterating and estimating the PSF until the convolved SR image matches the reference image.

B.4. Decorrelation Analysis

The algorithm follows these steps: First, the Fourier transform of the image is normalized as $I_n(k) = \frac{I(k)}{|I(k)|}$. By performing a cross-correlation between the Fourier-transformed image and the input image $I(k)$, values between 0 and 1 are obtained. In the second step, the process is repeated, but this time the normalized Fourier-transformed image is further filtered using a binary circular mask of radius $r \in [0, 1]$. This operation is iterated, allowing the computation of $d(r)$:

$$d(r) = \frac{\int\int Re\left\{I(\boldsymbol{k})\, I_n^*(\boldsymbol{k})\, M(\boldsymbol{k}; r)\right\} dk_x dk_y}{\sqrt{\int\int |I(\boldsymbol{k})|^2\, dk_x dk_y \int\int |I_n(\boldsymbol{k})\, M(\boldsymbol{k}; r)|^2\, dk_x dk_y}} \tag{B.9}$$

where $k = [k_x, k_y]$ denotes Fourier space coordinates, $I(k)$ is the Fourier transform of the input image, $I_n(k)$ is the normalized Fourier transform, and $M(k; r)$ is the binary mask with radius r.

As the mask radius is reduced from 1 to 0, an attenuation peak appears at a specific radius r, indicating the highest correlation for that spatial frequency, as shown in figure B.5 b. However, this peak represents the highest spatial correlation for the entire image. To refine this, we apply a series of high-pass filters, ranging from weak to strong, to the input image. By repeating the procedure for each filtered image, we can identify the frequency with the highest correlation, the attenuation peak, until the curve flattens and no peak is observed. For each filtered image, a decorrelation function is calculated, and the peak position r_i and amplitude A_i are extracted, yielding a set of $[r_i, A_i]$ pairs (see figure A.3 c). The resolution is then compute as:

$$Resolution = \frac{2 \times pixel\,size}{k_c} \tag{B.10}$$

where k_c is the local maximum normalized highest frequency, obtained from:

$$k_c = \max\left[r_0, \ldots, r_{N_g}\right] \tag{B.11}$$

where r_{Ng} represents the peak position for the N_g-th high-pass filter.

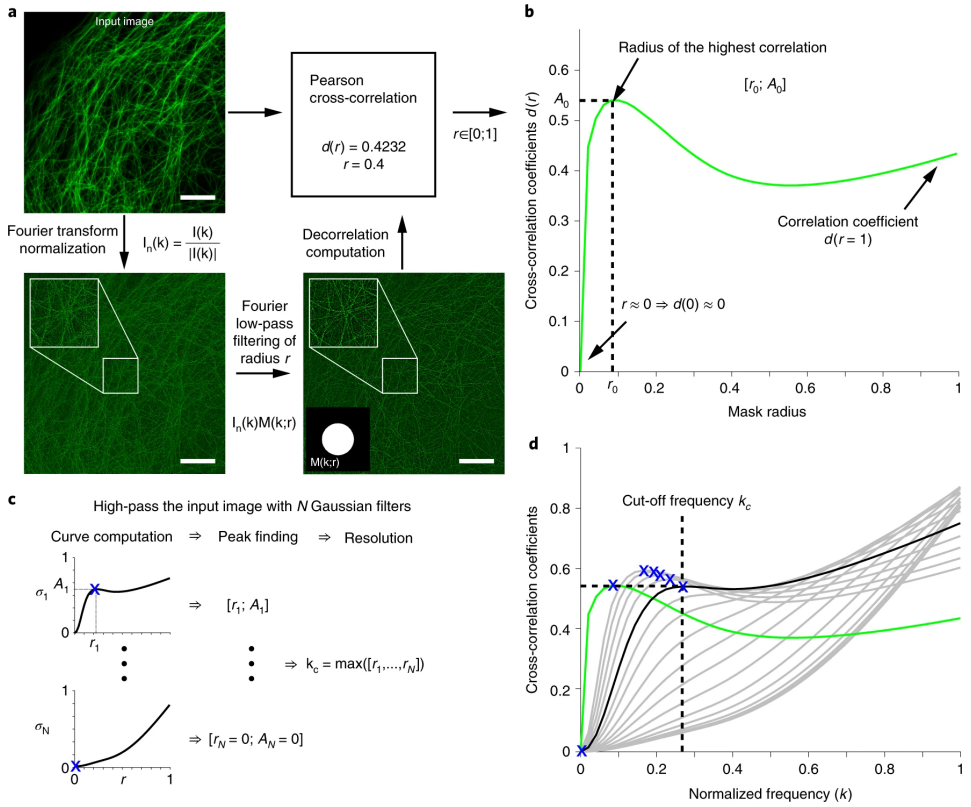


Figure B.5: a) Cross-correlation of the image with its normalized version after applying a Fourier filter. b) Cross-correlation coefficient plotted as a function of the mask radius. c) High-pass filtering applied to the input image, followed by resolution estimation. d) Plot showing all the decorrelation functions calculated for the image along with resolution estimation. The green line represents the decorrelation function without high-pass filtering, while the grey lines correspond to those with high-pass filtering. The blue crosses indicate local maxima, and the black line highlights the decorrelation function at the highest frequency peak. The vertical dashed line marks the cut-off frequency k_c. Scale bar: 5 µm.

C

Spatial Domain Loss

Spatial domain loss functions, such as the Mean Absolute Error (MAE or L1) and Mean Squared Error (MSE or L2), are popular in image SR reconstruction [36]. However, SR is closely associated with the frequency domain. Therefore, we opted for the loss function proposed in [37]. Here, we show the difference between the Fourier and spatial domain-based loss functions. Specifically, the L1 loss function is used, as the SR SOFI image sparsely contains outliers in the form of very bright pixels because of the nature of the non-linear response to the molecular brightness levels in the cumulant calculation[17], which can hinder the training process. Refer to the histogram depicted in figure C.1.

To assess the differences between the two loss functions, we trained the model on 20 frames using these loss functions. We used a synthetic dataset containing microtubules based on the physical model of [48]. The test set including a simulated ground truth. The training set comprises 2000 samples, while the evaluation and test sets each contain 480 samples.

During training, the Adam optimizer is employed with a learning rate (η) of 0.001 and beta values (β_1, β_2) set to 0.9 and 0.999, respectively. Model weights are initialized with Xavier Initialization[49] for guaranteed convergence, rather than random initialization as it tend to fail in our case. Additionally, an early stopping procedure with 400 epochs and a patience of 10 is implemented to prevent overfitting. A batch size of 10 is utilized based on memory constraints and computational efficiency. Training is conducted on an NVIDA GeForce RTX 3090 of 24 GB of memory.

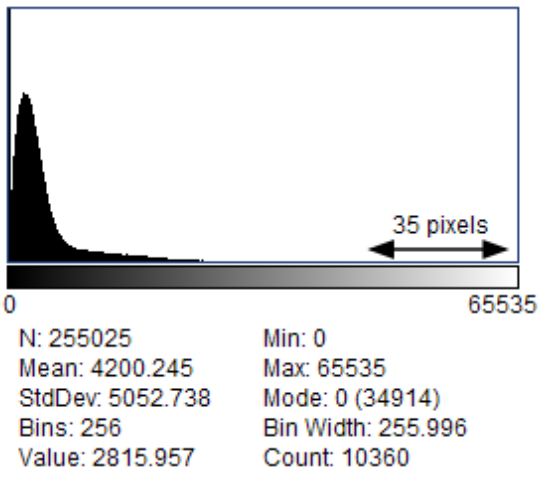


Figure C.1: Histogram of a SOFI image, showing that there are 35 pixels with relatively higher values compared to the rest, ranging between 43k and 65553.

C.1. Results

Result are depicted in figure C.2. Based on the Pearson coefficient with respect to the ground truth and RSP with respect to the STD widefield image using SQUIRREL, the L1-based loss function appears to perform better than the Fourier-based model. However, this is because the spatial domain loss is better at estimating the brightness levels of the pixels. These metrics consider this aspect as they reflect the correlations between the pixels of the predicted SR image and the corresponding ground truth. This is further confirmed by figure C.2 (d), where the Fourier-based model is better at predicting the true structures compared to the L1-based model. Hence, the L1-based model performs better in the TNR as shown in figure C.2 (e) because it generally predicts more background pixels as signal pixels, as depicted in the confusion matrices in figure C.3. Ultimately, the Fourier-based model outperforms the L1-based model as it is better at recovering the structures and achieves spatial resolution closer to the theoretical resolution, as depicted in figure C.2 (a).

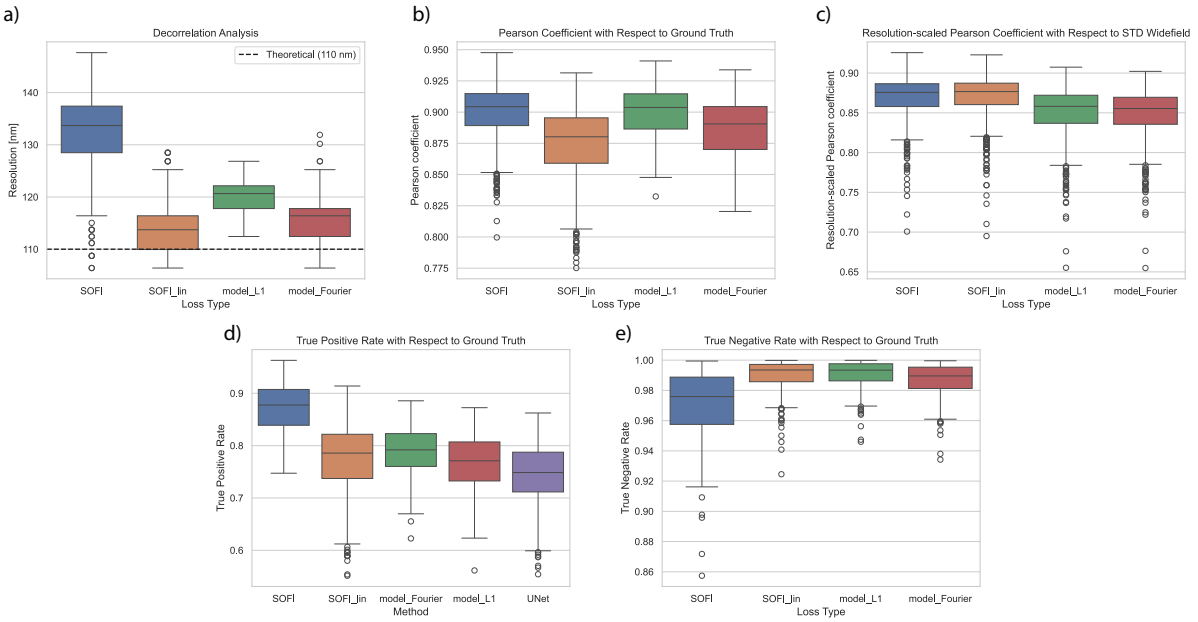


Figure C.2: (a-e) Results of the Fourier and L1-based loss functions for the model trained on 20 frames. Additionally, SOFI results based on 100 frames are provided for comparison. (a) Decorrelation analysis shows the Fourier-based loss reaching closer to theoretical spatial resolutions. (b) Pearson coefficient with respect to ground truth, with L1 scoring higher than the Fourier-based loss function. (c) RSP results showing L1 scoring higher. (d) TPR results showing Fourier relatively higher. (e) TNR results showing L1 scoring higher.

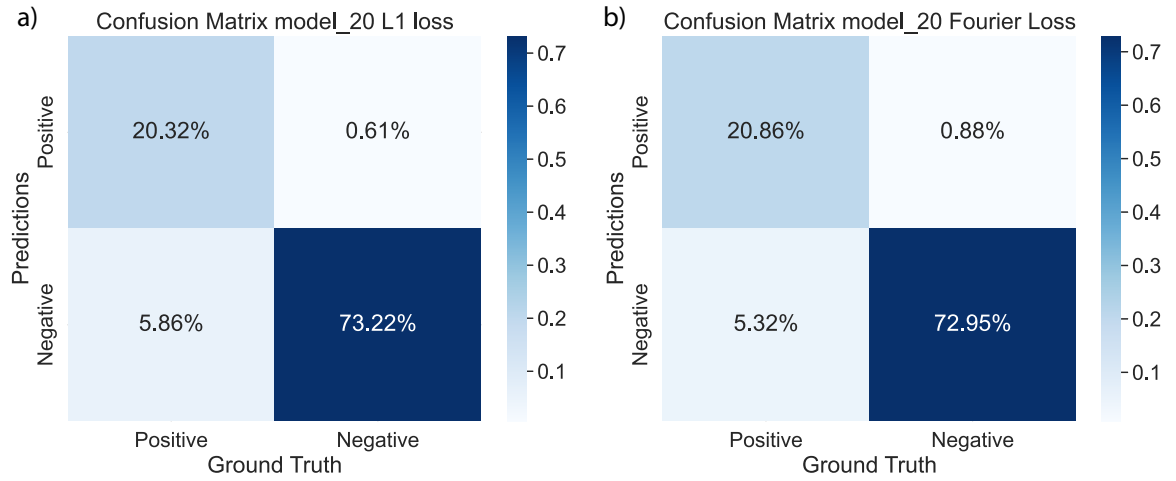


Figure C.3: (a-b) Confusion matrices of the model based on 20 frames trained with the L1 and Fourier-based loss functions. In the confusion matrices, it can be observed that the model based on the Fourier loss (b) is better at predicting the structures compared to the L1-based model (a), which predicts fewer structures and more background.

D

SOFI U-Net

The U-Net architecture is a popular choice in the field of SR microscopy[25, 50–52]. However, in the context of fluorescence microscopy, where there are correlated distributed blinking emitters over time[10], the U-Net architecture does not capitalize on this temporal information. Although it is used in the cumulant calculation of SOFI[10], this highlights the need for incorporating temporal information into the network. Here, we demonstrate this claim by comparing the results of our proposed model with the U-Net architecture, focusing primarily on spatial resolution, Pearson coefficient, RSP from SQUIRREL, and the TNR and TPR, while ignoring the computation time of the U-Net architecture.

To create a mapping from LR frames to a second-order SOFI image with U-Net architecture, the LR frames need to be up-sampled by 2 times, as the network only allows sizes by the power of 2^n[23]. To perform this up-sampling, spatial methods (such as bilinear interpolation) are commonly used[24]. However, these missing pixels are now based on the noisy-corrupted pixels of the LR frame. This can hinder the learning process and potentially produce background artifacts, as these new pixels do not conform to the randomness of noise, which is problematic in background areas where there is no fluorescent signal[25]. Hence, Fourier interpolation is used as it takes advantage of the finite support of the optical transfer function (OTF) of a microscope, which is the Fourier transform of the point-spread-function (PSF). By padding the Fourier-transformed image beyond its OTF support with zeros and then back-transforming it, the resulting image effectively doubles its pixels in both height and width. Afterwards, the SOFI images are padded with zeros to fit in the network due to their uneven sizes.

The U-Net architecture is trained based 20 frames using the L1 loss function. We used a synthetic dataset containing microtubules based on the physical model of [48], generated using SOFI simulation tools (see chapter: 5). The test set including a simulated ground truth. The training set comprises 2000 samples, while the evaluation and test sets each contain 480 samples.

During training, the Adam optimizer is employed with a learning rate (η) of 0.001 and beta values (β_1, β_2) set to 0.9 and 0.999, respectively. Additionally, an early stopping procedure with 400 epochs and a patience of 10 is implemented to prevent overfitting. A batch size of 10 is utilized based on memory constraints and computational efficiency. Training is conducted on an NVIDA GeForce RTX 3090 of 24 GB of memory.

D.1. Results

Results are depicted in figure D.1. From the decorrelation analysis, it can be observed that the U-Net architecture reaches closer theoretical spatial resolutions compared to our model based on the Fourier loss. This is likely due to the fact that our model has 131,510 parameters to train, whereas the U-Net architecture has 31,041,537 parameters for an input size of 20 frames. Generally, more complexity (more trainable parameters) results in better generalization given enough training data[20]. However, the Pearson correlations in figure D.1 (b-c) depict that the U-Net is underperforming compared to the

other models. This is due to the fact that the trained U-Net model is missing structures, which is further confirmed in figure D.1 (d) and the confusion matrix in figure D.2. These missing structures are visually depicted in figure D.3, where arrows indicate the areas of interest. Lastly, figure D.1 (e) shows that the TNR performs slightly worse than model based on the Fourier loss. Inspecting the confusion matrix in figure D.1 (c) for the U-Net more closely, it can be observed that it contains more artifacts, causing for a lower TNR performance. Additionally, depicted by the red arrows in figure D.3, the U-Net shows poorer distinction of the filaments, especially when they are close together, compared to the proposed model.

Ultimately, using temporal information demonstrates beneficial for better filament reconstruction in synthetic fixed-cell microtubule data across various blinking and density conditions. Although the U-Net model is more complex and therefore may generalize better, it lacks the temporal information needed for accurate filament structure prediction. In contrast, our proposed model, with significantly fewer trainable parameters, effectively leverages temporal information for improved performance.

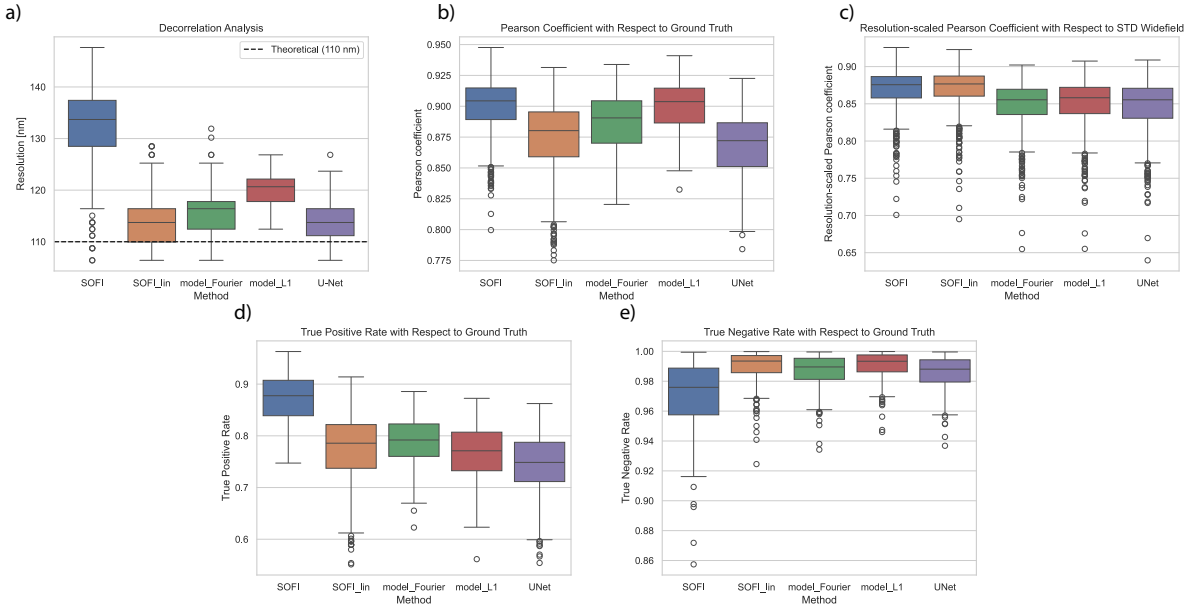


Figure D.1: (a-e) Results of the U-Net architecture trained on 20 frames compared with the model trained on the L1 and Fourier loss function based on 20 frames. Additionally, SOFI results based on 100 frames are provided for comparison. (a) Decorrelation analysis shows the U-Net architecture is reaching closer to theoretical spatial resolutions. (b) Pearson coefficient with respect to ground truth, with L1 scoring higher than the other models. (c) RSP results showing L1 scoring higher. (d) TPR results showing Fourier relatively higher. (e) TNR results showing L1 scoring higher.

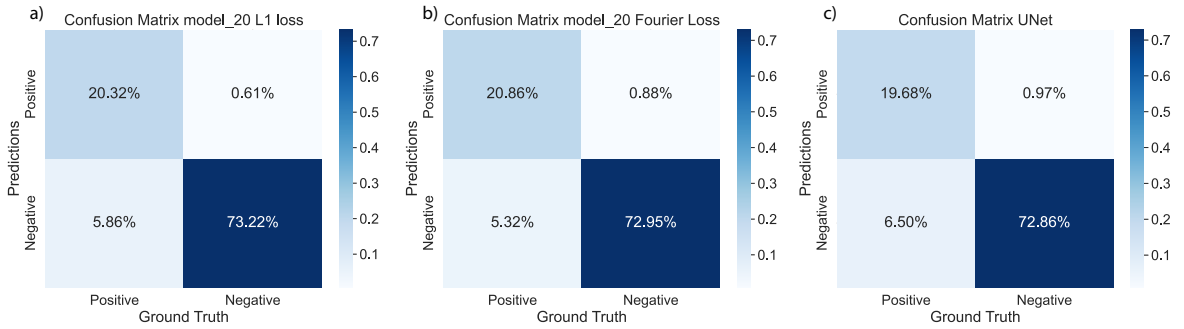


Figure D.2: (a-c) Confusion matrices of the models trained with the L1 and Fourier-based loss functions and the U-Net architecture, all based on 20 frames. In (c), it can be observed that the U-Net architecture contains more artifacts and missing structures in contrast with the models based on Fourier and L1 loss.

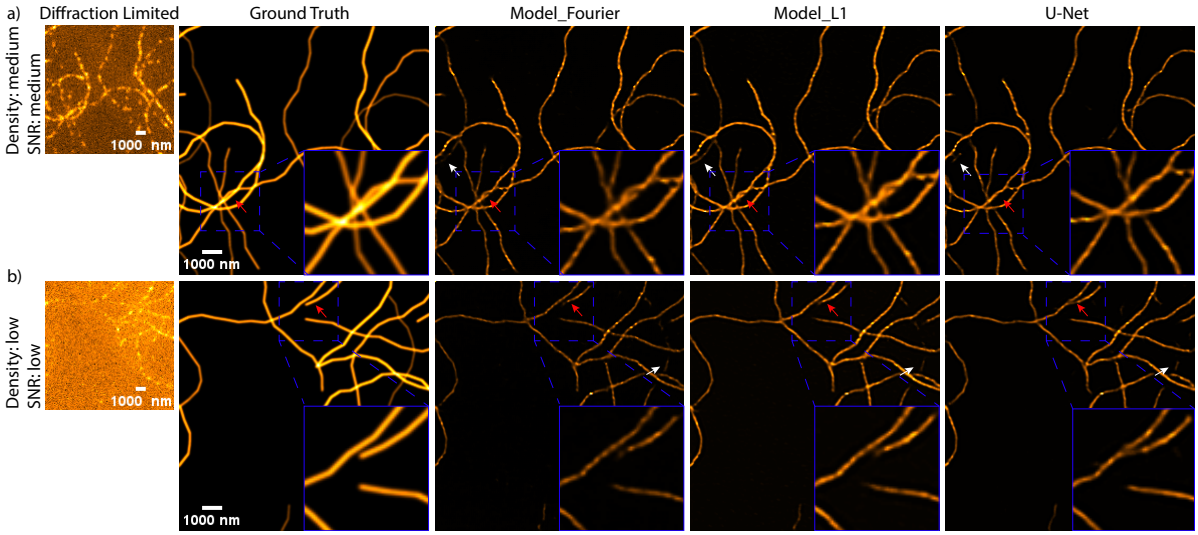


Figure D.3: Left to right: single diffraction-limited frame of a simulated microtubule at different emitter densities and SNR levels. Here, the emitter density is 5000 emitters/μm^2 with on times of 10 ms and off times of 1200 ms and 2400 ms, respectively, to simulate different densities. The SNR level is additionally set by the Ion value, which in this case is 60 and 40, respectively. Each row represents different emitter densities and SNR levels of a microtubule structure; model trained on Fourier loss based on 20 frames; model trained on L1 loss based on 20 frames; U-Net-based SR image based on 20 frames. Scale bar: 1000 nm. The white arrows indicate missing structures in the U-Net-based SR images, which our proposed models predict more accurately. The red arrows highlight better estimations of the filaments compared to the U-Net architecture SR images.

E

Model Trained on Default Linearization SOFI

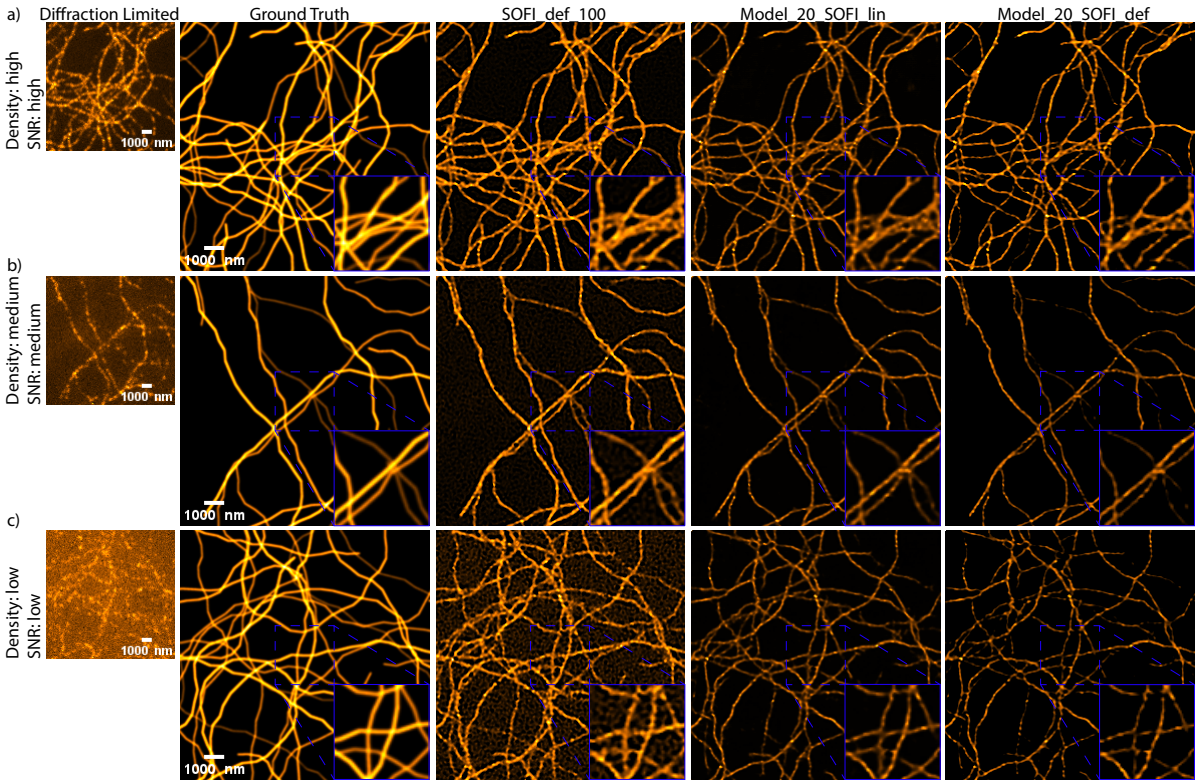


Figure E.1: Comparison of SR reconstructions between the GT, default linearized SOFI, model trained on adaptive linearization, and model trained on default linearization. Left to right: single diffraction-limited frame of a simulated microtubule at different emitter densities and SNR levels. Here, the emitter density is 5000 emitters/μm^2 with on times of 10 ms and off times of 600 ms, 1200 ms, and 2400 ms, respectively, to simulate different densities. The SNR level is additionally set by the Ion value, which in this case is 100, 60, and 40, respectively. Each row represents different emitter densities and SNR levels of a microtubule structure; SR ground truth; Default linearized SOFI SR image based on 100 frames; model-based SR image based on 20 frames trained on adaptive linearized SOFI images; model-based SR image based on 20 frames trained on default linearized SOFI images. Region of interest (ROI) marked by a blue dashed line, showing no background artifacts for both the models based SR reconstructions. However, the model trained on default linearized SOFI images appears to miss structures. Scale bar: 1000 nm.

Table E.1: Results of the model trained on adaptive and default linearized SOFI based on 20 frames, noted as the mean and STD value. It can be clearly seen that the model performs better when it is trained on adaptive linearized SOFI compared to the default, reaching closer to the theoretical value and missing the least amount of structure. Compared to the other metrics, they perform similarly.

Target Type	Resolution [nm]	Pearson	RSP	TNR	TPR
SOFI Adaptive Linearization	115.74 ± 4.03	0.89 ± 0.02	0.85 ± 0.03	0.99 ± 0.01	0.79 ± 0.05
SOFI Default Linearization	146.18 ± 2.06	0.91 ± 0.02	0.83 ± 0.04	0.99 ± 0.00	0.63 ± 0.06

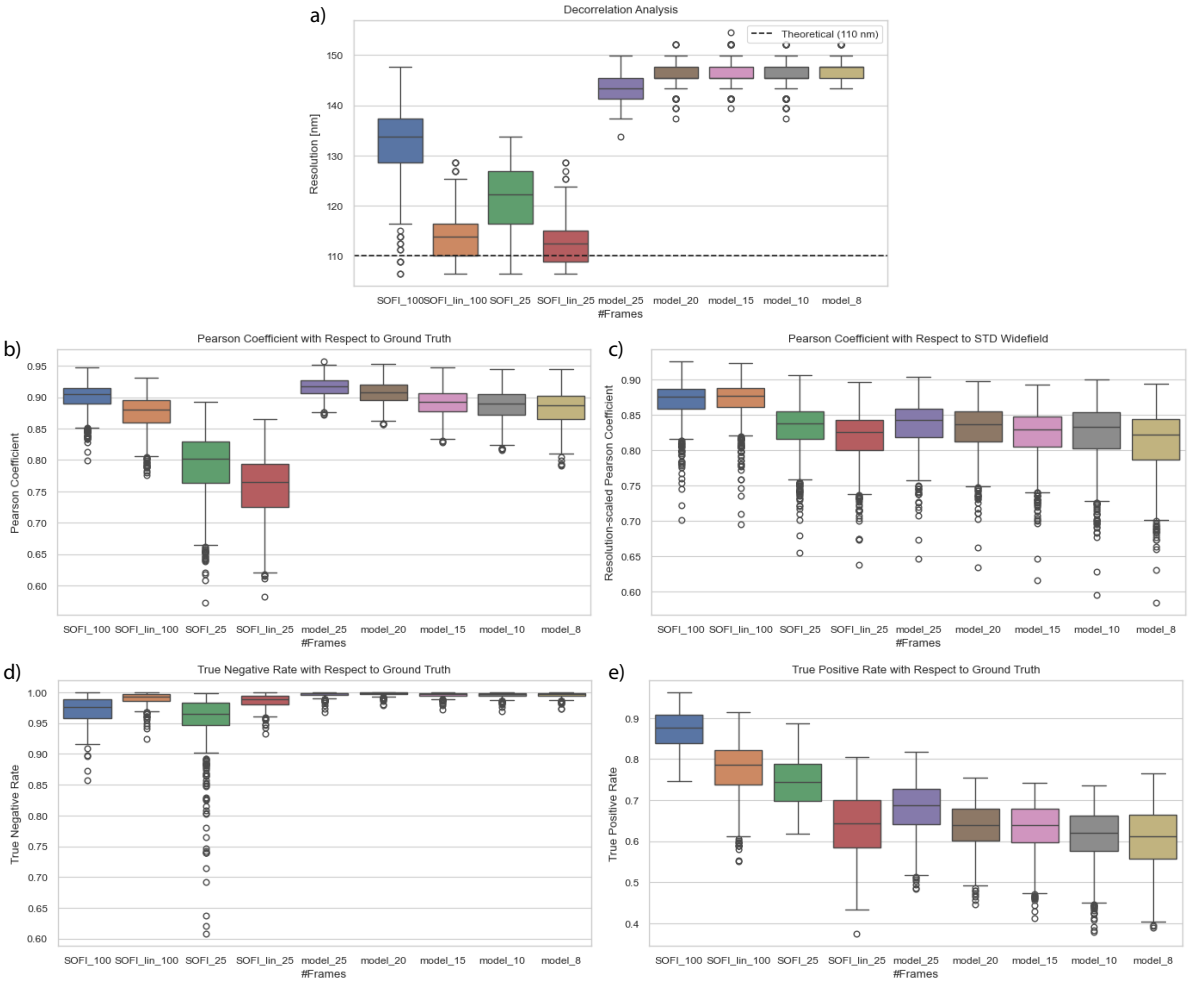


Figure E.2: (a) Decorrelation analysis to measure spatial resolution. The dashed line represents the theoretical spatial resolution after second-order SOFI analysis, which is half of the diffraction limit of 220 nm. The models deviate the most from the theoretical resolution, with deviations ranging from 38 nm for models using 20 frames to 8 frames. (b) Pearson coefficient relative to the ground truth, measuring the similarity between SR reconstruction and the ground truth. The model based on 25 frames scores the highest with a value of 0.92. (c) Pearson coefficient relative to the standard deviation (STD) widefield using SQUIRREL, comparing the rescaled SR image convolved with the estimated PSF of the widefield image to the STD widefield image. The same pattern appears as with the Pearson coefficient relative to the ground truth. (d) True Negative Rate relative to the ground truth, showing that the models and adaptive linearized SOFI score relatively the highest, indicating fewer background artifacts in the SR reconstruction compared to default linearized SOFI. (e) True Positive Rate relative to the ground truth, showing that the default linearized SOFI based on 100 frames misses the least amount of structure, while the models based on 10 and 8 frames miss the most structure.

F

Figures SOFI Acceleration Experiment

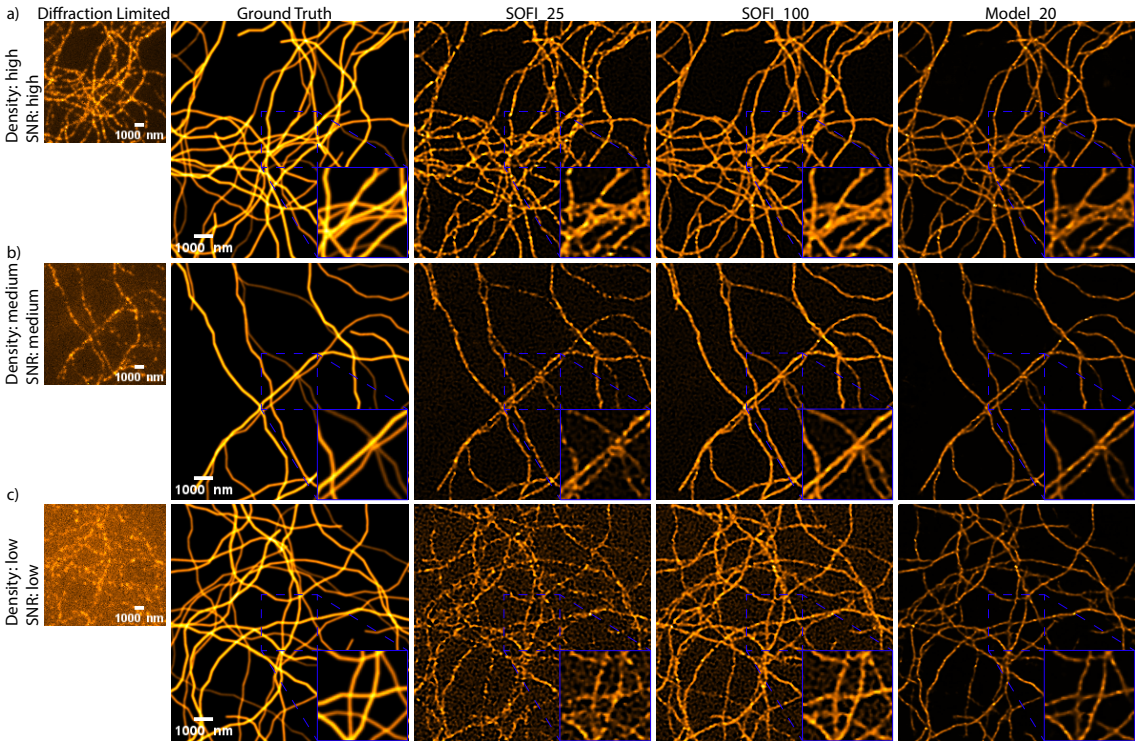


Figure F.1: Comparison of SR reconstructions between the GT, default linearized SOFI, and model. Left to right: single diffraction-limited frame of a simulated microtubule at different fluorophores densities and SNR levels. Here, the fluorophores density is 5000 fluorophores/μm^2 with on times of 10 ms and off times of 600 ms, 1200 ms, and 2400 ms, respectively, to simulate different densities. The SNR level is additionally set by the Ion value, representing illumination intensity, which in this case is 100, 60, and 40, respectively. Each row represents different fluorophores densities and SNR levels of a microtubule structure; SR ground truth; SOFI SR image based on 25 frames; SOFI SR image based on 100 frames; model-based SR image based on 20 frames. Region of interest (ROI) marked by a blue dashed line, showing no background artifacts for the model based SR reconstructions. Scale bar: 1000 nm.

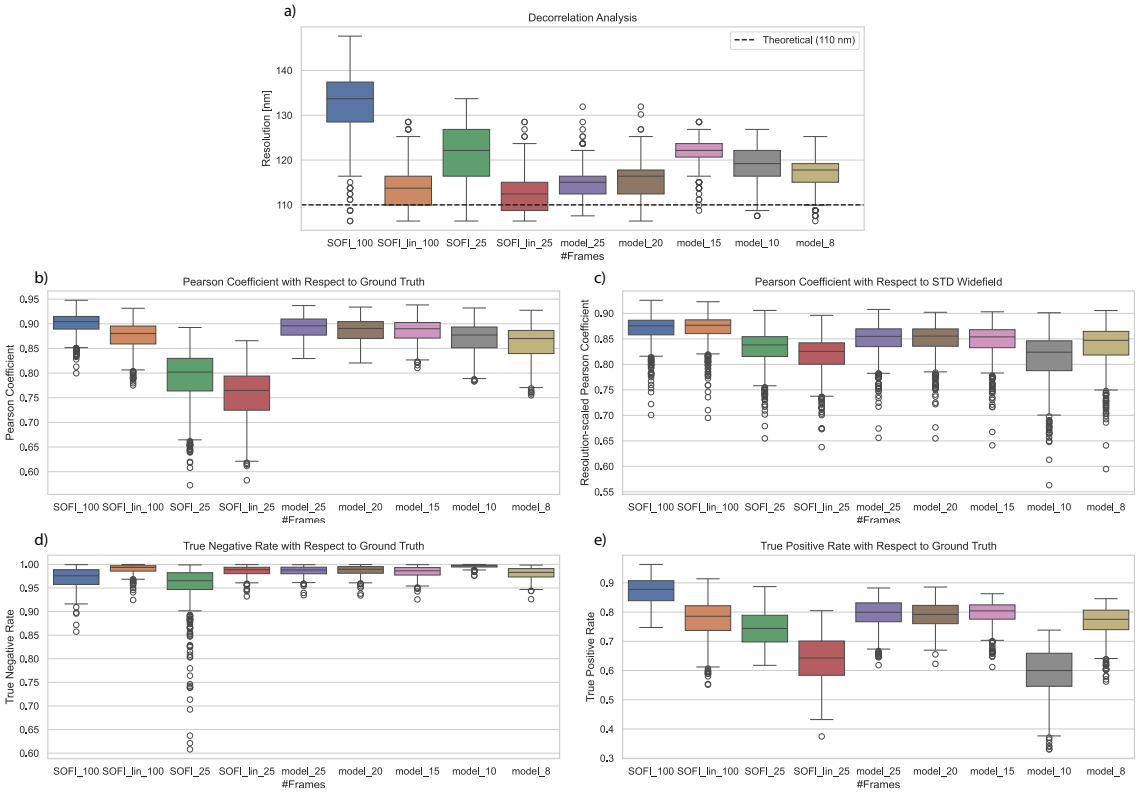


Figure F.2: (a)Decorrelation analysis to measure spatial resolution. The dashed line represents the theoretical spatial resolution after second-order SOFI analysis, which is half of the diffraction limit of 220 nm. Among the models, the one based on 25 frames comes closest to the theoretical value, deviating by around 6 nm, while the default linearized SOFI based on 100 frames performs the worst, deviating by around 25 nm. (b) Pearson coefficient relative to the ground truth, measuring the similarity between SR reconstruction and the ground truth. The default linearized SOFI based on 100 frames has the highest score, while the adaptive linearized SOFI based on 25 frames has the lowest score. For the models, the score worsens after the model based on 15 frames. (c) Pearson coefficient relative to the standard deviation (STD) widefield using SQUIRREL, comparing the rescaled SR image convolved with the estimated PSF of the widefield image to the STD widefield image. The same pattern appears as with the Pearson coefficient relative to the ground truth, with only the model based on 10 frames deviating from the trendline and scoring lower than the model based on 8 frames. (d) True Negative Rate relative to the ground truth, showing that the models and adaptive linearized SOFI score relatively the highest, indicating fewer background artifacts in the SR reconstruction compared to default linearized SOFI. (e) True Positive Rate relative to the ground truth, showing that the default linearized SOFI based on 100 frames misses the least amount of structure, and the model based on 10 frames misses the most structure. Models based on 25 and 20 frames score relatively equally, with the median TPR of the 25-frame models being around 0.2 higher than that of the models based on 20 frames.

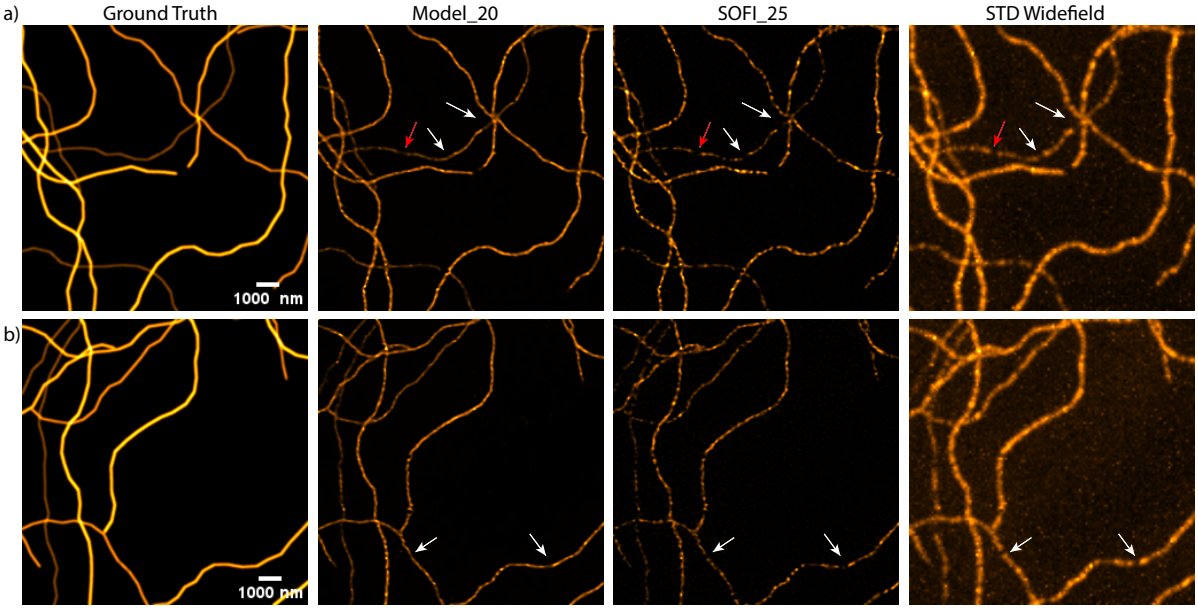


Figure F.3: (a-b) Example of the model showing better filaments reconstructions. Scale bar: 1000 nm. The average widefield images are up-sampled using bilinear interpolation to match the SOFI and model resolutions. The white arrows in the columns of model based on 20 frames show that the filaments are connected, whereas in the column of SOFI based on 25 frames, they are disconnected due to an absence of signal, which can be seen in the column of the average widefield image based on 20 frames. This demonstrates that if the filaments are closely located but there is a narrow gap indicating an absence of signal, the model will connect them. However, if the gap is too large (see red arrow), the model will leave them disconnected.

G

Figures Fixed-Cell Experiment

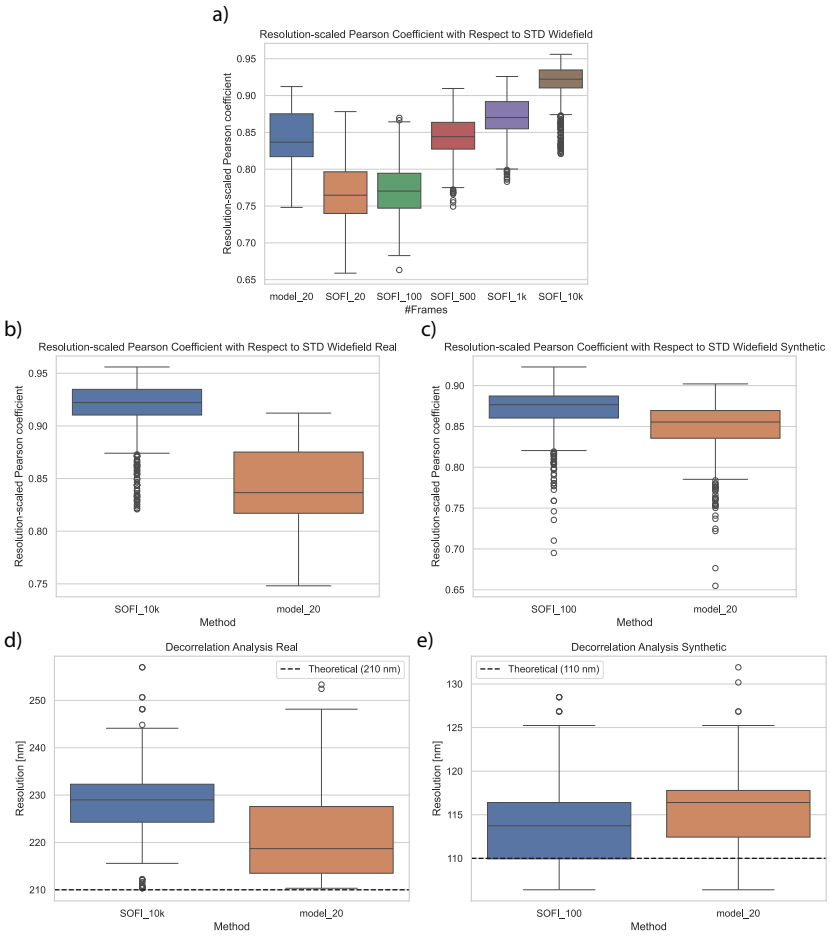


Figure G.1: (a) Resolution-scaled Pearson (RSP) coefficient relative to the STD widefield image based on 10k frames using SQUIRREL analysis. The model trained on 20 frames achieves RSP levels comparable to those of SOFI based on 500 frames. (b-c) RSP coefficient comparison for real and synthetic data using SQUIRREL, with the real data based on 10k frames of STD widefield and the synthetic data based on 100 frames. Both SOFI-based images demonstrate superior RSP performance compared to the models, though the models themselves exhibit similar performance, with RSP values of 0.84 for real and 0.86 for synthetic data. (d-e) Decorrelation analysis results for SOFI and the model trained on 20 frames for both real and synthetic data. The model on real data achieves a resolution closer to the theoretical value of 210 nm, deviating by 10 nm, whereas default linearized SOFI shows a larger deviation. For synthetic data, adaptive linearized SOFI achieves a resolution closest to the theoretical value of 110 nm, with the model deviating by approximately 5 nm. When comparing the models trained on real and synthetic data, they achieve a 1.90-fold and 1.91-fold spatial resolution improvement, respectively.

H

Motion-controlled Mitochondria Experiment

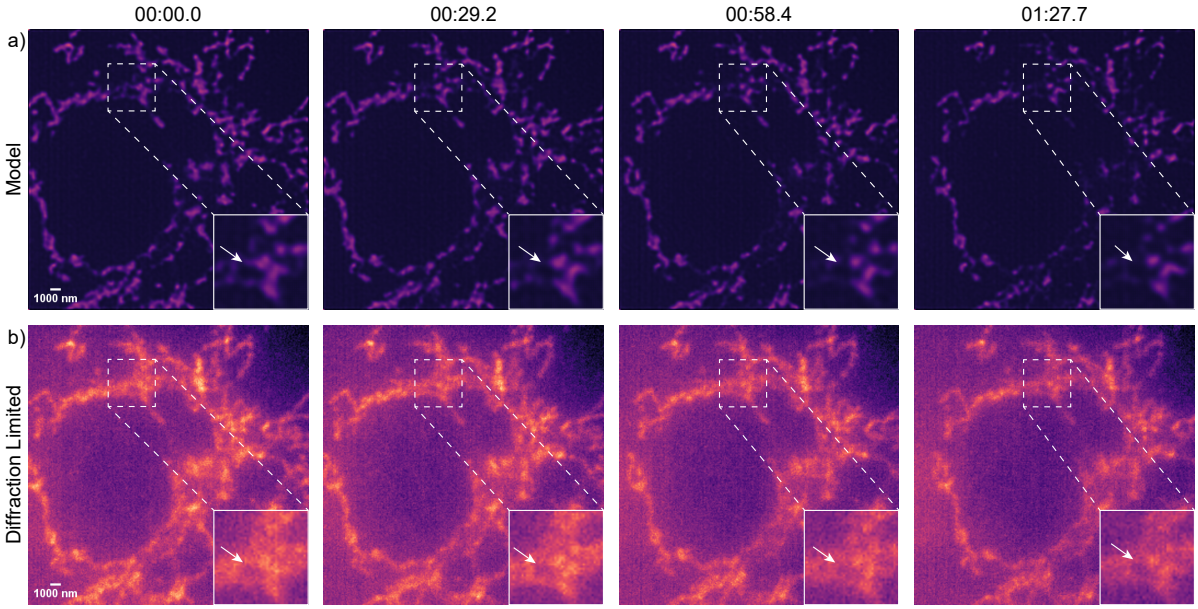


Figure H.1: Reconstruction of a 1.2-minute video of mitochondria in a motion-controlled environment. Using a rolling window, we can output SR images equal to the camera's frame rate. (a) Reconstruction of model without using background subtraction step using 20 frames, showing no SR improvement. (b) Diffraction-limited frame bilinearly interpolated to match the resolution of the model. The white arrow highlights the forming of a "bubble," resembling the fission process of the mitochondria (splitting into two). Scale bar: 1000 nm.

Table H.1: Comparison of the model's super-resolution (SR) results for mitochondria in dynamic and static (motion-controlled and fixed-cell) environments. The model's performance in the dynamic environment shows similar results to that in the static counterpart.

Enviroment	Spatial Resolution [nm]	RSP
Dynamic	221.41 ± 9.45	0.66 ± 0.02
Static	214.86 ± 7.32	0.63 ± 0.11

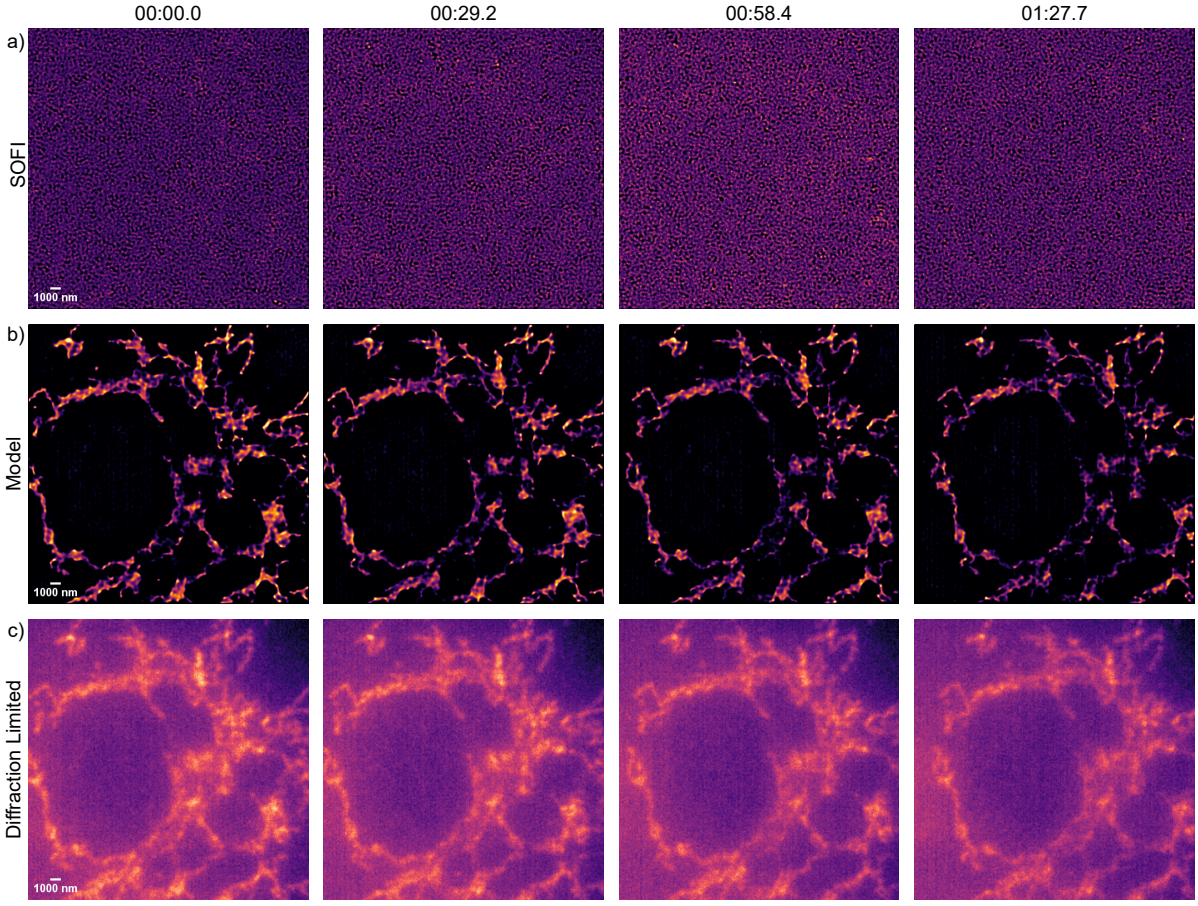


Figure H.2: Reconstruction of a 1.2-minute video of mitochondria in a motion-controlled environment. Using a rolling window, we can output SR images equal to the camera's frame rate. (a) Reconstruction of SOFI using 20 frames, showing that it is not able to reconstruct a SR image. (b) Reconstruction of the model, showing almost no background artifacts. (c) Diffraction-limited frame bilinearly interpolated to match the resolution of the model. Scale bar: 1000 nm.

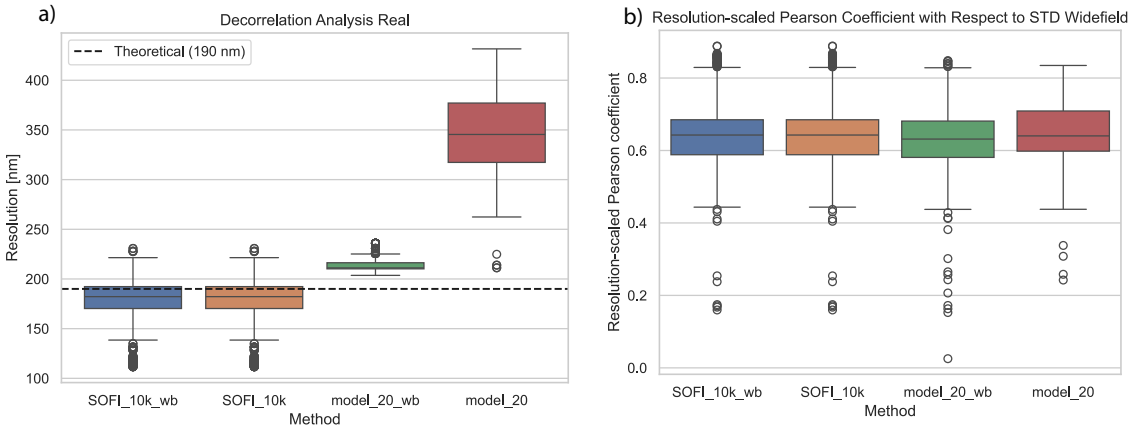


Figure H.3: Results of test set. (a) Decorrelation analysis shows that the SOFI model, which was trained without the background subtraction step, does not exhibit any spatial resolution improvement and remains approximately diffraction-limited. (b) On the other hand, RSP indicates that the model performs slightly better compared to the version with background subtraction. This is expected, as SQUIRREL modifies the SR image to match the widefield standard deviation image. Since the model without background subtraction does not show SR improvement, it performs slightly better in RSP. Additionnaly, background subtraction also resulted in the loss of finer structures, making the remaining features appear somewhat averaged compared to the original diffraction-limited frame, resulting in lower RSP score.

I

Figures Latency Experiment

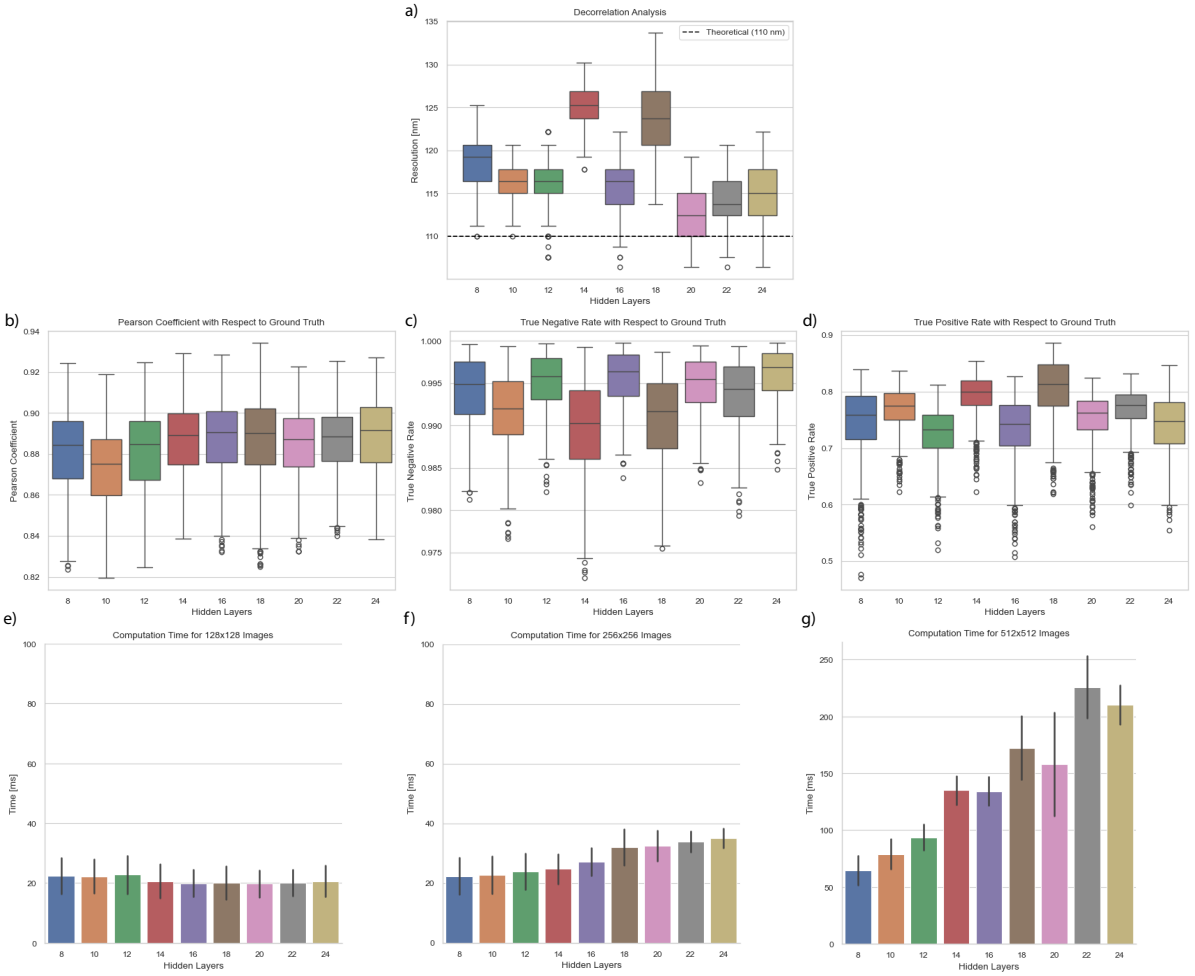


Figure I.1: The model, based on 20 frames and trained on adaptive linearized SOFI images, is compared across different numbers of hidden layers. (a) Decorrelation analysis shows the 20-layer model deviates by 2.5 nm from the theoretical value. (b) The Pearson coefficient peaks with 24 hidden layers, with a decline after 18 layers. (d) True Negative Rate is similar for 12, 16, 20, and 24 layers. (h) The 18-layer model misses the least structure (True Positive Rate). (e-g) Latency trends upward, with 22 hidden layers showing the longest latency across different input sizes.

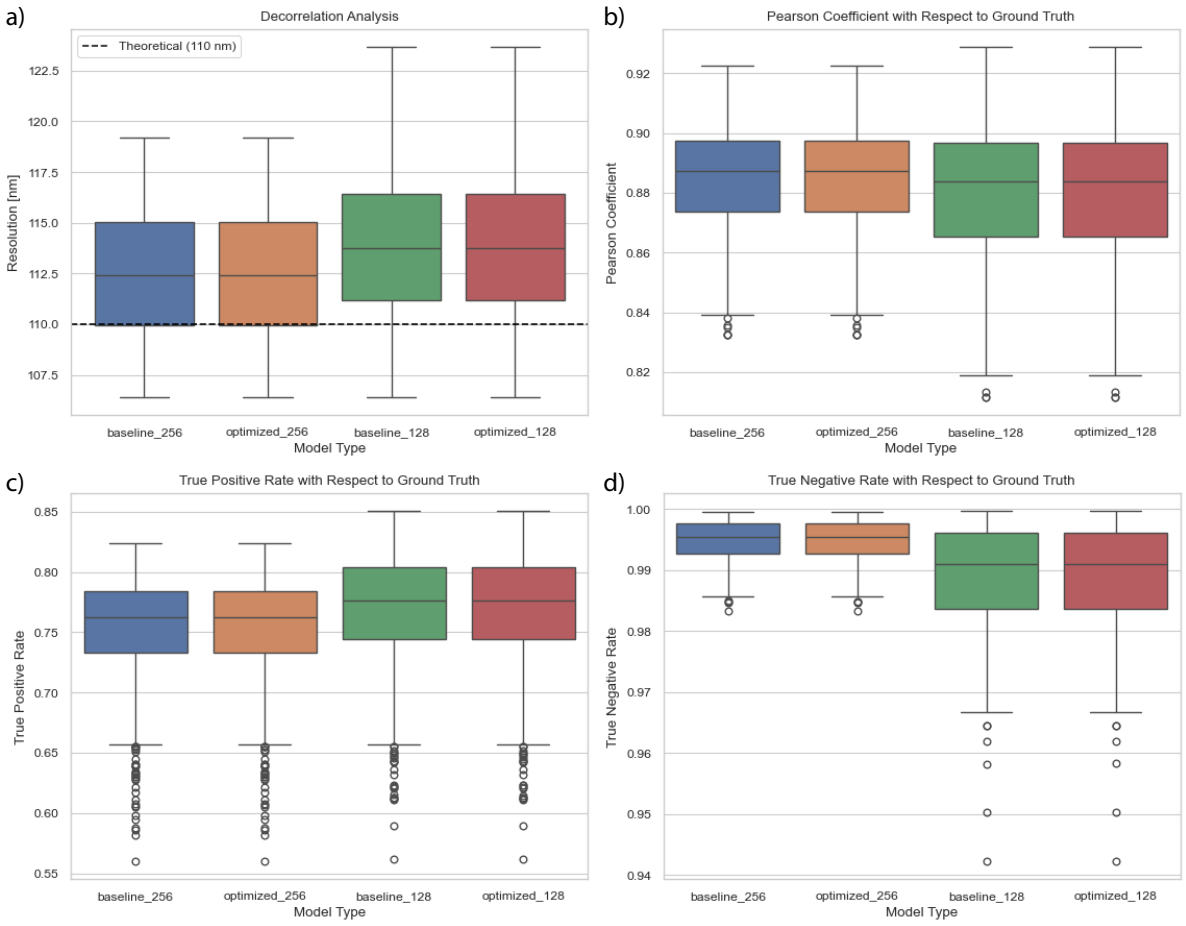


Figure I.2: There is no degradation in performance when optimizing our model using PyTorch TensorRT. Note that the input size of 512×512 is not provided here because of memory constraints when simulating the given test set. It is assumed to have similar performance.

J

Synthetic Movie

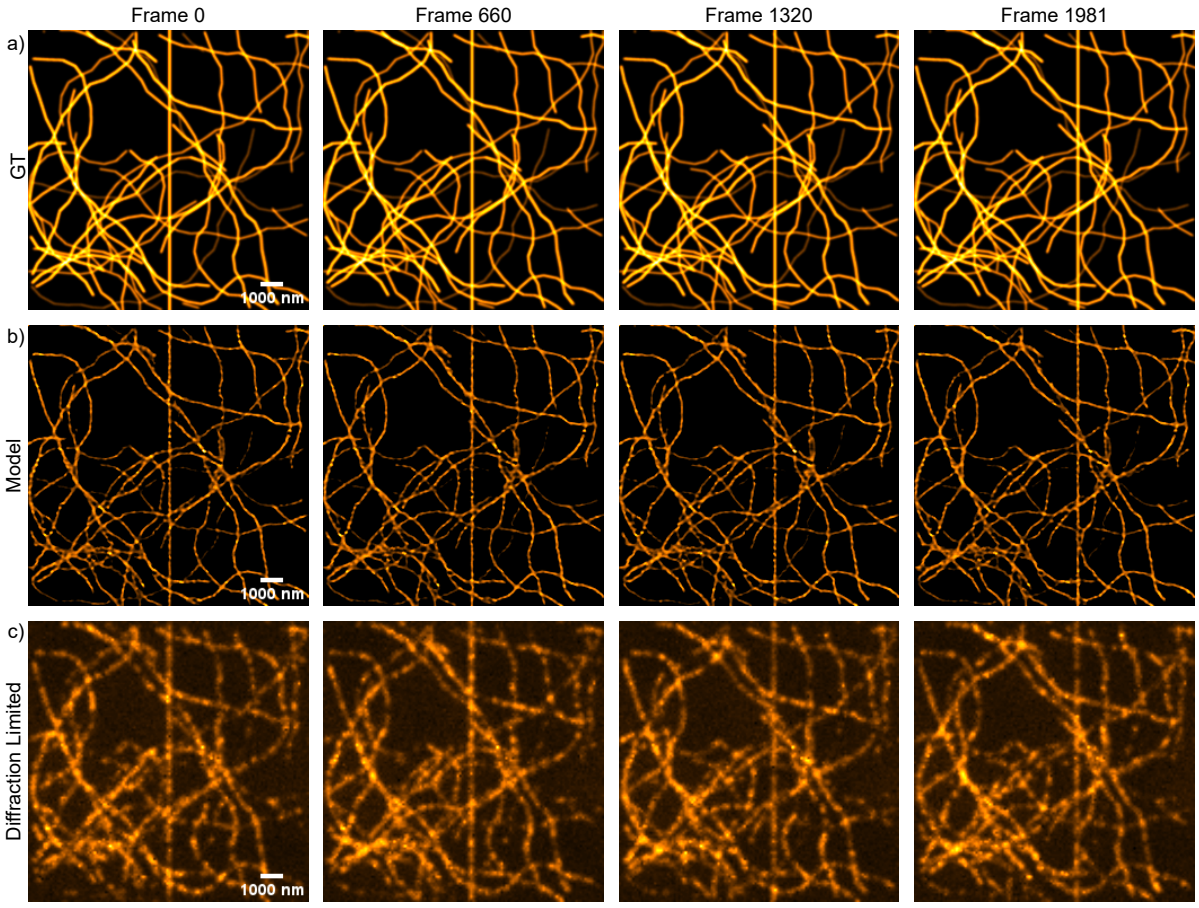


Figure J.1: Comparison of SR reconstructions between the ground truth (GT) and the model for the synthetic microtubules movie. The simulated microtubules feature high fluorophore densities and signal-to-noise ratio (SNR) levels. The fluorophore density is set at 5000 fluorophores/μm^2, with on times of 10 ms and off times of 600 ms. The SNR level is further determined by the Ion value, representing illumination intensity, which is set to 100 in this case. (a) shows the GT SR reconstruction, (b) displays the model's SR reconstruction using a rolling window, and (c) presents the diffraction-limited frame, which is bilinearly interpolated to match the resolution of the model and GT. In the movie, the bar moves one pixel every 20 frames, demonstrating that the model can successfully reconstruct the scene in a dynamic environment.

Table J.1: Comparison of the model's SR results for dynamic and static environments. The model's performance in the dynamic environment shows similar results to that in the static counterpart. Note that the true positive rate (TPR) is higher in the dynamic case due to the increased SNR and higher fluorophore density, which result in better predictions.

Enviroment	Spatial Resolution [nm]	Pearson	TPR	TNR
Dynamic	112 ± 3.78	0.92 ± 0.00	0.73 ± 0.00	0.99 ± 0.00
Static	115 ± 4.31	0.89 ± 0.02	0.63 ± 0.06	0.99 ± 0.00

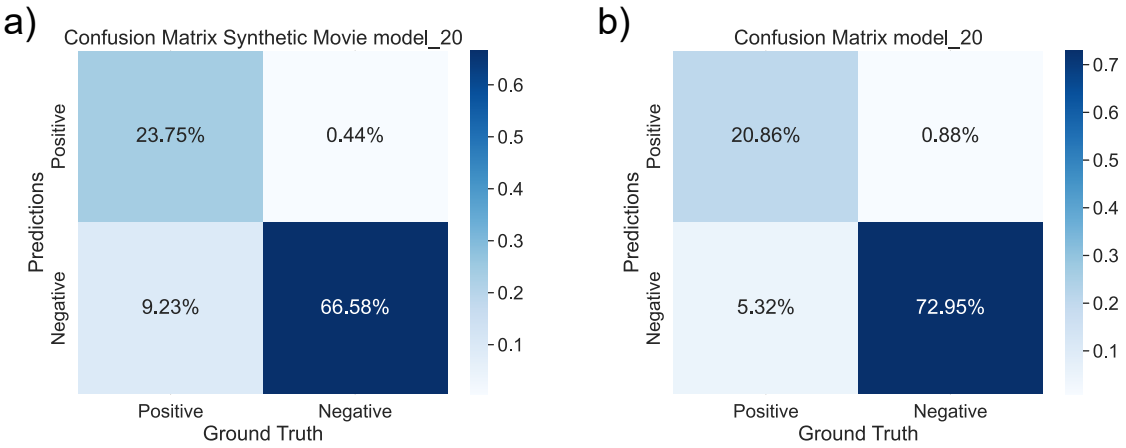


Figure J.2: (a-b) Confusion matrices of the models evaluated on synthetic data. (a) Confusion matrix for the synthetic dynamic movie, characterized by high signal-to-noise ratio (SNR) and high fluorophore density. (b) Confusion matrix for static data with a varied range. Despite the differences between the two data types, both scored similar results.