Maximising the effectiveness of high-density 3D single-molecule localisation microscopy

Ishan Costello

A thesis completed at

Randall Centre for Cell & Molecular Biophysics
School of Basic and Medical Biosciences
Faculty of Life Sciences and Medicine
King's College London
United Kingdom

Under the supervision of
Dr Susan Cox
Prof. Maddy Parsons

Abstract

Single-molecule localisation microscopy (SMLM) separates fluorophore emissions in time to create a tenfold increase in spatial resolution along all three axes. It exploits fluorophore blinking to individuate and localise individual fluorophores, and is usually employed for imaging structures in single cells. In terms of hardware, it is relatively simple, requiring only a high-performance widefield or TIRF fluorescence microscope with high-power lasers. However, the formation of the image is fundamentally reliant on image analysis, making the reliability of that analysis critical to performance.

Another major limitation on SMLM is long acquisition times (typically several minutes at a minimum), which make live-cell imaging challenging. Acquisition times can be shortened by acquiring data with a higher number of fluorophores per frame. However, this often leads to the generation of artificial structures in the reconstructed image, which are very difficult to detect. This risk is significantly increased, often in difficult-to-predict ways, when imaging in three dimensions.

In 2D data, the risk of such artefacts can be minimised by using Haar Wavelet Kernel (HAWK) pre-filtering of the data, which improves the accuracy of a wide variety of fitting methods. The development of a method without artefacts also offered the possibility of detecting artefacts in SMLM images, in particular the ability to check whether an analysis method gives artificial sharpening artefacts on a particular dataset. This is a capability lacking in other super-resolution image assessment techniques, and as it is one of the most common types of artefact, it is likely the lack of a test is leading to artefactual data being published. Using the HAWK approach, we developed HAWK Method for the Assessment of Nanoscopy (HAWKMAN), a fast and resource-inexpensive technology that uses HAWK to assess the validity of local structures in an image. HAWKMAN is unique in that it can perform this assessment across multiple length scales, allowing the user to determine the precision with which they can confidently take measurements.

While the above methods work well on 2D data, 3D data is much more challenging, coming in a variety of forms each prone to different artefacts. To extend

HAWK to 3D, we ran evaluations of the performance of HAWK with regard to artefact reduction in astigmatic, double-helix, and biplane SMLM datasets. Results revealed the complexity of artefact generation in 3D and strong variation in performance between different modalities. These results guided the development of new post-filtering fitting approaches to improve accuracy.

It should be noted that 3D fluorescence microscopy almost always has a poorer resolution in z than in the x/y plane. This is true both for most diffraction-limited methods, and most SMLM 3D approaches. In general, therefore, structures may have similar types of structure in x, y, and z, but when imaged, the information quality in z is strongly reduced. Improvement of z-resolution has been attempted with deep learning approaches, by learning particular types of structure and reproducing them. Here we take a different approach, training to reproduce similar characteristics in the x, y, and z directions, allowing enhanced resolution images to be generated without training to specific structures. The development of these rigorous data assessment tools and more generalisable methods offers an important tool to those working with SMLM microscopy, and suggests new avenues for the development of generative deep learning tools which are less reliant on directly reproducing learned structures.

Declaration

I, Ishan Costello, confirm that the work presented in this thesis is my own. Where information or data analysis has been derived from other sources, I confirm that this has been indicated in this thesis.

Acknowledgements

Though this PhD was undertaken during an usually difficult period, I consider myself lucky to have been so well-supported throughout, by friends, family, and colleagues. There is nowhere near enough room here to thank and applaud all of the people who have made this possible, but to those who have been here through the most instrumental stages:

Many thanks to Pauline and especially to Liisa, who held my hand through my first wobbly steps working in a laboratory. To Hugo, James, and Nick, for offering your wealth of experience and patience to me when I was still learning my way around a light sheet. To my second supervisor Maddy, and to Simon, Baljinder, and Sergi – thank you so much for your guidance, your expertise, and your kindness when I was feeling uncertain.

I have been exceptionally well-supported and well-loved by friends and family throughout this process. Tom and Alex: between you both, you taught me most of what I know about cells. You kept me level-headed and focussed, and you didn't let me sweat the small stuff. Thanks for your support, and for making me laugh when I didn't think I could. To Anneliese and Dylan – you make me feel less alone when I'm alone, and less mad when I'm... mad. Thank you for your warmth, your advice, your wit. I hope we keep making the time to see each other.

A few extra-special thanks are also in order. To my mum and dad, thank you for being there whenever I've needed you, for egging me on when I've had doubts, for making me feel worthy and capable. I will never have the words to express my gratitude to you both. You are, very literally, the best. To my partner, Lucy, who could not have been more generous, more supportive, or more understanding when I was so wobbly and stressed. God only knows what karmic justice has been satisfied by my having the love and support of someone so wonderful. There is no way I could have done this without you – thank you.

To Richard, for your exceptional generosity, incredible ingenuity, and for teaching me so much. I didn't actually know people could do all of the things that you can, all at once. This thesis wouldn't be here if you hadn't been exactly who you are – so thank you! You've been a truly exceptional and inspiring colleague.

And last, to Susan, my first supervisor. Thank you for believing in me when I didn't, for your unerring patience, for the hours you've given me, for your ideas, for your kindness. I have learned more from you than anyone else in my professional life. You are an exceptional researcher and an even better mentor; an inspiration to us all, and a credit to your profession. Please, keep on doing exactly what you're doing.

Contents

1 Introduction **9**

1.1 The Compound Microscope 9

1.2 Fluorescence Microscopy 11

1.2.1 Mechanisms & use of fluorescence 11

1.2.2 Resolution & the diffraction limit 15

1.3 Super-Resolution 18

1.3.1 Nanoscopic fluorescence microscopy 18

1.3.2 Principles of localisation microscopy 19

1.3.3 Blinking and Bleaching of fluorophores 22

1.3.4 Accessibility of SMLM 25

1.4 What makes a good super-resolution image? 26

1.4.1 Resolution as localisation precision 26

1.4.2 Resolution as self-similarity 29

1.4.3 Artefacts 31

1.5 SMLM in live cells 34

1.5.1 The problem of temporal resolution 34

1.5.2 Overcoming artefacts 35

1.5.3 The Haar Wavelet Kernel 38

1.5.4 Detecting errors in SMLM 41

1.6 SMLM in three dimensions 43

1.7 Deep learning & SMLM 46

1.7.1 Principles of deep learning 46

1.7.2 The convolutional NN 49

1.7.3 Principles of generative adversarial networks 51

1.7.4 Deep learning in SMLM 52

2 HAWKMAN **56**

2.1 Introduction 56

2.2 Materials & Methods 59

2.2.1 HAWKMAN sharpening detection 59
2.2.2 Cell preparation for experiments: tubulin & clathrin 63
2.2.3 Cell preparation for experiments: mitochondria 64
2.2.4 Reconstruction software 66
2.2.5 Statement on data selection 67
2.3 Results 68
2.3.1 The HAWKMAN algorithm 68
2.3.2 Testing with benchmark data 72
2.3.3 The effect of non-linear transformation SMLM algorithms on image resolution metrics 83
2.3.4 Performance comparison to SQUIRREL 87
2.3.5 HAWKMAN can indicate the presence of unresolved structure 90
2.3.6 Common biological structures & two-colour data 92
2.4 Discussion 96

3 HAWK in 3D 98
3.1 Introduction 98
3.2 Materials & Methods 102
3.2.1 Astigmatic sample preparation & imaging protocol: nuclear pore complex 102
3.2.2 Astigmatic sample preparation & imaging protocol: tubulin-AF568 103
3.2.3 Simulations of resolving 3D structure 104
3.2.4 Astigmatism simulations 104
3.2.5 Statement on the choice of ThunderSTORM as fitting algorithm 104
3.2.6 Astigmatic data analysis 105
3.2.7 DHP sample preparation & imaging protocol: lamin 105
3.2.8 DHP simulations 106
3.2.9 DHP data analysis 107
3.2.10 Biplane simulations 107
3.2.11 Biplane sample preparation & imaging protocol: mitochondria 107
3.3 Results 108
3.3.1 Note on assessment of 3D images 108
3.3.2 The Astigmatic PSF 109
3.3.3 Loss of z-precision in HAWK-processed astigmatic reconstructions 119
3.3.4 The Double-Helix PSF 120
3.3.5 Multifocal Plane SMLM 126

3.4 Discussion 133

4 AFFIRM3D **136**

4.1 Introduction 136

4.1.1 Anisotropy of microscope resolution 136

4.1.2 Convolutional neural networks in biological imaging 138

4.1.3 Overcoming resolution anisotropy with generative NNs 141

4.1.4 Use of Fourier-space information 147

4.1.5 Choice of training data 148

4.2 Materials & Methods 149

4.2.1 Engineering an appropriate cost function 149

4.2.2 The axial Fourier-space loss 150

4.2.3 Handling anisotropically sampled images 151

4.2.4 Training data 152

4.3 Results 153

4.3.1 AFFIRM3D v0.1 153

4.3.2 Assessment of images generated by deep learning algorithms . 156

4.3.3 Visual assessment by a human 157

4.3.4 Line scan 157

4.3.5 Normalised root mean squared error 157

4.3.6 Intersection over union 158

4.3.7 Structural similarity index 159

4.3.8 BRISQUE 160

4.3.9 Performance of the AFFIRM3D algorithm without GT loss . . 161

4.3.10 Optimisation 161

4.3.11 Performance of initial optimisation 164

4.4 Discussion 164

5 Conclusion & Outlook **166**

5.1 Accessibility & efficacy of SMLM imaging 166

5.2 HAWKMAN provides advanced and comprehensive image assessment 167

5.3 HAWK successfully removes artefacts from 3D datasets 169

5.4 Improving the resolution of 3D imaging 170

5.5 Concluding remarks 170

List of Figures

1.1 The Abbe sine condition of magnification 10
1.2 The numerical aperture of three theoretical imaging systems 11
1.3 A Jablonski diagram depicting the fluorescence cycle. 13
1.4 A figure showing the breadth and variety of available fluorescent proteins 14
1.5 Simulated pairs of PSFs to visualise the resolution limit. 17
1.6 Super-resolution imaging principles 20
1.7 Operating principle of single-molecule localisation microscopy. 21
1.8 Effectiveness of SMLM for imaging sub-diffraction limit structures . . 22
1.9 The operating procedure for the fourier ring correlation 30
1.10 Change in the apparent structure of a SMLM image is affected by high activation densities 32
1.11 Artefacts can arise from the sample parameters 33
1.12 Local dimensionality of the imaged structure impacts the acquisition time needed to achieve a particular resolution. 36
1.13 Simplified schema of HAWK's effect on a pair of blinking fluorophores 39
1.14 Error mapping with SQUIRREL 42
1.15 Diagram explaining the principles of artificial neural networks 48
1.16 Basic 2D convolution such as might be found in a CNN 50
1.17 Architecture of a basic GAN 52

2.1 HAWK removes artefacts from SMLM datasets of mitochondria for a range of emission densities 69
2.2 The processing steps of the HAWKMAN algorithm, producing the sharpening map, structure map and confidence map. 70
2.3 A comparison of SE and ME fitting of high emitter density data and equivalent LD data. 73
2.4 Results of HAWKMAN analysis on experimental microtubule data from the Localisation Microscopy Challenge comparing SE and ME fitting. 76

2.5 HAWKMAN recognises performance improvements in SRRF when parameters are optimised. 78
2.6 HAWKMAN results for Single-Emitter (SE) fitting of microtubule data 80
2.7 HAWKMAN results for Multi-Emitter (ME) fitting of microtubule data 82
2.8 Performance comparison when using alternative HD algorithms to HAWK to create the reference image. 84
2.9 A random distribution of molecules 85
2.10 SRRF cannot render high-frequency data in a reconstruction of spatially random data 86
2.11 A comparison of SQUIRREL and HAWKMAN error map sensitivity. 89
2.12 HAWKMAN analysis of experimental data of Z1Z2 antibody labelling of muscle sarcomeres demonstrates HAWKMAN's ability to reveal the possibility of structure unresolved in both inputs. 91
2.13 HAWKMAN can assess reconstructions from a variety of different structures. 93
2.14 HAWKMAN is sensitive to artefacts likely to be encountered in two colour imaging. 95

3.1 An illustration of the symmetry ambiguity 100
3.2 A simplified schematic of the astigmatic 3D localisation approach . . 110
3.4 HAWK removes artefacts from high-density astigmatic 3D SMLM image crops of simulated tubulin from the Microscopy Challenge . . . 115
3.5 HAWK removes artefacts in all three dimensions from ultra-high density astigmatic datasets of microtubules 117
3.6 HAWK removes artefacts from high-density astigmatic 3D SMLM image crops of the nuclear pore complex 118
3.7 Diagram demonstrating how a double-helix imaging set-up localises molecules in z 121
3.8 An exhibition of the many localisation ambiguities that can arise from two laterally adjacent emitters when using the DHP. 123
3.9 HAWK removes artefacts from simulated DHP SMLM datasets at three different densities 125
3.10 HAWK restores undersampled data to a lamin DHP dataset 127
3.11 HAWK removes artefacts from high-density biplane SMLM simulations of straight line pairs 129
3.12 Simulated mitochondria data at LD and MD show HAWK greatly reduces mislocalisations in 3D biplane datasets 131

3.13 Reconstructions of 3D mitochondria data using a TOM20 stain show the distortions in z-position arising from excess emitter density, and how they are greatly reduced by HAWK 132

4.1 The PSF is much more extended in the axial direction than the lateral directions. 137
4.2 Data from the CARE paper demonstrates effective image improvement using deep learning. 143
4.3 Typical failure modes of generative networks 144
4.4 A diagram of the training loop for the first version of AFFIRM3D . . 154
4.5 Output z-slices and y-slices of the initial version of AFFIRM3D after training on simulated data 155
4.6 Inputs and outputs of optimisation on assymmetrically sampled data 162
4.7 Test of Adam vs LBFGS optimisers for AFFIRM3D 163

Abbreviations

3B	Bayesian analysis of blinking and bleaching
4πM	4-pi microscopy
AI	artificial intelligence
AFFIRM3D	AI for forging isotropic resolution in 3D micrographs
ANNA-PALM	artificial neural network-accelerated photoactivated localisation microscopy
BRISQUE	blind/referenceless image spatial quality evaluator
BSA	bovine serum albumin
CARE	content-aware image restoration
CCD	charge-coupled device
CMOS	complementary metal–oxide–semiconductor
CNN	convolutional neural network
COT	cyclooctatetraene
CS	compressed sensing
DAISY	dual-view astigmatic imaging with SAF yield
DHP	double-helix point-spread function
DMEM	Dulbecco's modified eagle Media
DNA	deoxyribonucleic acid
DONALD	direct optical nanoscopy with axially localized detection
DeconSTORM	deconvolutional stochastic optical reconstruction microscopy

EM	electron microscopy
FALCON	fast localization algorithm based on a continuous-space formulation
FA	formaldehyde
FBS	fetal bovine serum
FM	fluorescence microscopy
FRC	Fourier ring correlation
FRET	Förster resonance energy transfer
FWHM	full width at half-maximum
FZ	axial Fourier-space (loss)
GAN	generative adversarial network
GLOX	glucose oxidase with catalase
GT	ground truth
HAWK'd	Haar Wavelet Kernel pre-processed
HAWKMAN	Haar Wavelet Kernel Method for the Assessment of Nanoscopy
HAWK	Haar Wavelet Kernel
HD	high-density
HWT	Haar wavelet transform
HeLa	Henrietta Lacks cervical cancer cells
IM	interferometric microscopy
L1	absolute error (loss)
L2	squared error (loss)
LBFGS	limited-memory Broyden-Fletcher-Goldfarb-Shanno
LD	low-density
LED	light-emitting diode
LEF	localisation error function
MATLAB	Matrix Laboratory
MD	medium-density
MEA	2-aminoethanethiol
ME	multi-emitter
MNIST	Modified National Institute of Standards and Technology
MS-SSIM	multi-scale structural similarity index
NA	numerical aperture
NLM	nonlocal means
NN	neural network
NRMSE	normalised root mean squared error
PAINT	points accumulation for imaging nanoscale topography
PALM	photoactivated localisation microscopy

PBS	phosphate-buffered saline
PCA	protocatechuic acid
PCD	protocatechuate 3,4-dioxygenase
PFA	paraformaldehyde
PSF	point-spread function
RCAN	residual channel attention network
RMSE	root mean squared error
ROI	region of interest
RSF	resolution scaling function
RT	room temperature
SAF	supercritical angle fluorescence
SE	single-emitter
SIMPLER	supercritical illumination microscopy photometric z-localization with enhanced resolution
SIM	structured illumination microscopy
SMLM	single-molecule localisation microscopy
SNR	signal-to-noise ratio
SOFI	super-resolution optical fluctuation imaging
SQUIRREL	super-resolution Quantitative Image Rating and Reporting of Error Locations
SRRF	super-resolution radial fluctuations
SR	super-resolution
SSIM	structural similarity index
STED	stimulated emission depletion (microscopy)
STORM	stochastic optical reconstruction microscopy
TIRF	total internal reflection fluorescence
TRIS	tris(hydroxymethyl)aminomethane
dSTORM	direct stochastic optical reconstruction microscopy
fPALM	fluorescence photoactivation localisation microscopy
hGAN	hybrid generative adversarial network
iPALM	interferometric photoactivated localisation microscopy
sCMOS	scientific complementary metal–oxide–semiconductor

Chapter 1

Introduction

1.1 The Compound Microscope

The standard compound microscope is the basis on which all modern designs are based. White light is used to illuminate a sample and provide contrast with which to view structures. Light that has interacted with the sample is captured by the objective lens, which focusses a real image of the sample to the eyepiece or camera, allowing a virtual image of the object to be passed to the user. The sample can be backlit (brightfield illumination) or lit through the objective lens (epi-illumination). The objective lens' breakthrough achievements are its ability to achieve a high level of magnification and collect a large volume of light. Considerations in the design of the objective lens can be thought of as a trade-off between the amount of light that the objective can capture, and the level of magnification it can achieve. These limitations are best thought of through an understanding of the Abbe sine condition [1]:

$$\frac{\sin(\alpha_o)}{\sin(\alpha_i)} = \frac{\sin(\beta_o)}{\sin(\beta_i)} = |M| \tag{1.1}$$

The condition states that the sines of the object-space and image-space angles must be proportional to achieve a clear image, and that the ratios of these sines is equal to the magnification produced by the optical system (Figure 1.1). A larger object-space angle can only be achieved if the space between the sample and the objective is reduced. This reduces the amount of light available to the system, as high-magnification objective lenses require smaller apertures to ensure that the objective does not capture out-of-focus light. The amount of light an objective can

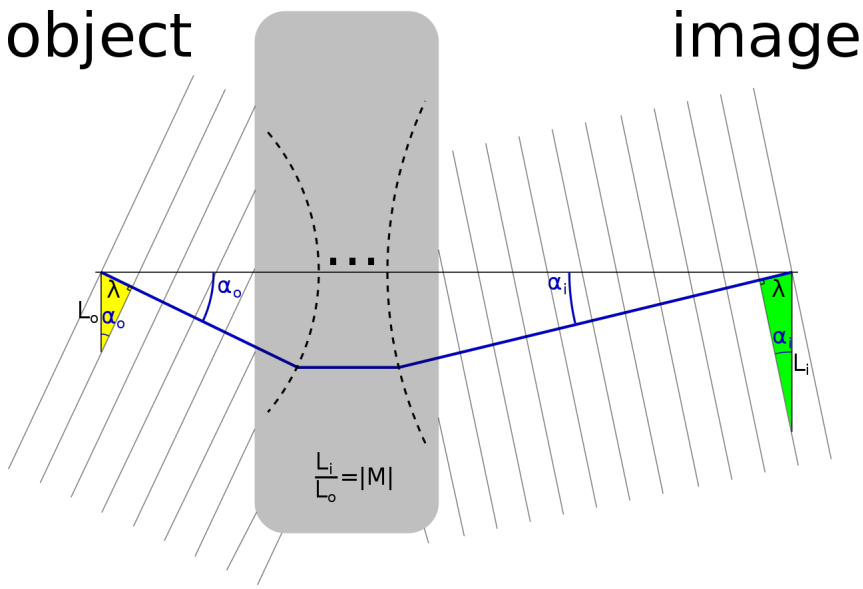


Figure. 1.1. The Abbe sine condition of magnification as demonstrated by a ray trace diagram of an imaging system (grey box). The imaging system obeys the Abbe sine condition, meaning the ratio between the angles α_i and α_o denote the magnification of the object. Figure obtained from *Abbe sine condition* [2].

capture is denoted by its numerical aperture [1], NA:

$$NA = n \cdot \sin(\theta) \tag{1.2}$$

Where n the refractive index of the medium in which the sample and objective are immersed, and θ the angular aperture, or the maximum angle to the normal from which light incident to the objective can reach its focus.

At this point one might think that objective design is the only limiting factor when conducting light microscopy. But while it is certainly the workhorse of a microscope, the objective does not single-handedly govern what can and can't be imaged. The other essential component of an imaging system is the light itself, and how it is employed in the system to render the image. In brightfield illumination, a sharp variation in opacity across the objects of interest is required, because regardless of the kind of objective used, an experimenter is simply looking at a relative contrast map between those parts of the sample that allow light to pass through easily, and those that don't. In some cases, these difficulties can be resolved using illumination schema that promote contrast, such as darkfield illumination, which illuminates the sample at a sharp angle so that only light scattered by the sample is imaged; or phase contrast and differential interference contrast microscopy, which introduce a phase shift or polarisation difference between scattered and unscattered light resulting in higher contrast via destructive interference. While these developments

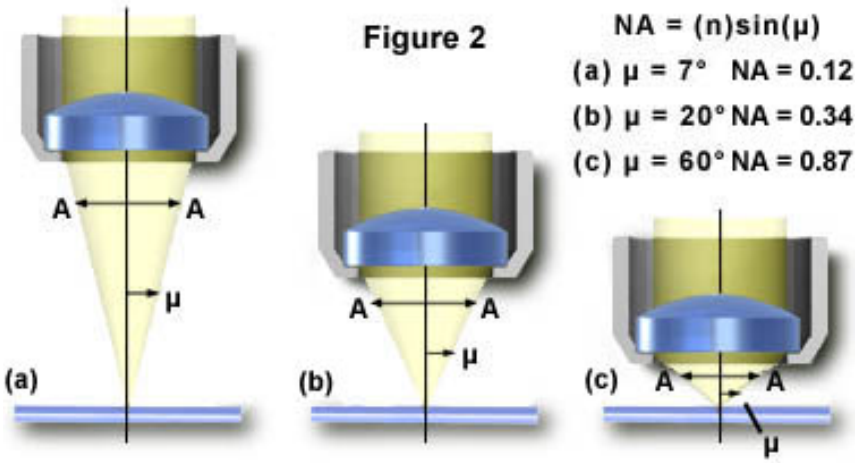


Figure. 1.2. The numerical aperture (NA) of an imaging system denotes its ability to collect light. Despicted are three theoretical objective lenses with different NAs, showing their differing abilities to collect light from a sample. This is because the NA increases with the angular range over which the objective can collect light. As a result, it impacts the distance from the sample required to achieve a focussed image, with higher-NA objectives requiring a shorter distance to avoid collecting out-of-focus light. Figure obtained from *Numerical Aperture and Resolution* [3].

have greatly improved the contrast and thus the available information in optical microscopy images, they are not sufficiently powerful for quantitative microscopy of the smaller organelles, molecule tracking, or 3D imaging, for instance. For these goals, the molecules of interest need to be individuated and illuminated against a dark background.

1.2 Fluorescence Microscopy

1.2.1 Mechanisms & use of fluorescence

The assemblage and articulation of many of the underlying principles governing the operation of a modern fluorescence microscope is often credited to Sir George G. Stokes, who coined the term fluorescence in describing the mechanism by which certain compounds, when irradiated with a discrete wavelength of light, emit light in turn. He additionally observed that this emitted light is of a greater wavelength (and thus, as was later discovered, of a lower energy) than the illuminating photons; the Stokes' Shift phenomenon. This energy loss occurs because electrons excited from the ground state to the singlet excited states by irradiation will only fluoresce when relaxing from the lowest vibrational level (ν_0) of the first excited singlet (S_1) state, meaning electrons at any energy level greater than the $S_1(\nu_0)$ state must lose energy non-radiatively before fluorescing, if they fluoresce at all. This can occur

by one of three mechanisms: through intersystem crossover, internal conversion and vibrational relaxation, or by transfer to another molecule [4]. During intersystem crossover, energy is lost in transitioning the excited singlet-state electron into the much longer-lived (and therefore highly reactive) excited triplet state, which can then relax to the ground state via phosphorescence. Vibrational relaxation of the electron distributes excess kinetic energy between the vibrational modes of the fluorophore and its environment, bringing it down to $S_1(\nu_0)$, from which fluorescence by radiative transition down to the ground state is possible. If the molecule is in the S_2 or S_3 states, diabatic 'internal conversion' of the electron's electromagnetic energy will enable it to relax non-radiatively to S_1 before vibrational relaxation occurs.

To visualise these structures with fluorescence microscopy (FM), a cellular molecule of interest can be tagged by various means with a fluorescent molecule. If this fluorophore is illuminated with light of wavelength λ_i, it will only emit light of wavelength λ_e because the $S_1(\nu_0) \rightarrow S_0$ transition is of discrete energy for a given fluorophore, where $\lambda_e > \lambda_i$ in accordance with Stokes' Shift. Thus, in a fluorescence microscope, the emitted light λ_e is passed through a filter selective for this wavelength, removing excitatory (λ_i) and ambient light from the image so that only the molecule of interest remains.

The phenomenon of fluorescence has long been exploited in biological research because it enables high-contrast imaging of specific cellular structures. It is fantastically easy to implement, as it requires no more than a sample tagged with a fluorophore, and a microscope with an appropriate illumination wavelength, granting a high degree of reproducibility in experiments. A particular advantage of FM is the imaging of live processes: living cells can be tagged with fluorescent proteins, allowing long-term imaging of biological processes with impressive detail. FM is not only confined to *in vitro* experiments, either, but can be performed on living organisms to study cells and tissues in their native *in vivo* environment. This is extremely important as *ex vivo* experiments can produce misleading results, since cells sometimes exhibit different behaviours in a native context. For these reasons it is one of biology's *main* tools.

Most commonly, FM is carried out using a **widefield** [6] or a **confocal** [7] microscope. In the former approach, a column of laser or LED light illuminates the whole sample at once. It is foolproof, cheap, and very quick, enabling imaging of living or fixed samples with relatively little light, in a relatively low-maintenance system. Despite these advantages, the use-case of widefield fluorescence microscopes is generally limited to thin samples. Thicker samples are more of a challenge for this set-up because the column of illuminating light is easily scattered by the sample,

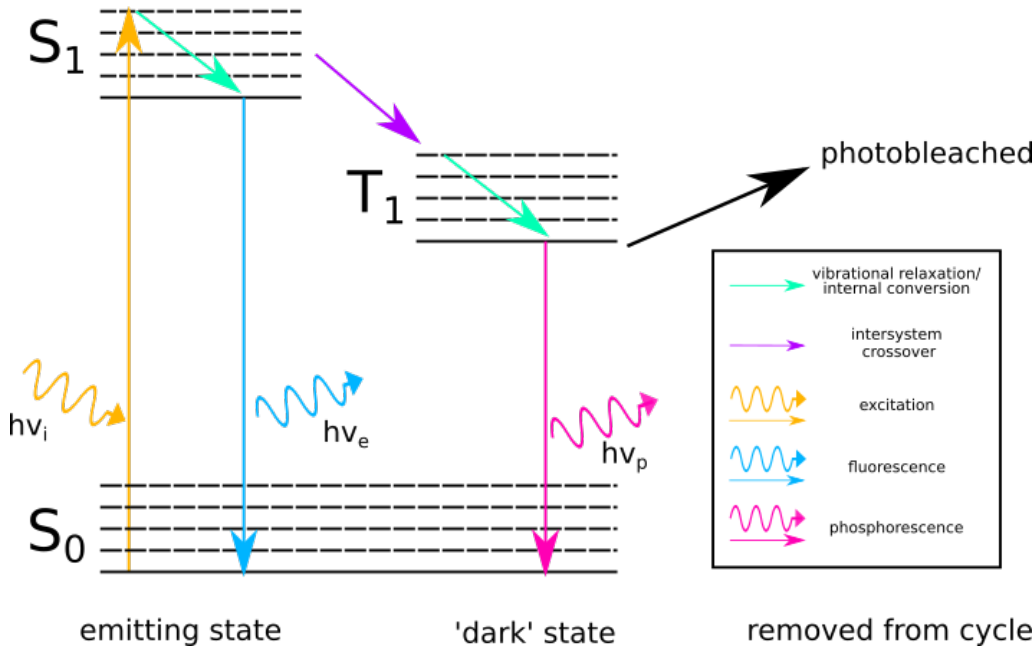


Figure. 1.3. A Jablonski diagram depicting the fluorescence cycle. A photon $h\nu_i$ incident on the fluorophore can excite it from the ground state S_0 to the excited state S_1 (orange arrow). Fluorescence only occurs during the transition from the lowest vibrational level of the excited state to the ground state (blue arrow), so some energy is lost non-radiatively (green arrow) before fluorescence can occur – this loss is responsible for the 'Stoke's shift' in wavelength between $h\nu_i$ and $h\nu_e$. When excited, electrons can also transition between spin states to the triplet state T_1, which has a much longer-lived decay process (called 'phosphorescence' when emitting) and is therefore much more vulnerable to photobleaching. Here the electron is permanently removed from the absorption/emission cycle by reaction with another molecule, and fluorescence ceases. Triplet state decay is sometimes non-emitting, and usually not visible in a fluorescence image. This is due to the increased decay time: because phosphorescence lasts longer, more energy is dissipated, meaning less energy is dissipated per photon – the Stoke's shift is more pronounced, resulting in much longer-wavelength photons that are filtered out before imaging. Hence, some microscopists refer to it as the 'dark sate'. If the electron has relaxed to S_0 the cycle can begin again.

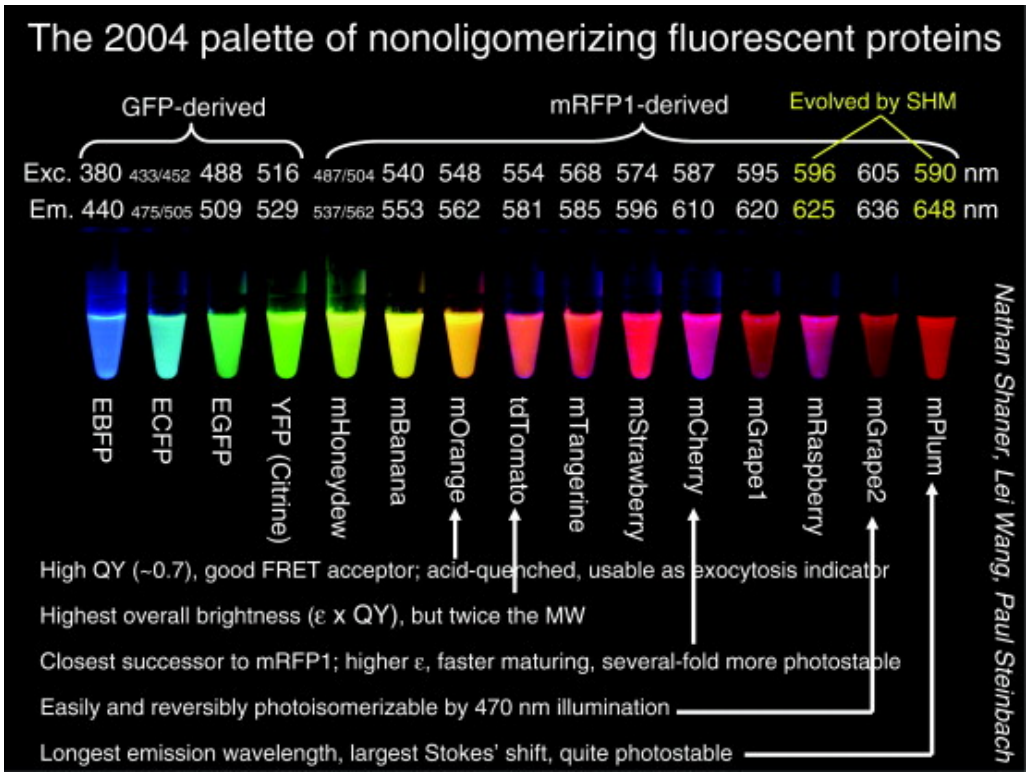


Figure. 1.4. A figure showing the breadth and variety of available fluorescent proteins, and some of the many applications available. This image contains only a handful of the full array of the full spectrum of available fluorescent molecules. These are specified as nonoligomerizing, meaning they do not form oligomers, which is important as oligomer formation would change the fluorescent properties and imaging potential of the molecules. Fluorophores vary in their wavelengths, brightness, photostability, photoisomerisibility (required for photoswitching) and duty cycle, and propensity for donating or accepting electrons from nearby molecules. Having a large variety of available combinations of these properties ensures a wider array of possible experiments. For instance, blinking or photoswitching properties, as well as brightness, are usually essential for SMLM. Image from *Building and breeding molecules to spy on cells and tumors* [5].

and out-of-focus light can dominate the image. In this regard, confocal microscopes offer a substantial advantage over widefield systems. This is due to their utilisation of a point-scanning laser beam to illuminate the sample, and a pinhole aperture to ensure that only in-focus emitted light is imaged [8]. The dwell time of the illuminating laser can be adjusted for faster scanning but weaker illumination and vice versa, resulting in a flexible modality that can be used for thick tissue slices or live-sample imaging. The 'point' that is scanned through the sample has a certain dimensionality that is limited by the diffraction limit (section 1.2.2) and affects the resolution of the final image. A major advantage of the point scanning system is that 'layers' of the sample can be scanned in sequence allowing for sectioning – that is, the acquisition of 3D datasets with significantly higher z-resolution and z-range than in the widefield set-up. It should be noted, however, that confocal resolution is not isotropic, being poorer in the axial than the lateral direction.

1.2.2 Resolution & the diffraction limit

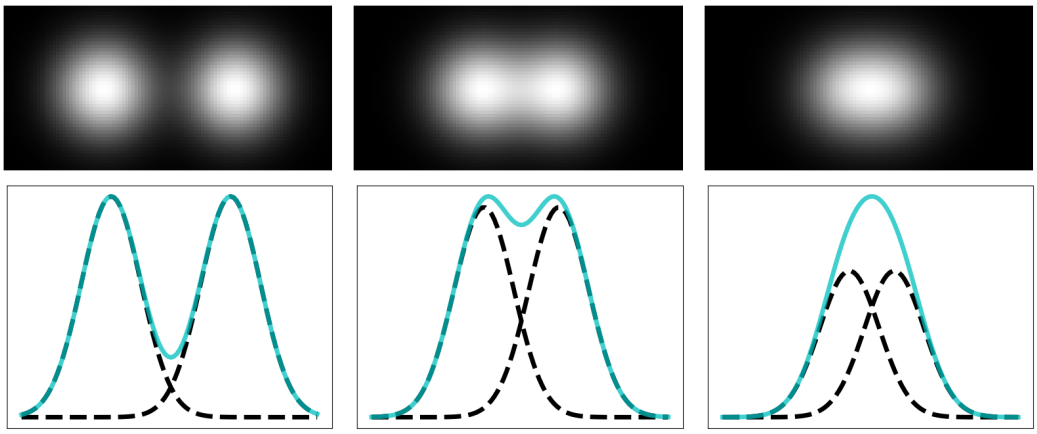
'Resolution' simply means that **there is enough information present to distinguish between multiple entities**. In microscopy, two molecules are resolved if their separation is sufficiently greater than their respective sizes when imaged. In this context, we might say resolution refers to the sharpness of imaged structures. Optical systems such as fluorescence microscopes are 'diffraction-limited'; this means that embedded into their physical functioning is a maximum attainable resolution. That is, there is a proximity beyond which two point-like molecules cannot be visually distinguished from each other in an image, regardless of the specific parameters of the microscope – the 'spot' to which a light signal can be focussed down has a minimum possible diameter. This occurs because of the quantum mechanical uncertainties in a photon's direction, position, and consequently, origin. These are implicit in the 'wavefunction' describing a photon's trajectory, whose form enables a photon can interfere with other photons, and with itself. The Huygens principle can be used to visualise this concept: every point on the 'crest' of a 'wave of light' (a wavefront) can be physically treated as its own point source of diffracting light. These secondary wavelets propagate in all directions, and interfere with each other. The quantum uncertainty of light in a microscopic system is mapped according to the 'point spread function' (PSF), whose shape and size is determined by the light's wavelength, and the system of lenses used to magnify and focus it. Thus, even in a perfect microscopic system with no aberrations, asymmetries, or distortions, a 'point' source of light, once imaged, appears as a central maximum surrounded by hazy concentric rings known as an 'Airy Disc' [9]. The width of the Airy Disc is lim-

ited to a minimum determined by the wavelength of light, which incurs a minimum distance below which the separation of two points cannot be resolved. This distance is called the 'diffraction limit' or 'Abbe limit' for the scientist who quantified it as:

$$d = \frac{\lambda}{2 \cdot n \cdot \sin(\theta)} = \frac{\lambda}{2 \cdot \mathrm{NA}} \tag{1.3}$$

Where d is the minimum resolvable distance, and λ is the wavelength of light being imaged (other terms from equation 1.2). It should be noted that the above equation is Abbe's original formulation [10], in which the criterion of resolvability is 0.5 times the ratio between the wavelength of the light and the numerical aperture. However, the value of this number varies according to the criterion of resolvability used. The commonly employed 'Rayleigh criterion' of resolution, for instance, defines the criterion of resolvability as 0.61 [11], when the central maximum of one Airy disc is aligned with the first minimum of its neighbour's. This is because the point at which two objects are resolved from one another is ultimately a human judgement, and this is reflected in the variability of the constant (Figure 1.5). However, the Rayleigh and Abbe criteria constitute widely used and accepted values for resolution criteria – this is important so that researchers can consistently make judgements about the amount of information that is available in an image.

If NA and λ are variables, then equation 1.3 only describes a 'limit' if their values are tightly constrained. NA has a 'hard' limit, imposed by a) the refractive index of the highly transluscent immersion medium (of which there is a fairly limited selection), and b) the sine of the half-angle over which light can be transmitted by the objective, which must be less than $\sin(90°) = 1$ *a priori*. λ has a 'softer' limit, in that though a wide range of wavelengths can theoretically be imaged, not all of them can be used in any context. This is because the energy of a photon of light is inversely proportional to its wavelength, in accordance with the quantum mechanical Planck relation $E = h/\lambda$. Thus, illuminating cells with (ultra)violet or shorter light wavelengths will quickly lead to cellular degradation, making any resulting micrographs worthless if the intention is to study their native structures and behaviours. Because of these two limits, fluorescence imaging offers a maximum lateral resolution of roughly 250 nm when imaging cells, even fixed ones. Therefore, the use of typical widefield and confocal far-field techniques to resolve the structures of many features of known organelles is physically implausible.



(a) Resolved **(b)** Resolution Limit **(c)** Unresolved

Figure. 1.5. Simulated pairs of PSFs to visualise the resolution limit. Here, the PSFs are approximated by a Gaussian function. The 2D images of the fluorescing molecules (above) were visualised as 1D plots by projecting them vertically and plotting their intensity. The black dashed lines represent the individual PSFs, while the cyan lines depict their summed signal. In (a) the two fluorescing molecules are completely unambiguously separated because the FWHM of each PSFs is much less than the distance between their maxima. (b) exhibits a 'limit case' wherein any further proximity would result in the two molecules becoming indistinguishable. This is of course a judgement call, which is sensitive to the definition of resolution employed (e.g. Abbe, Rayleigh, Sparrow). It is important to note that in this example, though the molecules can be individuated, their locations are biased towards their mutual centre due to the summing of signal between their peaks – that is, the maxima of the summed PSFs (cyan) are slightly closer together than either would be alone (black, dashed). In (c), this has become so severe that the two PSFs give way to single, enlarged PSF, and the number of emitting molecules has become completely ambiguous.

1.3 Super-Resolution

1.3.1 Nanoscopic fluorescence microscopy

Over a hundred years after the original quantification of Abbe's limit, attempts to overcome it and enter the world of nanoscopic super-resolution imaging began to bear fruit. The 'big three' archetypical fluorescence super-resolution methodologies are stimulated emission depletion (STED) [12], structured illumination microscopy (SIM) [13, 14], and single-molecule localisation microscopy (SMLM) [15]. These methods differ in many of their practical considerations but share a core aspiration, which is to extract more information from the point-spread function than is natively present. Until the development and widespread uptake of these technologies, imaging below the visible light limit microscopy was confined to electron microscopy (EM). This is because electrons have a much shorter De Broglie wavelength than photons, meaning that their diffraction limit is much smaller and they can resolve much smaller structures. However, electron microscopes have many disadvantages when compared with fluorescence microscopes. They are very expensive to buy, use, and maintain when compared with fluorescence microscopes. Effective usage of these machines also is highly technical, requiring much more training and time to produce an image [16]. Because of this, EM is considered a very 'low-throughput' method compared with FM, meaning scientific conclusions must generally be drawn from smaller sample sizes. Additionally, sample preparation for electron microscope is generally much more difficult, error-prone, and abrasive. Biological samples cannot be imaged with EM alive and certainly not *in vivo*, as they must often be cut into razor-thin slices and stained with heavy metals. Ultimately, EM's only major advantage over FM is the attainable resolution, meaning super-resolved FM presents an enormous step forward in research as well as medical diagnosis, allowing 3D, live-sample imaging of sub-diffraction limit structures with much more accessible modalities [17–19]. STED and SIM can be thought of as achieving this using innovations in illumination, with STED drawing particular inspiration from the confocal microscope. SMLM and its descendent methods are more the products of innovations in image processing, labelling, and tuning fluorophore behaviour [20–23]. A diagram outlining the modal differences between these three methods is displayed in Figure 1.6.

In STED, the point-scanning excitatory laser that causes fluorescence is supplemented by a structured depletion laser (often doughnut-shaped) which selectively de-excites the fluorophores in the surrounding area, confining the imaged fluorescence to an area that is smaller than the diffraction limit (Figure 1.6B) [24]. As with

confocal microscopy, the parameters of the STED lasers – including the dwell time of the scanning point, and the intensity of each laser – can be fine-tuned, meaning that the trade-off between spatial and temporal resolution, as well as the functional imaging depth, can be managed to suit the experimenter's needs. It attains resolutions in the area of ∼50 nm [25], and with a sufficiently short dwell time, can even achieve live-cell imaging [26]. STED's primary shortcoming it its tendency to bleach fluorophores quickly, due to its scanning the sample with two extremely powerful lasers, going over each area multiple times due to the imaged area being much smaller than the illuminated area. Aside from this, it requires a highly specialised microscope, limiting its accessibility somewhat.

SIM works by doubling the amount of information available in reciprocal, or Fourier space, before performing an inverse Fourier transform on this data to retrieve the extra information. This effectively doubles the resolution of the final image. We can think of the diffraction limit as a sort of 'low-pass filter' on the Fourier transform of the image data, meaning that higher-frequency information – meaning the sharper, higher resolution structures in the image – cannot propagate through. The central thesis of SIM is that a low-frequency illumination structure, such as Moiré Fringes (Figure 1.6A), can reveal higher-frequency structure in the illuminated object, which can be retrieved from the Fourier domain via transform. There are many ways to implement SIM, but most boast a two-fold increase in resolution over the diffraction limit (to around ∼120 nm) at live-sample-compatible temporal resolution [28, 29] and extremely low phototoxicity. Though it generally requires a fairly specialised set-up and only offers a modest increase in resolution when compared with the other two super-resolution methods, its suitability to the live-sample and 3D imaging contexts makes SIM an exceptionally valuable and flexible tool.

1.3.2 Principles of localisation microscopy

SMLM was independently conceptualised and achieved by three separate research groups in 2006 as stochastic optical reconstruction microscopy (STORM) [30] photoactivated localisation microscopy (PALM) [31] and fluorescence photoactivation localisation microscopy (fPALM) [32]. Among its peer super-resolution FM techniques (SIM and PALM), SMLM is unique in its ability to resolve exceptionally super-resolved distances of 10-30 nm while – at the most basic level – requiring no specialist hardware [33]. Instead, SMLM software typically uses standard widefield fluorescence microscopes, coupled with computational fitting to approximate the properties of individual PSFs and discern their locations. The centroids of the PSFs

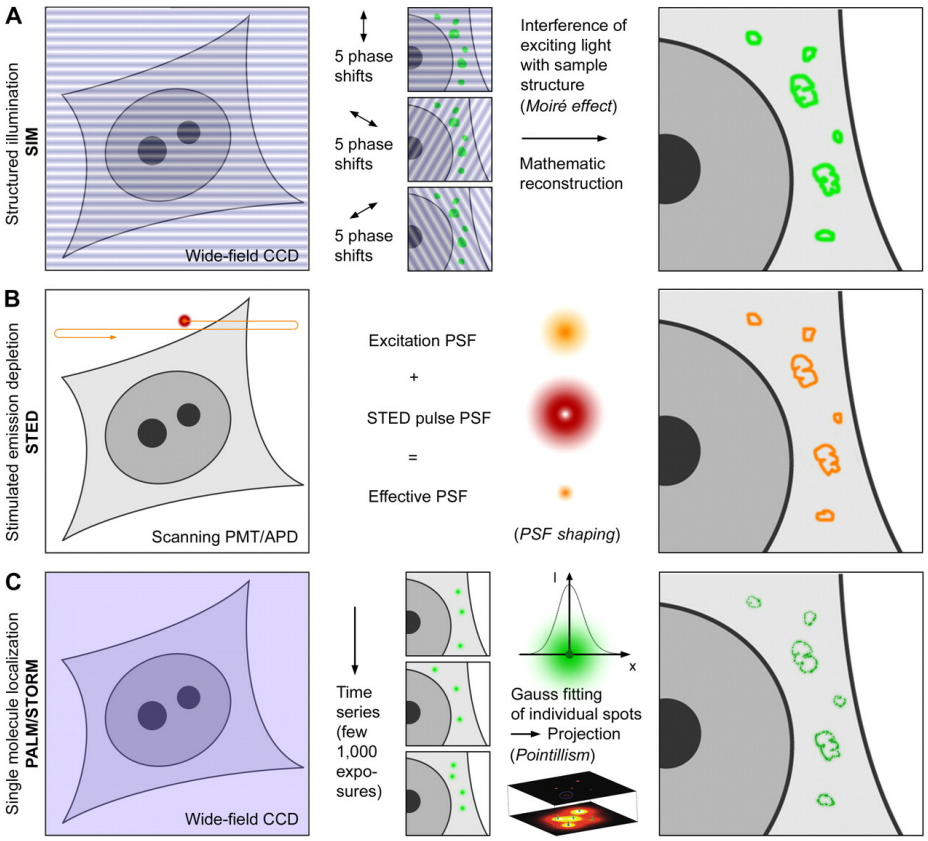


Figure. 1.6.
Super-resolution imaging principles. (A) In SIM the sample plane is excited by stripe patterned widefield illumination. The combination of the sample information with the stripe pattern generates moiré fringes, which are recorded (as part of the image) by a camera. The high spatial frequency sample information present in the moiré fringes is then retrieved mathematically, using a series of images which are phase-shifted and have the grating at diferent angles. A super-resolution image using this information is then reconstructed. (B) In STED microscopy the focal plane is scanned with two overlapping laser beams. The first laser excites the fluorophores, while the second longer wavelength laser drives the fluorophores back to the ground state by the process of stimulated emission. A phase plate in the light path of the depletion laser generates a donut-shaped energy distribution, leaving only a small volume from which light can be emitted that is then being detected. The PSF is therefore a volume smaller than the diffraction limit, resulting in a super-resolution. (C) Single molecule localization microscopy takes a series of images, where in each image only a small proportion of the fluorophores are emitting light. This is achieved either by photoactivation, photoswitching, triplet state shelving, or blinking. These fluorophores are detected as diffraction-limited spots, whose lateral position is determined with very high accuracy by a fit, most often of a Gaussian PSF. Single molecule positions from several thousand raw images, each with a different subset of emitters, are then used to generate a reconstructed image. Figure from *A guide to super-resolution fluorescence microscopy* [27].

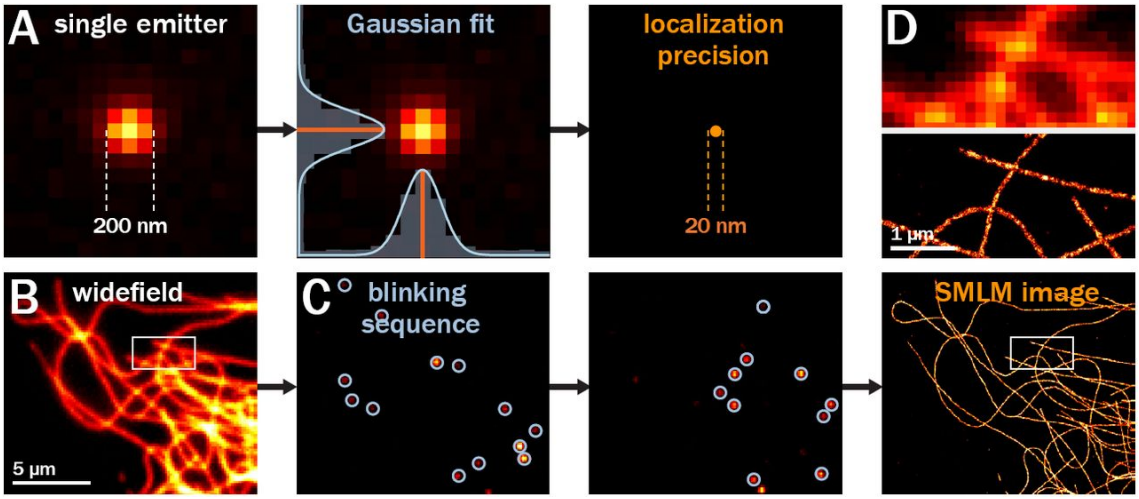


Figure. 1.7. Operating principle of single-molecule localisation microscopy. In standard FM, the size of the PSF has a minimum diameter according to the diffraction limit (a) However, a more precise rendering of the molecular location can be obtained by using gaussian fitting. This means that the coordiantes of a single molecule can be computed to a greater precision than the diffraction limit because subpixel displacements in the molecule location can be quantified from alterations to the pixel values. Without localisation, the diffraction limit means that finer cellular substructures can't be resolved due to ambiguities in the fluorophore positions arising from PSF overlap (b). If the fluorophores in an image can be made to blink independently, then the molecules can be localised from a time sequence (c) by approximating each PSF with a gaussian. The structre can then be algorithmically reconstructed from these localisations to produce a super-resolution image. Figure adapted from *About samples, giving examples: Optimized Single Molecule Localization Microscopy* [34].

can then be determined computationally, and a super-resolution image can be reconstructed from those centroids by treating the fluorophores as point-like emitters. Of course, such images cannot be reconstructed from a single widefield image, since the overlapping PSFs within would make centroid localisation impossible. SMLM circumnavigates this by exploiting the phenomenon of 'photoswitching' or simply 'blinking', whereby the sample's physical conditions are tweezed to encourage fluorophores to remain in a stable dark (non-emitting) state, but with a small probability of 'switching on' for an instant.

A high frame-rate image series – typically to the order of 10^4 images – is taken, with most fluorophores in each frame remaining in the dark state. So long as the simultaneously fluorescing molecules are further apart than the minimum resolvable separation given by equation 1.3, they can be accurately fitted and localised in each frame, allowing for a full super-resolution image reconstruction from a large enough image stack [30]. The time normally needed to obtain enough images for a clean SMLM reconstruction is high when compared with other optical imaging methods [33]. Thus the technique can be thought of as a way to trade temporal resolution for

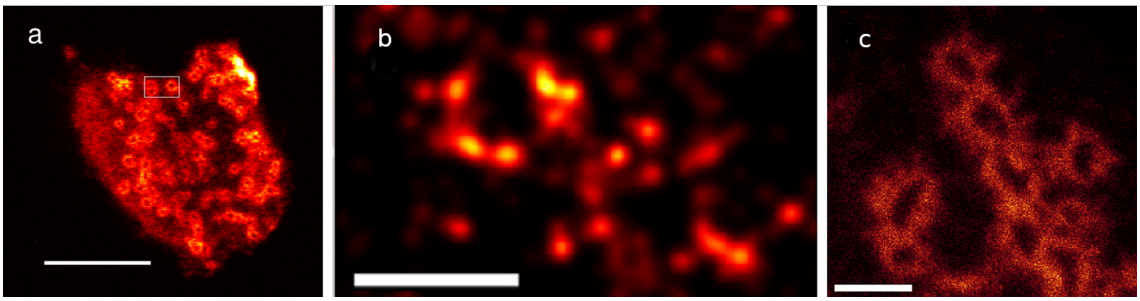


Figure. 1.8.
Effectiveness of SMLM when imaging the actin-associated protein ring of podosomes, an almost-2D structure with a diameter of $\sim$ 500 nm. Vinculin imaged in fixed cells using confocal microscopy (a). A SMLM reconstruction of a sub-region is displayed in (b), and a STED image of podosomes is displayed in (c). The SMLM reconstruction exhibits a very high resolution, but also shows the severe impact that undersampling can have on the technique. The apparent resolution of depicted structures is very high, yet the full podosome 'rings' appear punctate and discontinuous due to missing structure arising from undersampling. Scale bars are (a) 10 µm, (b, c) 1 µm. Figure adapted from *Imaging cells at the nanoscale* [35].

increased spatial resolution. Because of this, it is easily misconstrued as a method that is exclusively well-suited to fixed-sample imaging. This can be seen in SMLM images of the podosome [35] in Figure 1.8. Here, an attempt to take an image with SMLM over an insufficient timescale has resulted in a visibly undersampled final image that could be missing important structural information.

1.3.3 Blinking and Bleaching of fluorophores

The development of the three SMLM techniques differ in a number of experimental details, though the underlying principles of photo-switchability and localisation are the same. While PALM and fPALM used photo-switchable protein expression as their source of fluorescence [31, 32], the original STORM developers employed immunohistochemistry to image using photo-switchable dye conjugation pairs [30]. Research prior to these first SMLM experiments reported that red laser light, which normally produces fluorescence in cyanine dye Cy5, may also switch it to a stable dark state, but that this state can be de-stabilised back into fluorescence with green laser light. Previously, the STORM research group also found that the power of green laser light required to perform this 'switch on' was three orders of magnitude lower if the Cy5 dye was in close proximity to a Cy3 dye molecule due to a undetermined mechanism functionally analogous (but not physically identical to) FRET [4, 36].

Two years later, research found that this 'activator fluorophore' is not a necessity for STORM imaging, and the technique was refined into 'direct' STORM

(dSTORM), so named for its non-reliance on a meticulously balanced and separated conjugation of dyes. The experiments demonstrate that some organic dyes' blinking (most notably Cy5 and its cousin Alexa Fluor® 647) can be easily controlled by tuning laser power. The probability with which these fluorophores 'turn off' scales linearly with the 'excitatory' laser power (the 647 nm laser). Similarly, the probability with which they 'turn on' scales linearly with the 'activation' laser power (the 568 nm laser) but without a Cy3 conjugate, the laser power required is 200 times higher to achieve the same effect [20, 37, 38]. The advent of dSTORM enables a great increase in the accessibility of SMLM. Combining the operational simplicity of dSTORM with the freely accessible, open-source ImageJ plugins on offer to perform image reconstruction, such as QuickPALM [39] and ThunderSTORM [40] makes the potential accessibility of a tenfold resolution improvement over the diffraction limit is very high. Here, it should be noted that the increased accessibility of SMLM presented by a range of localisation algorithms to choose from means that researchers need to be careful to maintain consistency across their experiments by ensuring their fitting algorithms are consistent when comparing different imaging protocols. Different localisation software allow for different adjustable parameters during reconstruction and use different algorithms for calculating e.g. the uncertainty of the Gaussian fit of the molecule.

A major issue with the increased laser power required to employ dSTORM effectively is its rapid photobleaching effect. Excited electrons have a tendency to react with molecular oxygen, degrading them and removing their fluorescent properties – a process accelerated by increased laser power. If the fluorophores in a sample bleach too quickly, the quality of SMLM images will plummet due to their low temporal resolution, since ordinarily a large stack of images must be obtained. This is believed to occur at least in part because of the possibility of electron intersystem crossover from the singlet excited state to the triplet state, though there is some debate about whether this is the only mechanism at play.

The 'dark' (non-emitting) state that a fluorophore undergoes during a STORM experiment is thought to occur because the electron is forced into the triplet state. The theory suggests that because electrons in the triplet state have a lifetime of the order 10^6 times as long as the singlet state due to the $T_1 \rightarrow S_0$ relaxation transition being spin-forbidden by the Pauli Exclusion Principle [41], they have a proportionally higher chance of reacting with the constituents of their chemical environment, molecular oxygen in particular. These electrons then stop participating in the excitation-emission cycle, resulting in irreversible covalent alterations in the fluorophore which contribute to the nullification of its fluorescence potential (see

Figure 1.3) [42].

Reactions with molecular oxygen form reactive oxygen species such as radicals, which are highly energetic and correspondingly destructive to a sample and its dyes, a phenomenon known as phototoxicity. The problem is two-fold in live-cell imaging, where living cellular substructures also are under threat from reactive oxygen species [43]. There are several means through which photobleaching and associated complications are minimised: these include the use of robust fluorophores, which have a lower tendency to bleach; more sensitive photon detectors, which minimise the activation laser power needed by reducing the photon emission required for image generation; enzymatic oxygen scavengers, which catalyse irreversible reactions between molecular oxygen and other compounds to form more innocuous chemicals; and quenchers, which prematurely de-excite electrons non-radiatively [44]. Glucose oxidase with catalase (GLOX) has been found to be an effective enzymatic oxygen scavenger, while primary thiols, such as 2-aminoethanethiol (MEA), are employed as 'dynamic quenchers' [45], meaning they de-excite dyes by forming complexes with them in the excited state. In forming an adduct with excited cyanine dyes, primary thiols not only reduce the rate of photobleaching, but encourage healthier blinking dynamics by shepherding them back into the dark state [46]. An assessment of 26 organic dyes found that including both GLOX and an experimentally determined appropriate thiol (usually MEA) in imaging buffers near-consistently reduced the rate of photobleaching and saw a corresponding increase in image quality [47]. Of particular note is Alexa Fluor® 647, a robust dye used in much of the dSTORM imaging in our experiments, which saw a performative increase from 'rapid bleaching (no image)' to 'robust photoswitching (high quality image)' with the addition of both GLOX and MEA in imaging buffer [48].

A notable shortcoming of employing GLOX as an enzymatic scavenger are the products of the reaction it enables: oxygen and glucose react to form hydrogen peroxide, a powerful oxidizer and hydroxyl radical precursor, and gluconic acid, the build-up of which leads to rapid acidification of the buffer [49]. The latter can adversely affect dyes' blinking kinetics as pH has been shown to correlate positively with many cyanine dyes' ability to switch to the dark state [46]. Because of this, imaging buffer should not be prepared until the sample and microscope concerned are prepared for an imaging session: in our experiments, a single STORM imaging session in GLOX buffer never exceeded 3 hours. Hydrogen peroxide is removed with the addition of catalase solution, which catalyses its conversion to oxygen and water.

It should be noted that blinking or photoswitching are not necessarily requirements for SMLM. A newer variant, DNA-PAINT [50] utilises short DNA probes

that are labelled with a fluorescent dye. These bind to specific complementary target DNA sequences that have been attached to the sample. The binding affinity for single-stranded DNA to short complementary sequences is such that these can unbind easily, meaning that a 'blinking' dataset can be achieved via the simple binding/unbinding of the DNA probes. This is because the unbound probes will diffuse out of the focal plane (and ideally, the illumination volume, which can be achieved using TIRF or light-sheet illumination). However, DNA-PAINT can be prohibitively slow to implement, since increasing the activation density is equivalent to increasing the concentration of fluorescent probes in the sample, which can result in high background.

1.3.4 Accessibility of SMLM

The advent of dSTORM constituted a significant step forward in terms of the accessibility of SMLM, because basic experiments on fixed cells could now be performed with nothing more a standard fluorescence staining protocol and a buffer recipe [20]. Though the raw hardware requirements required to perform an SMLM experiment are minimal when compared with its counterparts, STED and SIM, the various mechanisms that govern resolution and image assessment are more complex. They are governed not only by the illumination parameters (as in STED), or the physical limit of how much extra information can be extracted from the Fourier domain (as in SIM), but by the mathematics of localisation [51, 52]. This has major implications for SMLM's uptake, because experimenters who may have little knowledge of mathematics and physics will need to make a number of considerations. For instance, what is the resolution of their final reconstruction – or more appropriately, of a local region of interest (ROI) – and does the resolution increase relative to the diffraction limit actually make using SMLM worthwhile? Has their image has been reconstructed correctly (section 1.4.3)? Which localisation algorithm (of which there are a great number) is best-suited to their experiment? Is it compatible with the 3D localisation modality they have to hand, if needed (section 1.6)? How long will it take to process the data?

In other words, despite the advances represented by modalities like dSTORM and its minimal hardware requirements, there is a long way to go with regards to bringing SMLM into wider usage. There are broadly two strands of development in this area: one involves developing localisation software and data pre-processing techniques that are 'foolproof' and effective. The other is directed at the development of easy-to-use assessment tools and image post-processing steps that allow researchers easy access to experiment optimisation and information extraction. Both are essential and have

seen advances in recent years, but it is my view that SMLM is still too often seen as the dominion of physicists or senior researchers. Yet, two of its greatest assets as a super-resolution technique, alongside its outstanding attainable spatial resolution, are its availability and flexibility. The work in this thesis merely presents three instances of techniques developed by my research group with the aim of addressing the level of specialism that is often thought of as being required to adopt SMLM as a main workhorse of biological imaging. It is my hope that these techniques will ultimately be taken up and used by a broad church of bioscience researchers for their efficacy and ease of use.

1.4 What makes a good super-resolution image?

1.4.1 Resolution as localisation precision

In diffraction-limited microscopy, the criterion of PSF separability is what determines resolution, in accordance with equation 1.3. An SMLM micrograph, however, is not populated with PSFs, but what we might call 'localisation error functions' (LEFs). These represent the precision with which the PSFs were localised, which we can think of as a local proxy measurement for the resolution. This is much more difficult to quantify accurately than equation 1.3, because the localisation precision is not governed exclusively by the microscope hardware and the wavelength of light. We can derive the mathematical upper limit of the localisation potential for a given PSF from the Fischer information from the fit – more simply, it is determined by the curvature of the fitting function [53].

$$\Delta = \frac{\lambda_e}{2\pi n_a \sqrt{\gamma A t}} \tag{1.4}$$

Where γ denotes the efficiency of the optical system, A the emission rate of the fluorophore and t the total acquisition time for the 'blink'. However, in an experimental scenario, the precise shape of the LEF is also highly responsive to a number of other factors, whose impacts cannot be quantified precisely:

1. The specificities of sample preparation, including the fixation protocol [54], and the 'linkage error' – the distance between the fluorophore and the molecule of interest. This last quantity ranges from ∼2-10 nm for fluorescent proteins, to ∼10-20 nm using direct antibody labelling or proteinic ligand binders such as SNAP-tags, or even up to 40 nm when using indirect antibody labelling [55]. Thus, it is possible for the labelling method to become a limiting factor for

the resolution and the localisation accuracy. The position of the fluorophore can additionally be influenced by the fixation method used – some fixation methods are known to alter cellular structures [56].

2. The signal-to-noise ratio (SNR) of the acquisition stack, which is significantly reduced in a SMLM dataset. Because only the light of a single fluorophore per localisation is available for reconstruction, a high background can greatly reduce localisation accuracy [57]. While camera noise [58, 59] and sample imperfections contribute to this problem and can be particularly tricky to address, the greatest contributor often comes from out-of-focus light, or background. Even a small amount of background can become very quickly detrimental to an acquisition due to the low signal strength in SMLM. The severity depends on the fluorophore used and the sample geometry, but is also highly sensitive to the choice of illumination mechanism. Thus, in FM, this light is most easily reduced by simply illuminating a slice of the sample, so that the image is axially sectioned. There are a number of ways to achieve this, but by far the most accessible in terms of hardware is Total Internal Reflection Fluorescence (TIRF). The technique exploits the fact that a light wave's transmittance through a two-media-interface of differing refractive indices – in this case, the sample and its culture dish – diminishes with increasing angle of incidence in accordance with Snell's law. When incident at a sharp enough angle (greater than the 'critical angle' θ_c), all illuminating laser light is reflected off the bottom of the sample as it has a lower refractive index. However, in this case an 'evanescent wave' is generated at the substrate-sample interface [60]. While having the same wavelength as the original laser beam, the evanescent wave decays exponentially in intensity with propagation, such that it is sometimes unable to penetrate more than 100 nm into the sample. Tweaking the 'TIRF angle' of laser beam incidence thus gives an experimenter sub-diffraction limit sectioning power, even in an otherwise conventional widefield system. In confining dye excitation to the sample surface, the signal of interest is not distorted by fluorescence originating deeper within the sample, lowering the background signal by as much as 2000 times [61]. Additionally, the evanescent field can be many times more intense than the incident light in cases where the angle of incidence is within 10° of the critical angle, reducing the power of incident light required for strong fluorescence and boosting the signal-to-noise ratio even further [60]. It also reduces the amount of light that the sample is exposed to overall, reducing the phototoxicity [61] For its high accessibility, TIRF is very popular in SMLM, though it only allows imaging within roughly 100 nm

of the coverslip, limiting its practical applications somewhat.

3. Camera pixelation, which simply determined the number of data points available to the fitting algorithm. Larger pixels mean fewer data points and thus a lower localisation precision and resolution. However, larger pixels also receive more light, resulting in a higher SNR [53], which is also an essential factor in localisation precision. Most SMLM systems use pixels in the range of 100-160 nm, because this range satisfies the Nyquist sampling criterion that pixel width should be half that of the PSF.

4. Stage or sample drift – this is a particularly nefarious source of error in SMLM, as it heavily impacts both precision and accuracy, and is exacerbated by the longer acquisition times normally needed to get high-quality data. Lateral drift is generally corrected retroactively with various analytic methods, most commonly cross-correlation either with fiducial markers [62, 63] or sample fluorescence [64], while axial drift is most often actively compensated for by monitoring light fluctuations at the interface between the cover glass and the sample. Drift can also be corrected actively in all three dimensions by implementing feedback loops into the acquisition [65].

Because of these factors, there is no scientifically watertight, or even by-and-large agreed-upon way of defining resolution in cases where the diffraction limit has been broken. However, the above factors which contribute to localisation precision have been aggregated in the equation derived by Thompson et al. [57]:

$$\Delta^2 = \frac{\sigma + p^2/12}{N} + \frac{4\sqrt{\pi^2}\sigma^3 b^2}{pN^2} \tag{1.5}$$

Where Δ^2 is the two-dimensional spatial uncertainty in fluorophore localisation, σ is the standard deviation of the PSF profile, N is the number of photons collected per fluorophore, p is the pixel size, and b is the average background noise. The first term describes the uncertainty introduced by the Airy Disc itself: it decreases with N because an increase in the number of photons corresponds to an increase in the amount of information from which to determine a common origin, and increases with p because larger pixels equate to a larger area from which each photon could have originated from. The second term, which is usually much smaller, represents the error introduced by background noise [57]. In terms of its recognition of the contributing factors to the reduction of localisation precision, the equation is fairly comprehensive. However, should be noted that it is considered to under-report the position uncertainty of fluorophores [53].

Despite its central role in governing the resolution of SMLM images, localisation precision is complex to calculate. As we have seen, it is highly sensitive to a wider array of attributes than measures of diffraction-limited resolution. These attributes are often extremely difficult to quantify accurately, and the size of the LEF will always vary from molecule to molecule. This has two major implications. Firstly, resolution in SMLM must be measured as a quality of the image, not of the hardware. Secondly, one should be prudent when speaking of the 'global' image resolution of a SMLM reconstruction. In spite of this last point, one method for making global resolution measurements is still popular among many researchers for its intuitive methodology and ease of use.

1.4.2 Resolution as self-similarity

The most popular approach for obtaining a measure of image quality is the Fourier Ring Correlation (FRC) approach. It splits the localisations randomly into two datasets, then compares the similarity of those datasets over different length scales. The correlation of the two drops as the length scale decreases, and the length scale at which the correlation drops below a certain level that is said to be the resolution [51]. It thus uses self-similarity as a proxy for the global resolution measure for the image.

It should also be noted that density of imaged fluorophores will have a direct impact on any resolution measure that uses this criterion. Higher fluorophore density in finer structures will improve the cross-correlation measure of the Fourier rings towards the resolution threshold. This means that resolution is dependent on both the labelling density and the structure being imaged. Repeated localisations of the same fluorophore can lead to a falsely good resolution measure unless this effect is taken into account in the analysis [51]. A more recent alternative to FRC called 'decorrelation analysis' measures the correlation of the Fourier transform of the full image with its normalised counterpart. This method uses phase information and only requires one image, allowing direct comparison of SMLM with those other super-resolution techniques for which localisations cannot be randomly split in two [66].

These methods contribute to SMLM image assessment tremendously by offering an approach that does not involve a mathematical breakdown of the many contributing factors to localisation precision. However, when we speak of 'resolving' structures, the underlying implication is that the structures are known to have been *correctly* localised. It might be more correct to say that FRC and decorrelation analysis measure image *sharpness*, because molecule mis-localisation is indeed a major

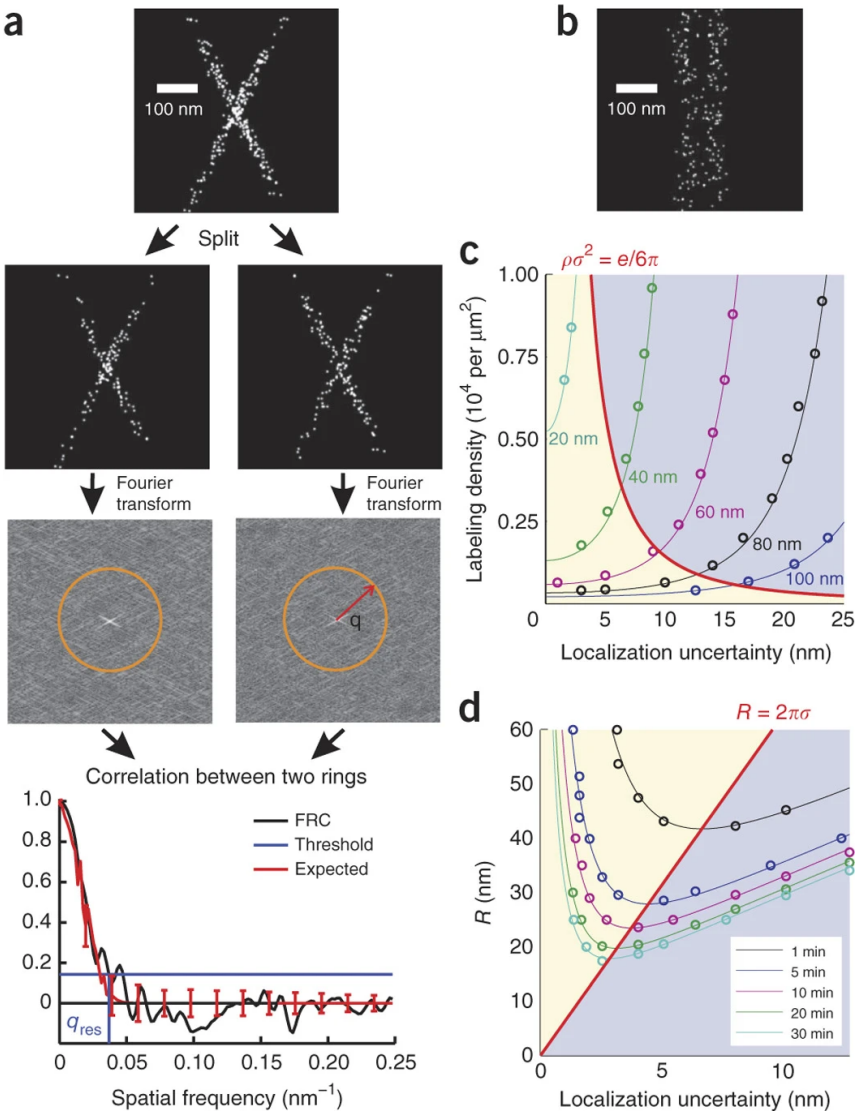


Figure. 1.9. The operating procedure for the Fourier ring correlation (FRC). (a) The localisations from the reconstruction are divided randomly in two. The FRC then uses a correlatory measure of their Fourier transforms. The correlation between the two image's Fourier transforms over the perimeter of a circle of radius q outputs a decay curve measuring correlation against spatial frequency. The resolution is then taken as the inverse of the spatial frequency for which the correlation value drops below 1/7. This means a threshold value at $q = 0.04\,\mathrm{nm}^{-1}$ is equivalent to a 25 nm resolution. (b) displays a SMLM reconstruction of a line pair with separation 70 nm, a mean labeling density of $\rho = 2.5 \times 10^3$ µm^{-2}, and a localisation uncertainty of $\sigma = 7.6$ nm. (c) Constant resolution in theory (lines) and simulation data (circles) for line pairs as in b as a function of localisation uncertainty and labeling density. Regions of localisation uncertainty–limited resolution (blue) and labeling density–limited resolution (yellow) are separated by the red line $\pi\rho\sigma^2 = e/(6\pi)$. (d) shows simulation results for localisation uncertainty versus image resolution for different fixed total measurement times. Camera frame rates were varied to match the on-times of the emitter. The minima of the curves fall on the line $R = 2\pi\sigma$ that separates the yellow region (not enough emitters localised) from the blue region (emitters not localised precisely enough). Figure obtained from *Measuring image resolution in optical nanoscopy* [51].

issue in SMLM.

1.4.3 Artefacts

While precision loss decreases the resolution of the final reconstruction, bias, or accuracy loss, erodes the fidelity of the final image without being visually obvious. It is quite possible, and indeed a frequent issue in SMLM that crisp, high-resolution reconstructions do not reflect the actual structure of the sample [67, 68]. 'Artefact' is an umbrella term that can denote any apparent details of an imaged biological structure that are *not* legitimate features of the sample. Here, we use artefacts to refer exclusively to misrepresentations of an imaged sample as arising from the mislocalisation of fluorescing molecules in SMLM. The labelling and activation density, fixation [56], as well as the dimensionality of the structure [67] can lead to artefacts – having a notable impact on their severity, the nature of their effects, and their detectability. Figure 1.10 exhibits the effects of increased emitter density on acquisitions of different configurations of simulated one-dimensional filaments. Here, the apparent resolution of the HD reconstruction (Figure 1.10b) – and thus its proximity to the GT when projected horizontally – is far superior to the LD and MD reconstructions, even though these more accurately capture the sample.

Artefacts can also arise from the configuration of the sample itself. Figure 1.11 displays an example of artificial clustering on images of the cell plasma membrane. Here, the artefacts arise simply because TIRF illumination results in a rapidly decreasing excitation with distance from the coverslip. This causes non-blinking background from deeper in the sample – be it because of fluorescently-labelled structures in the cytosol, or regions of the cell membrane not sitting flush against the coverslip.

A particularly prominent source of image artefacts in SMLM is overlapping emitters. This is because localisation algorithms will fail to fit molecules accurately if their emission profiles are sufficiently overlapping. In this case, the fitting algorithm will bias two or more molecules to be localised towards their mutual centre, creating false structures in the final reconstruction called 'artefacts'. If this causes images to appear as though they have higher resolution, it is referred to as 'artificial sharpening'. This term refers to the phenomenon of multiple structures being condensed into one, or a reduction in the dimensionality of the structure – in other words, though sharpened images appear to have higher resolution, they also have missing structure. When this happens, algorithms that measure image sharpness as a proxy for resolution, such as the FRC method, will report the image as having improved resolution. However, sharper images can just as often be indicative of artefacts that lead to fictitious depictions of the sample. This means that resolution measurements

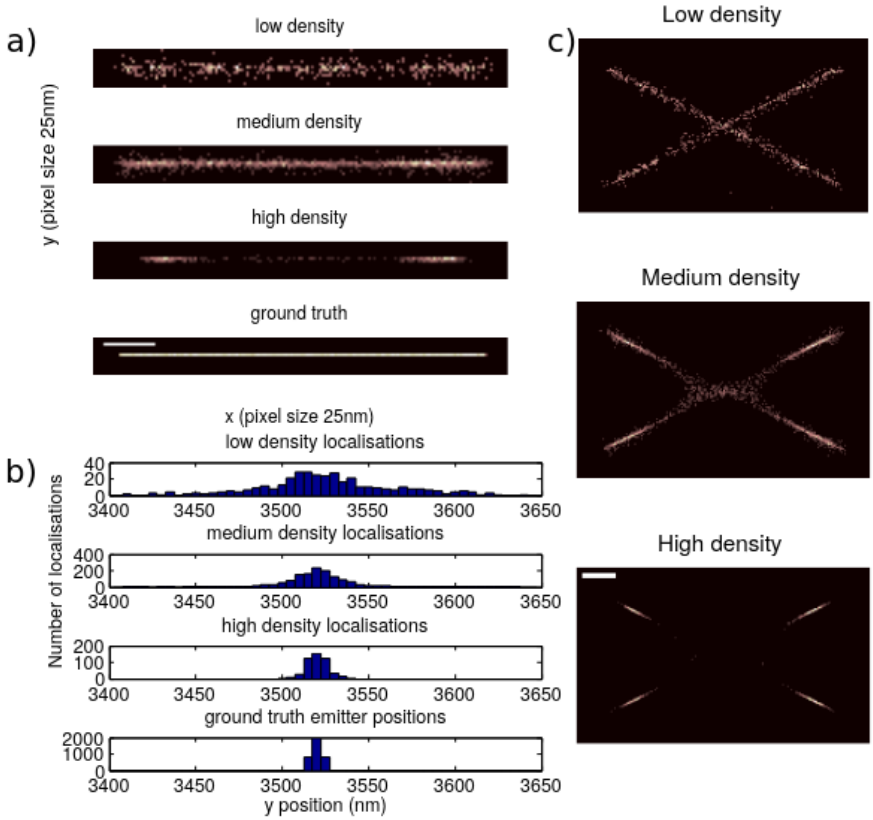


Figure. 1.10. Change in the apparent structure of a SMLM image is affected by high activation densities. (a) displays a simulation of a single one-dimensional filament reconstructed from datasets of various activation densities: LD (top, 15 μm^{-2}), MD (second, 150 μm^{-2}), HD (third, 1500 μm^{-2}). The GT is also displayed at the bottom of (a). A high degree of emitter overlap in the HD reconstruction leads to a poor reconstruction that only returns localisations at either end of the structure. (b) a horizontal projection of the distribution of localisations is plotted as a histogram. These suggest that even though the structure was rendered incorrectly, the HD localisations are on average much closer to the GT than the LD reconstruction. This points to the elusiveness of artificial sharpening: HD data can repeatably return false structure, yet may simultaneously appear to match the underlying structure better than LD, artefact-resistant reconstructions. (c) displays a simulation of a cross-shaped structure, evaluated at the same densities as the filament in (a). 'Pinching in', the premature crossover of the fibres, is clearly seen in the MD, but not the HD reconstruction, wherein the crossover region is completely absent. This indicates that sharper images can also have degraded resolution. All scale bars 500 nm. Figure from *Local dimensionality determines imaging speed in localization microscopy* [67].

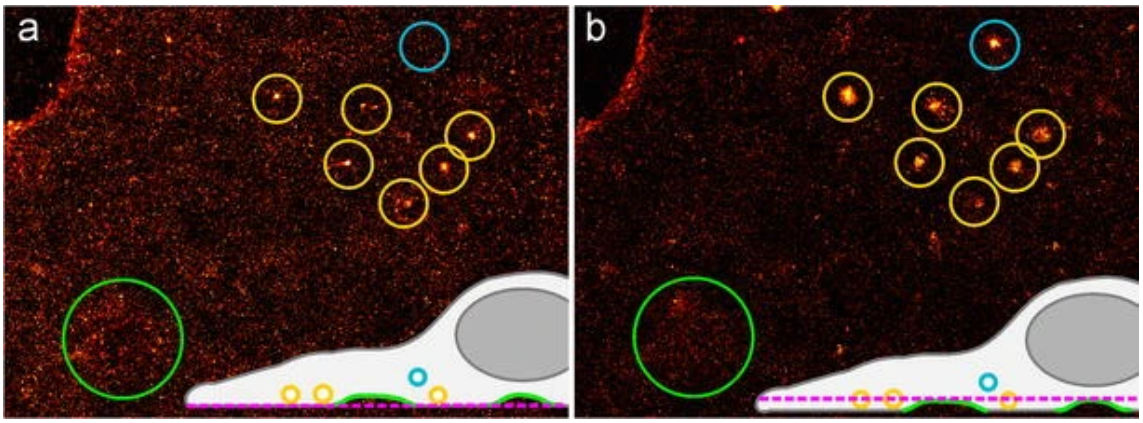


Figure. 1.11. Artefacts can arise from the sample parameters. Artificial clusters and 2D projections in the plasma membrane induced by vesicle-like structures and membrane modulations. dSTORM images of glycoconjugates in the basal plasma membrane of U2OS cells stained by metabolic labeling and click chemistry with Alexa Fluor 647. Images were recorded at an irradiation intensity of $\sim 7\,\mathrm{kWcm^{-2}}$ in TIRF (a) and epi-illumination (b) with focal planes slightly shifted (0.5-1 µm) into the cytosol (dashed magenta line). The artificial clusters in a (yellow circles) are generated due to non-blinking background signals resulting from endosome-like vesicles filled with fluorophores which are located right above the plasma membrane (yellow circles) or further up (blue circles) in the cytosol (b). Two-dimensional projections of plasma membrane modulations lead to artificial inhomogeneities in the emitter pattern (green circles). Figure from *Artifacts in single-molecule localization microscopy* [69]

such as the FRC, when taken alone, are insufficient to measure to quality of SMLM reconstructions.

There are three major reasons for which overlapping fluorophores are likely to be encountered by most researchers using SMLM techniques. Firstly, as implied by the STORM acronym, fluorophore activation is stochastic. Even if the activation density of the fluorophores is relatively low, there is no known way to ensure that they will never emit simultaneously & in close proximity to one another. Secondly, experiments that look at the correlatory or otherwise related behaviours of organelles require that the molecules of interest are stained in different colours. However, high-density data is a common occurrence when attempting SMLM with any but the most specialised fluorophores. There is a distinct lack of organic dyes and fluorescent proteins that have a 'low duty cycle' i.e. that are known to go into the dark state easily [48]. That is, the clear majority of fluorophores tend to spend enough time in the emitting state that they frequently overlap and give rise to artefacts. It should also be noted that when imaging two colours (and thus with two lasers) it can the case that the blinking dynamics of the longer-wavelength dye will be affected by the other laser. This suffers from another artefact-producing issue, photoblueing, wherin the fluorescent molecule photoconverts to a blue-shifted intermediate [70].

Thirdly, artefacts become an unavoidable issue when attempting to image live cells with SMLM (section 1.5).

1.5 SMLM in live cells

1.5.1 The problem of temporal resolution

There is one area in which STORM and PALM is ordinarily greatly outperformed by more conventional, diffraction-limited FM methodologies: the imaging of living cells. Because effective use of SMLM typically incurs such a massive temporal resolution penalty, active cells can become impossible to image, limiting use of the technique to fixed cells. This imposes a significant limitation on the kind of data that can actually be acquired using these techniques. We may be able to take precise quantitative measurements on the size and scale of structures, but we cannot do so in anything resembling a native context, or perform any kind of monitoring of cellular activity.

While faster data acquisition can be enabled with a higher frame rate [18], the challenges of achieving a low activation density in densely labelled samples [67] mean that the most natural solution is to reduce the number of frames used to generate a reconstruction. This incurs an obvious problem – if we reduce the number of frames in the acquisition, we reduce the amount of information available in the final image. There are two options for overcoming this: either the imaged structure must be interpolated from undersampled data, or the number of molecules that are 'switched on' in each frame must be increased. The former method is trickier, as it involves adding information to the image that is not there natively, meaning it typically requires some form of machine learning [71]. As a result, latter approach has seen many more techniques developed. The central observation here is that if the density of information in each frame is higher, then the number of frames required for a well-sampled reconstruction will be lower.

Unfortunately, this solution leads to a more subtle issue (section 1.4.3), that is, an increase in emitter density will lead to a reduction in image accuracy. Higher emitter densities correspond to an increased probability of emitter overlap in the unprocessed data. As we have seen, emitter overlap biases the localisation software, leading to artefacts in the final reconstruction. Maintaining emitter density at a level that does not incur artefacts makes it too slow for live-sample experiments. It should be noted here that the local sample dimensionality will additionally impact the number of frames required to generate a well-sampled SMLM reconstruction and thus the acquisition time. Surfaces will take longer to image than fibres, and fibres more than point-like or clustered structures, in accordance with the number of

localisations required *and* the increased likelihood of emitter overlap for higher dimensional structures. Figure 1.12 from Fox *et al.* [67] exhibits the effects of sample dimensionality on SMLM dataset acquisition. Here, the effects of sample dimensionality on acquisition times for localisation algorithms that analyse one frame at a time were modelled according to the equation:

$$T(x) = \frac{Qe^{a|S_\epsilon(x)|}}{a} \tag{1.6}$$

Where $T(x)$ is the time taken for a well-sampled acquisition, Q is the quality of the SMLM reconstruction defined as the activation density for an acquisition of successfully localised fluorophores, S is the sample surface, a is the probability of an emitter switching on per unit area per unit time, and ϵ is the 'exclusion zone' This is defined as the area surrounding an emitting fluorophore within which no other simultaneous emissions can occur if the fluorophore is to be successfully localised, artefact-free. $S_\epsilon(x)$ is the intersecting area between the sample surface and the exclusion zone – this is the term that causes sample dimensionality to affect the feasible $T(x)$. This equation is modelled in Figure 1.12d, e for samples for 0D, 1D, and 2D [67].

1.5.2 Overcoming artefacts

The problem of multi-colour imaging could perhaps be overcome by the development of new fluorophores that willingly enter the dark state and blink brightly [72]. But the need for live-cell imaging means that high-density data is not something researchers can simply find new ways to avoid. Rather, it demands that we develop more extractive methods for obtaining information from raw acquisition data, because the correct locations of molecules are more heavily obscured.

A Gaussian fitting algorithm can more effectively make use of information in the spatial domain by iterating over the possibility of multiple signal peaks for each 'connected component' (i.e. intensity-thresholded region), as in multi-emitter (ME) fitting [73]. Similarly, DAOSTORM extracts even more information from the spatial dimension by also examining the residuals image (the image containing only the noise that has been removed from the original) for additional, 'unnoticed' peaks [74]. These methods have been shown to reconstruct images from high-density datasets with much better fidelity than traditional single-emitter (SE) fitting [73, 74].

Alternatively, information taken from the temporal domain can be used to distinguish between signal contributions to each pixel. These methods not only utilise

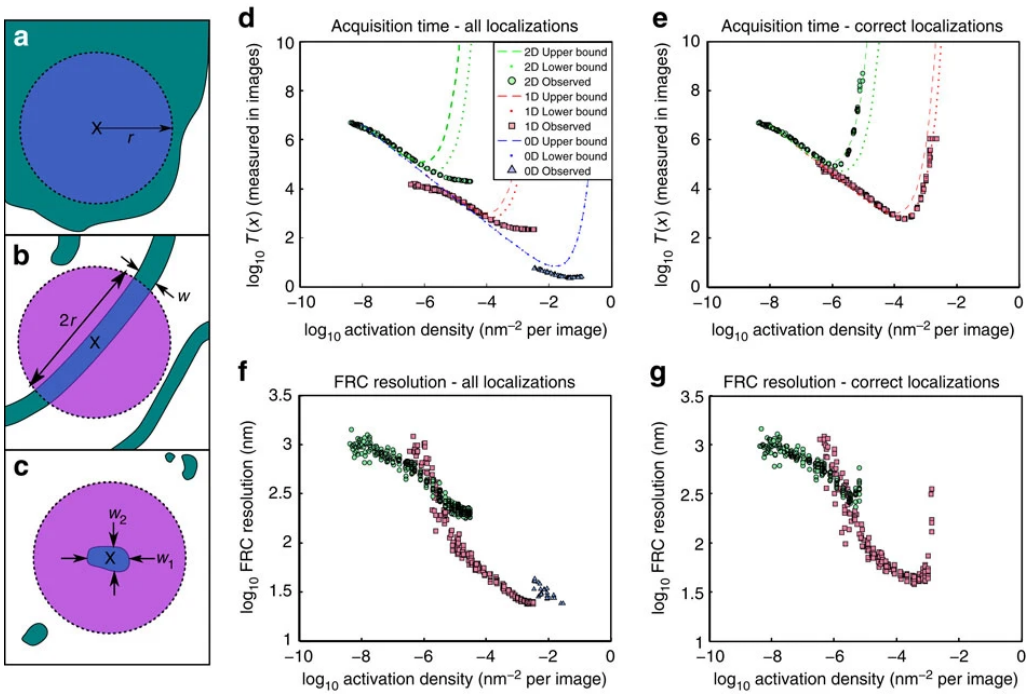


Figure. 1.12. Local dimensionality of the imaged structure impacts the acquisition time needed to achieve a particular resolution. (a-c) display illustrative examples, from top to bottom, of a 2D, 1D (i.e. filament-like) and 0D (i.e. point-like) structures. In SMLM, fluorophore emissions can be thought of as bearing an 'exclusion area' within which any further fluorophore emissions would result in artificial sharpening due to the proximity of the emissions. The structures are shown in green, the exclusion area in purple, and their intersection in blue. For a 2D sample (a), the intersection area is given by $\sim \pi r^2$; for a 1D sample (b), it is $\sim 2wr$; for a 0D sample (c), $\sim w1w2$. The the number of acquisition frames, $T(x)$, required to achieve a particular resolution varies with the activation density are shown in (d, e). Here, (e) shows only the localisations from only the 2D and 1D structures in (d) that were deemed to be good fits after re-classification, because the 0D structures were too small for user re-classification. These are shown theoretically for the estimated lower and upper-bounds of the exclusion radius, 400 nm (dashed line) and 600 nm (dotted line), and by simulation (triangles). The resolution can also be evaluated for a given number of frames using the FRC (f, g). Here, the FRC resolution is graphed for each structure dimensionality using all localisations returned by the analysis algorithm (f) and again using only those localizations classified as good (g). Figure obtained from *Local dimensionality determines imaging speed in localization microscopy* [67].

spatial information present in each frame, but couple this analysis with a quantitative assessment of the temporal signal fluctuations for each pixel. The SOFI algorithm [75] obtains molecule locations from the integrals of the second order correlation function for the signal/time trace of each pixel with its neighbours. 3B bins molecules and employs a maximum likelihood algorithm model to estimate the positions of individual molecules in each bin by taking into account the blinking and bleaching [76]. SRRF uses the divergence trace of image sub-pixels to calculate molecule locations as the peak centroids of radial symmetry [77].

Though these methods have been shown to significantly reduce the incidence of artefacts in reconstructions from medium-to-high-density acquisitions, they all suffer from the same failure mode as single-emitter Gaussian fitting. That is, once they reach a certain activation density, they will start to produce artefacts. This activation density will certainly vary between different methods, making different algorithms appropriate to different experimental scenarios. However, in terms of structure intricacy and activation density, *which* scenarios each algorithm is best-suited to cannot be ascertained from the raw acquisition data, meaning their use remains somewhat risky (if less risky than standard Gaussian localisation). Thus, they tend to fail and produce serious artefacts at sufficiently high activation densities, with images often appearing very sharp.

1.5.3 The Haar Wavelet Kernel

HAWK is an algorithm for processing high-density SMLM data that offers two unique advantages over its counterparts. First, it is a pre-processing step that employs temporal band-pass filtration to artefact-proof the dataset. This makes it entirely compatible with other localisation modalities that don't utilise information from the signal time trace, such as ME fitting and DAOSTORM. Second, the failure mode of HAWK is unique in that the penalty for increased activation density in the original dataset is not localisation accuracy, but *precision*. That is, rather than fabricating artificial structures once the emissions become too dense, HAWK pre-processed ('HAWK'd') reconstructions simply display images with a decreased resolution – not a decreased fidelity. This is because HAWK does not actually localise the molecules itself. When faced with an overlapping cluster of PSFs, other localisation algorithms make an assumption about what the underlying structure will be. Conversely, HAWK will separate the PSFs forming these clusters into interpolated 'virtual frames', allowing them to be fitted and localised individually [78].

To achieve this, HAWK operates on each pixel separately by treating them as

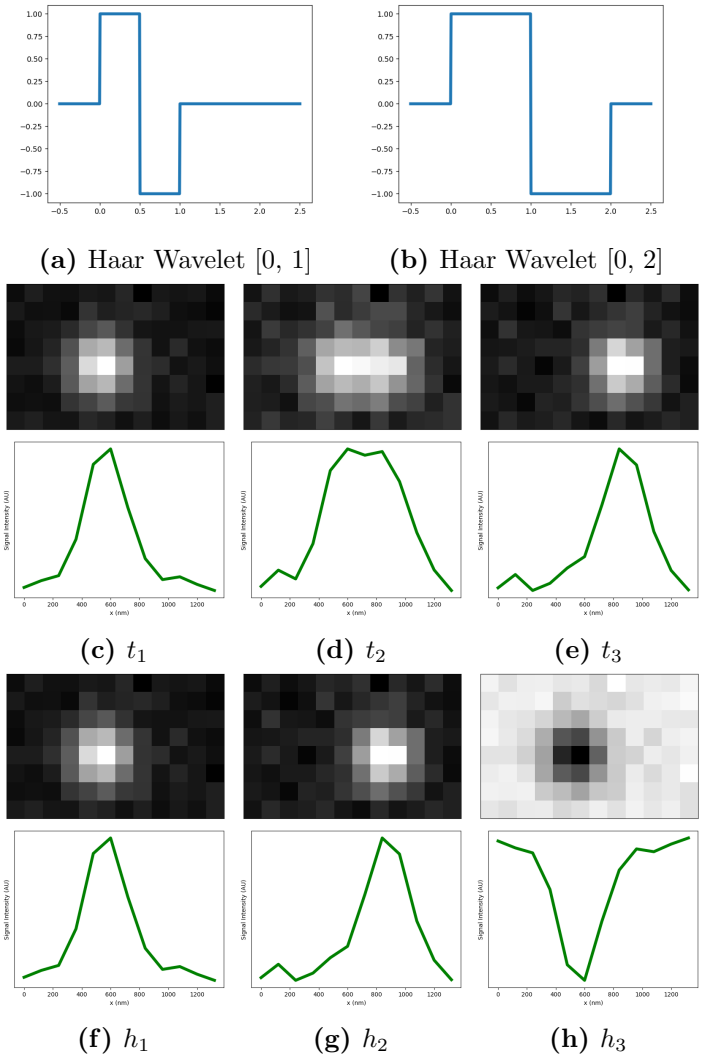


Figure. 1.13. Simplified schema of HAWK's effect on a pair of blinking fluorophores. (a, b) depict two Haar wavelets that might be used in HAWK processing – (a) is a Haar wavelet of length 1 and (b) is of length 2. HAWK processing will utilise a range of lengths in the set [1, 2, 4, 8...] to account for different 'on' and 'off' times for the emitters in the dataset. (c-e) displays three frames from a SMLM time sequence of two adjacent molecules from a time sequence t, and their respective intensity plots when the images have been projected vertically. At t_1 (c), the left fluorophore turns on; at t_2 (d), the right fluorophore turns on; at t_3 (e), the left fluorophore turns off. Looking at the intensity profile at t_2, an attempt to localise the molecules in this frame (d) with gaussian fitting would likely result in an artefactual localisation between the two actual molecule maxima. (f-h) displays (c-e) after HAWK processing. The time sequence t has been replaced with a HAWK'd sequence h which represents the blinking events, rather than the present state of the fluorophores. As a result, at h_2 (g), the molecules from t_2 have been disambiguated. It also results in a negative peak at h_3 (h), representing the turning off of the left molecule. In the final HAWK output, the absolute values of these negative peaks will be obtained through various means (section 3.3.3). The pixel size has been simulated as 120 nm.

time sequences, which must be truncated to an integer power of 2 due to the shape of the Haar wavelet. It applies a Haar wavelet transform to each of these using multiplication with the Haar matrix, which represents a Haar wavelet of a specific length or 'level', m. The elements of the output of the Haar wavelet-transformed intensity sequence that don't correspond to the specific level of the Haar matrix are then set to zero. Lastly, the inverse Haar wavelet transform is applied to this transformed & filtered sequence. The sequence of transforms and filters required to create HAWK'd dataset Z from an N-frame acquisition X, for a single filter level m, is shown in equation 1.7:

$$\begin{aligned} X(t) &= [A_{xy}(n=1), ...A_{xy}(n=N)] \\ Y(t') &= H \cdot I(t) \\ Z^{(m)}(t) &= H^{-1} \cdot Y^{(m)}(t') \end{aligned} \tag{1.7}$$

Here, x and y are the pixel indices, t is the frame number, and A is the signal amplitude. Y is the Haar-transformed input sequence, while $Y^{(m)}$ is Y for which only m-level elements are non-zero. H is the Haar matrix and H^{-1} is the inverse Haar sequence, which is easily found as $H^{-1} = H^T$. Lastly, $Z^{(m)}$ is the m-level filtered intensity trace, whose values will almost only correlate (given an appropriate value for m) with nearby pixels containing the same fluorophore. The final output image sequence is obtained by repeating the above process for each desired filter level m and concatenating each output $Z^{(m)}$. This produces a much longer image sequence than the input acquisition, but when $Z^{(m)}$ has been obtained for all pixels for appropriate m values, the overlapping emitters will mostly have been separated into different frames. The sequence output from the Haar Wavelet Transform can thus be said to depict blinking *events*, rather than signal. That is, a molecule switching on will correspond to a positive peak in the Haar-convolved sequence, and a molecule switching off will correspond to a negative peak. Negative peaks in Z can be handled in a number of ways, each with their own advantages and disadvantages, and that are each best-suited to different localisation approaches (section 3.3.3). This means that HAWK will only fail to separate adjacent, simultaneously emitting fluorophores if their emissions fully correlate – that is, that throughout the acquisition they switch on at the same times, for the same durations, which is highly unlikely. Its only other assumption is that pixels corresponding to individual fluorophores will be correlated in time (i.e. all parts of the PSF will be switched on or off at the same time) [78], which is perfectly reasonable.

Generally, this process needs to be repeated for multiple filter levels. The experimenter can choose fewer filter levels for less memory usage, but more filter levels will

successfully separate overlapping emitters in datasets wherein the fluorophore might stay 'on' for more frames – this may be due to the electrochemistry of the fluorophore being used, or the frame rate of the camera. This is because the Haar wavelet transform for each filter level in essence results in an output of which the signal of each frame is displayed as a comparison with its neighbouring frames. Each filter level essentially corresponds to an increased 'range' of neighbouring frames across which the comparison will be performed. Thus the first filter level will only compare signal in the input time sequence for directly adjacent frames, the second will do so for the surrounding four frames, the third for the surrounding eight, etc. By default, three filter levels are used. This constitutes a sensible starting point for a broad range of structures, labelling densities, and fluorophores, and uses a not-excessive amount of computer resources. However, if an SMLM researcher needs to use fluorophores which are less likely to 'turn off', or if the labelling density is particularly high, they might consider using four or five. If computer resources become a limiting factor and HAWK is unable to complete its transform, the experimenter need simply crop their acquisition stack into sub-stacks and perform HAWK on each one separately before recombining the outputs [78]. This can easily be achieved using the ImageJ macro language [79].

1.5.4 Detecting errors in SMLM

To an experienced SMLM researcher dealing with familiar structures, a severe loss in localisation accuracy leading to heavy artefacts is easy to recognise. In other, more common and experimentally relevant scenarios, however, localisation algorithm performance is tricky to evaluate. There are several methodologies that enable algorithm evaluation in a more abstract sense. One option is to statistically assess an algorithm's ability to create a reconstruction from spatially random data without creating artificial structures [80]. This allows researchers to obtain a quantification of the 'resolution limit' of an algorithm. There is also the 'Localisation Microscopy Challenge', a set of simulated benchmark SMLM datasets (i.e. in which the ground truth (GT) is already known) that developers can test their algorithms on. The creators of The Challenge also developed a set of quality metrics, including localisation accuracy, that can be used to assess the strengths and weaknesses of each localisation methodology.

The above approaches constitute important developments in SMLM research for comparatively evaluating reconstruction algorithms. However, they are conspicuously lacking an important feature: they cannot be directly applied to novel experimental data. The super-resolution quantitative image rating and reporting

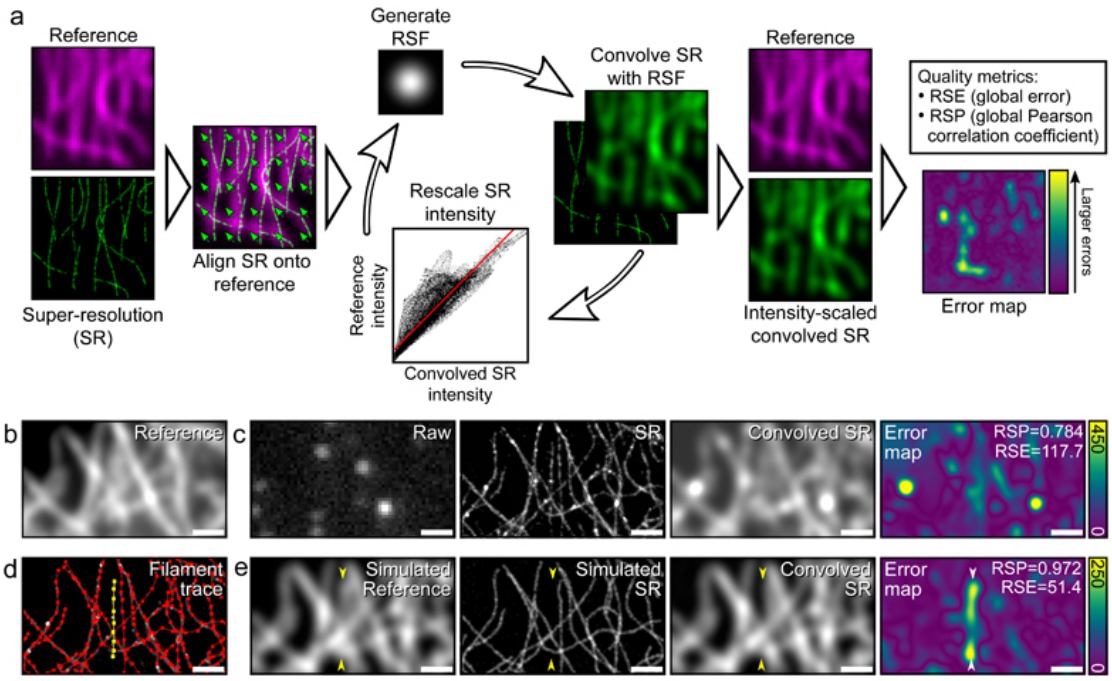


Figure. 1.14. Error mapping with SQUIRREL. (a) Representative workflow. (b) Fixed microtubules imaged in TIRF. (c) Raw – single frame from raw dSTORM acquisition of structure in (b), SR – super-resolution reconstruction of dSTORM data set, Convolved SR – super-resolution image convolved with appropriate RSF, Error map – quantitative map of errors between the reference and convolved SR images. (d) Filament tracing used to generate (e), yellow filament is made to be present in reference image but absent in super-resolution image. (e) Simulated reference image, super-resolution image, and super-resolution image convolved with RSF and error map. Scale bars = 1 µm. (b-d) represents data from 1 of 5 independent experiments showing similar results. Figure from *NanoJ-SQUIRREL: quantitative mapping and minimisation of super-resolution optical imaging artefacts* [81].

of error locations method (SQUIRREL) [81] is a simple and effective method that compares the reconstruction with the most accessible unbiased representation of the GT structure available: the widefield image (taken simply as the sum of all frames in the raw data). SQUIRREL iteratively generates a 'resolution scaling function' (RSF) using intensity information from the widefield. The RSF is convolved with the super-resolved reconstruction in order to bring it to the PSF scale for a suitable comparison. Artefactual structures that occur at the PSF length scale can thus be mapped out using local cross-correlation measures of the RSF-convolved super-resolution image with the widefield (Figure 1.14). These measures can also provide an indication of global bias.

Though immensely useful, because SQUIRREL uses a diffraction-limited image to evaluate the severity of artefacts, it is incapable of reliably quantifying or mapping bias that occurs below the PSF scale. Additionally, because the RSF is generated

using a linear fit of the pixel intensities in the image, SQUIRREL tends to favour algorithms that construct images with a linear intensity distribution over those that don't, even if the former creates images with more spatial artefacts. Indeed, there are no good reasons to think that the shape of a reconstruction's intensity distribution gives an indication of its structural fidelity to the GT. In particular, images containing artificial sharpening are actually often favoured by SQUIRREL analysis for its effects on the intensity distribution. As with FRC analysis, SQUIRREL is frequently beguiled by the apparent high quality of artificially sharpened images [82]. It should be noted that there are cases in which a linear intensity distribution is of experimental value, meaning that using an evaluation algorithm like SQUIRREL makes a lot of sense. In particular, certain types of quantitative imaging – for instance, measuring gene expression – benefit from using images with linear intensity distribution.

It is clear that a more comprehensive evaluation protocol is sorely needed. An ideal candidate would include several features:

1. An approach that enables algorithms to be assessed by application to experimental data, and not only to simulations.

2. The ability to consolidate a quantification of structure sharpness (to measure the maximum attainable resolution) with a quantification of structure fidelity (to measure of the severity of artefacts).

3. A sensitivity to the variation of these measurements across local regions, since artefacts and sharpness vary greatly with the local structure being imaged.

1.6 SMLM in three dimensions

In SMLM, each PSF is separated from the others and can be assessed as a discrete object. This introduces new and innovative possibilities for creating super-resolved 3D images, because molecular separation from the focal plane can be deduced from PSF diameter. However this is a scalar, not a vector-like measurement. That is, PSF diameter can serve as a measure of distance from the focal plane, but by itself cannot measure z-location since it does not tell us whether the molecule is above or below the focal plane. To obtain this information requires a spatial light modulator (SLM) or lens system to introduce an asymmetry in how molecular z-displacement is imaged. Notable examples introduce axial asymmetries into PSF growth [83, 84], image multiple planes at once [85], employ deep learning [86, 87], or use interferometry [88].

These techniques utilise PSF diameter to measure molecule z-locations, which makes emitter overlap – and thus, artefacts – more prevalent. Furthermore, those techniques that modify the PSF to make it grow asymetrically ([83, 84], among others) do so by making the PSF profile *even* larger. As a result, high-density analysis methods become even more important when processing 3D SMLM datasets, especially if attempting to image living cells. However, these approaches require a geometric assessment of individual PSFs to work. This means that algorithms that rely on a temporal analysis of PSF fluctuations to localise molecules in high-density datasets [75–77] cannot be coupled with the above 3D SMLM modalities. In fact, even high-density algorithms that obtain more accurate localisations from the spatial dimension [73, 74] cannot be guaranteed to work with 3D methods, because they don't necessarily 'know how to' fit the abnormal 3D PSF profiles presented. Consequently, only high-density processing methods that preserve PSF geometry can be used in conjunction with 3D SMLM.

Researchers must also contend with novel types of artefact when processing 3D SMLM data. The modulated PSF geometries in these datasets mean that they often present a much greater variety of artefacts than in 2D data. These are often far more challenging to diagnose due to their eclectic nature – they are highly method-sensitive, and synergistically reduce accuracy.

In the most rudimentary case, a cylindrical lens is introduced before the PSF is imaged. This restricts PSF growth to the x-axis when imaged molecules are above the focal plane, and restricts growth to the y-axis when they are below it. This means the PSF ellipticity can serve as a measure for its z-location. The so-called 'astigmatic method' is outstandingly implementable in terms of its simple hardware and computational inexpensiveness, and can attain a super-resolved axial localisation precision of roughly 50 nm. For these reasons, it is still the most popular method for 3D SMLM. However, it creates many new issues for researchers when compared with 2D SMLM.

The astigmatic method, like many others, reduces localisation precision and accuracy along the axis of ellipticity. The elongated PSFs have a lower curvature and are more likely to overlap along this axis. Both of these issues are rapidly worsened with distance from the focal plane, making for a small functional range ($<1\,\mu m$). Artificial sharpening induced by signal overlap will also result in severe mis-localisation in z, as it will dramatically change the ellipticity of the imaged PSF. In 2D SMLM, this kind of result is easily filtered out, as the standard deviation of the fitting function becomes excessively large. However, in astigmatic 3D SMLM, large PSFs with strong ellipticity are no longer necessarily an indication that something

has gone wrong – it could be that they are simply heavily displaced in z. Thus, they cannot be easily filtered out without also removing localisations that are accurate. In addition, two PSFs that cannot be resolved may simply present to a localisation algorithm as a single, highly ovular PSF. This means that situations that would ordinarily only lead to artificial sharpening now also lead to heavily displaced z-locations. These axial displacements can present much more dramatically than artificial sharpening, without being necessarily any easier to diagnose and remove.

Aside from the issues discussed above, the method suffers from relatively poor axial localisation precision, despite its ease of implementation. By contrast the 'double helix' and 'tetrapod' PSFs, which can be obtained by passing the emitted signal through a phase mask, have excellent localisation precision and range. The double helix phase mask splits each PSF into two lobes whose mutual angular rotation serves as a measure for the molecule's location in z. It has roughly double the functional z-range and precision when compared with the astigmatic method. This makes it highly effective at medium-to-low densities; however, the two lobes must always be resolvable from each other, meaning that a single PSF is always about three times as large as the diffraction limit. Moreover, at high densities, the double helix localisation algorithm struggles to determine which lobes belong to which pairs, which can result in quite severe artefacts in all three dimensions. And of course, splitting the photons between two lobes halves the intensity of the signal in each, which is a detriment to localisation precision.

Similarly, the tetrapod PSF [89] is iteratively designed to maximise localisation precision using the Fischer information. These PSFs boast a unique customisability: input a desired functional range up to 20 µm, and a tetrapod will be generated that can determine the axial location of a molecule at the limit of possible precision. In spite of this benefit, tetrapod PSFs are among the largest, spanning over 10 µm at the limits of their functional range, about forty times the diffraction limit in a normal microscope. This makes them so artefact-prone that they are all but wholly impractical for most biological experiments; additionally, distributing the signal across such a large PSF squanders xy-localisation precision. These cases testify to the frustrating fact that in 3D SMLM, there is often a trade-off between a method's ability to localise in x/y and in z.

Biplane microscopy extends its reach in the opposite direction, sacrificing localisation precision while remaining relatively artefact resistant. It obtains axial positioning information by imaging multiple planes at once, using the size ratios of imaged PSFs to determine their axial location with high precision [85]. There are clear strengths to this approach: first, the absence of PSF engineering means that

problems of increased bias arising from enlarged PSFs disappears; second, the axial range can be increased by imaging more planes at once. However, this method also suffers a trade-off – if the planes are too far apart, the z-precision is poor, but if they're brought closer together, z-range is lost instead. Furthermore, by partitioning the signal between the imaged planes, the SNR is degraded, meaning that biplane microscopy squanders localisation precision along all three axes with every plane added. Exceptionally, if performing biplane microscopy on no more than two focal planes using two antisymmetric imaging objectives, the signal does not need to be partitioned prior to imaging. In this case, 3D SMLM by biplane imaging arguably becomes one of the most error-resistant approaches currently available, though it requires a fairly specialised set-up that may not be compatible with certain sample holders and illumination techniques.

1.7 Deep learning & SMLM

1.7.1 Principles of deep learning

'Deep learning' refers to an algorithmic paradigm employing a machine learning architecture that is designed to mimic the complexity of human reasoning. This architecture of 'neural networks' (NNs) aims to emulate our ability to make complex inferences from limited raw information [90, 91]. Broadly, NNs employ a combination of linear transformations and high-dimension calculus to dissect or construct information from a multitude of formats, including (but not limited to) images, sound waves, even human language. In the simplest case of an image categorisation task (arguably the most recognisable exemplar of an NN task), the network will normalise each pixel's intensity value, and pass them through a set of linear transforms. The 'nodes' or 'neurons' conducting these transforms perform fairly simple transforms of the form

$$p_j = w_{ij} \cdot p_i + b_{ij} \tag{1.8}$$

meaning they multiply their input value p_i by a weight w_{ij} and a bias b_{ij} to obtain the output value p_j. These neurons typically number in the thousands (or even millions) which, coupled with extremely dense connectivity over multiple layers, enable highly complex classification tasks. The eventual output of the network depends on the nature of the categorisation task: for instance, a network trained to read handwritten single-digit numbers (e.g. using the now-famous 'MNIST dataset') will output ten numbers between 0 and 1, each representing the network's certainty

that the image is a drawing of each number. Passing data through a trained network is generally a mathematically straightforward and computationally inexpensive process. The ingenuity and developmental challenges of neural nets lie in the 'training loop', the cyclical application of the chain rule of differentiation [92] to calculate how neuron parameters w_{ij} and b_{ij} must change incrementally to optimise the network.

Let's again consider the training of a network designed to recognise the handwritten digits 1-9 using the MNIST dataset. In this instance in the training loop, the untrained network must recognise the handwritten digit 6. A 'perfect score' for this network's output would be represented by the column vector $[0, 0, 0, 0, 0, 0, 1, 0, 0, 0]$, which represents a 100% certainty that the input was a 6, and a 0% certainty that it was any of the other possible numbers. However, prior to training, the network cannot recognise the numbers at all, and when fed a handwritten 6, it outputs a column vector of numbers randomly distributed between 0 and 1. The first step to training a network involves choosing a *cost function*. The cost function is an algorithm that we can use to consistently calculate 'how wrong' the network's output is during the training loop. More practically, we can describe the cost function as measuring *how badly a configuration of weights and biases will perform a classification task for a given network architecture*. Its output is the *loss*, which is a single number quantifying the extent of the network's failure to correctly categorise the input – or alternatively, *how the weights and biases need to change to correctly classify that same input*. In this case, a researcher might use an 'L2' cost function – which would take the sum of squared values between the desired and actual output vectors – to calculate the loss.

The optimisation process of the NN training loop, *backpropagation*, uses the loss to formulate an attempt to improve the network. Here, the loss is utilised to calculate the relative change required in the network weights and biases. We can think of an NN's cost function as a highly complex function with thousands of parameters – the weights and biases – to be optimised. In defining a cost function, we effectively describe a manifold of network parameters plotted against loss, whose minimum represents the best performance an NN can exhibit on a given task. Thus, we can say that optimisation of the network occurs when the loss output by the cost function reaches a *global minimum*. However, global minima are extremely hard to obtain when the number of parameters is so high, so NN developers settle for finding the smallest *local* minimum they can access. The configuration of weights and biases corresponds to a position and a slope on the cost function manifold. This slope is the gradient of the cost function, meaning it provides 'the direction of steepest descent' for the function. It tells us how the network weights and biases

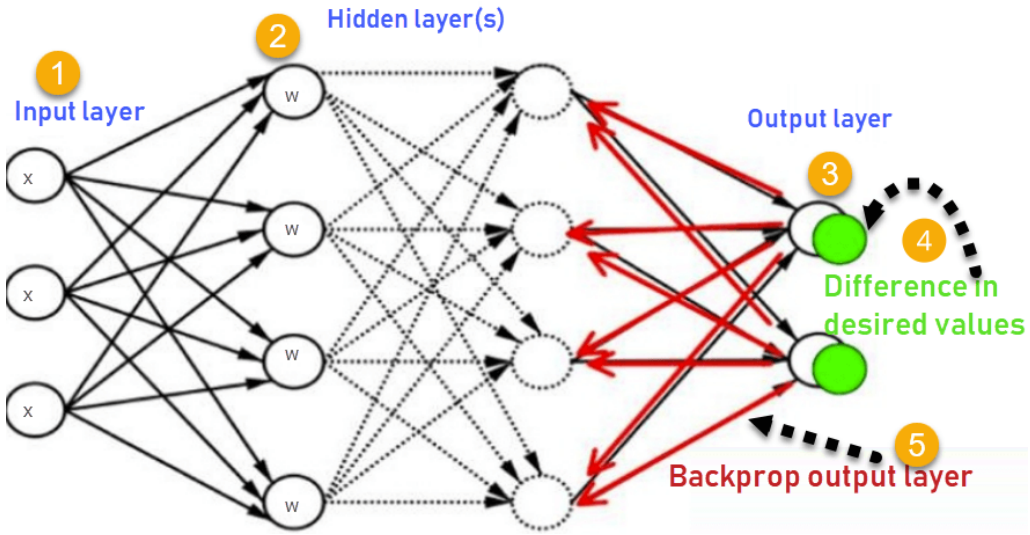


Figure. 1.15. Diagram explaining the principles of a feed-forward (i.e. non-cyclical) ANN. The input layer (1) consists in data that will be transformed by the network – here the nodes can represent, for instance the pixels of an image, or the amplitudes of a 1D signal. The hidden layers (2) transform the data further, and are required for the network to perform more complex tasks – however, they also quickly make training more computationally expensive by increasing the number of parameters that need to be optimised. The output nodes (3) can correspond to classifications in a categorisation task, or e.g. pixels if the network's task is to make synthetic data. To train the network, the output values are compared with desired output values (4, green), the differences are passed through the cost function, generating a 'loss' that indicates how poorly the ANN performed on the task in training. The magnitude of the loss, coupled with differential calculus, can be used to calculate a 'nudge' in the weights and biases of the previous layer that will push the actual output towards the desired values. Backpropagation (5, red) then exploits the chain rule of differentiation to obtain a nudge to the *next* previous layer to achieve the same end. A sufficiently large (but not too large) volume of trianing examples, and an appropriate network architecture, will result in a configuration of weights and biases that can consistently deliver an output that is close to the desired output on any novel relevant input. Image adapted from *Back Propagation in Neural Network: Machine Learning Algorithm* [93]

should change incrementally (in which 'direction' they should 'step') to reach a local minimum of the loss most quickly. From the loss, we obtain the gradient of the final layer L from the loss, then from here we obtain the gradient of the penultimate layer $L - 1$, using the chain rule to work backwards [94] through the network and obtain the required step for every weight and bias – hence the term 'backpropagation'. The 'size' of the step taken is scaled to be proportional to the gradient of the cost function at that point, ensuring that minima are not overshot. By repeating this process thousands of times in a process called 'gradient descent', the network parameters are changed to a point where it can theoretically perform its classification task nearly infallibly. It should be noted that training is performed in batches for the sake of training speed and computational efficiency, so only a single backpropagation step will be performed from the losses from several images. In this case researchers speak of 'stochastic gradient descent' [95–98] because processing the images in batches means that each step is not guaranteed to 'walk' down the angle of steepest descent.

1.7.2 The convolutional NN

Neurons in NNs commonly perform some iteration of the linear transformation in equation 1.8 because it does not contain too many parameters to optimise. Among such iterations, the most notable in image processing and computer vision is the convolutional layer [99]. Rather than fully connecting each neuron to every other in its surrounding two layers, convolutional layers scan a 'kernel' across the image. The kernel is a (usually) small n-dimensional array of numbers that represent the parameters of a simple linear transform, not unlike equation 1.8. This transforms the image according to the kernel parameters, allowing certain structural features to be emphasised, such as edges or corners, which is particularly useful in the context of NNs. Convolutional kernels are used extensively in image processing [100], and particular distributions of kernel parameters are used to perform common transforms, such as gaussian blurring. However, manually specifying the hundreds of convolutional kernels that would be best used in tandem to obtain a particular result would be tedious and exceptionally difficult. Fortunately, the parameters of these kernels can be optimised to perform a given task using the logic of NNs and backpropagation (section 1.7.1).

Including convolutional layers into a NN introduces advantages in terms of image feature-sensitivity and computational effectiveness. In the standard (more antiquated) feed-forward 'perceptron' NNs described in section 1.7.1, each layer of neurons is 'fully connected' to the neurons in the surrounding layers. This creates

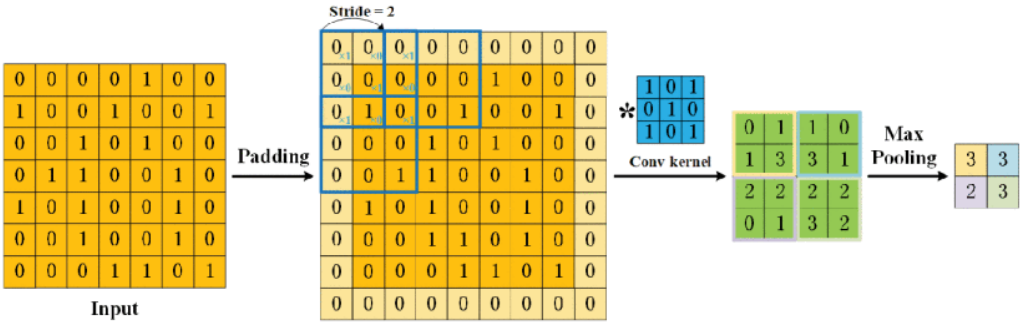


Figure. 1.16. Basic 2D convolution such as might be found in a CNN. A convolutional kernel (blue) is scanned through the array (orange) which can be the input (e.g. image) or the previous layer of the network. This means that each layer does not need to be fully connected. To control the dimensions of the output (green) and reduce edge effects, a 'padding' of ones or zeros is frequently added to the input. This process will usually be repeated with multiple kernels whose values are being learned by the network, which creates several outputs (alternatively 'output channels', not depicted). Each of these can be thought of as a kind of feature abstraction, with each output channel corresponding to an 'attempt' by the kernel to detect the prevalence of a particular structural feature of the input. However, the density of features that will be present in each output channel can lead to overfitting, where multiple features are over-correlated with each other. To combat this, pooling layers (kernels that take the maximum or average value of a region, rather than a convolution) are frequently employed to downsample the outputs from each layer. Additionally, rectified linear unit (ReLU) and normalisation transformations (not depicted), are often used to reduce or remove negative and extreme values from the output to simplify the output and speed up training. Figure images from *A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects* [101]

computational overhead during training by maximising the number of learnable parameters. In convolutional NNs (CNNs), however, the convolutional layers are optimised to extract local features and, in essence, convert the image into one of lower dimensions, which boosts computational efficiency without compromising training effectiveness. They are not fully-connected, rather, they re-sample local regions of their input layer into a local region of their output layer using convolutional kernels. Additionally, convolutional layers retain information about the relative positioning of image pixels when performing transforms. This 'local feature extraction' makes CNNs highly effective for processing spatial data, especially when we consider that each convolutional layer consists of multiple kernels that splits the image into 'channels' each representing one type of feature. A CNN comprising many convolutional layers can thus ultimately perform a kind of perfect 'feature abstraction' in which an image is condensed into a list of 0-dimensional numbers, each representing the prevalence of a particular feature. This can make it much easier to perform complex analytical tasks in image processing.

1.7.3 Principles of generative adversarial networks

While *discriminative* networks enable efficient categorisation, *generative* networks are used to transform data, or create new data entirely, even from 0-dimensional noise. Generative networks have seen a great breadth of applications across multiple bio-imaging disciplines. These include particle tracking [102], image augmentation including denoising and deconvolution [103, 104] and super-resolution [105, 106], image segmentation [107] including for medical diagnosis, for instance, by image-based assessment of retinal blood vessels [108].

The generative adversarial network (GAN) [110] is a particular configuration of generative network that learns to create synthetic data by being pitted against a discriminative network, the 'adversary'. To train a GAN to make images of a cat, the adversary is presented with a real picture of a cat, or a 'fake' one synthesised by the generator. The discriminator outputs a number between 0 and 1, representing its best guess at the probability that the image is not one of the generator's counterfeits. The loss for the generator is calculated as the extent to which its output was recognised as fake. Conversely, the adversary's loss is calculated as the extent to which it classifies the generator's output as real[1]. The loss for each network is thus calculated as the extent to which it beats the rival network in this contest. It is a zero sum game that forces both networks to improve until equilibrium is reached,

[1]The other component of the adversary's loss is obtained similarly from its failure to recognise real inputs as such.

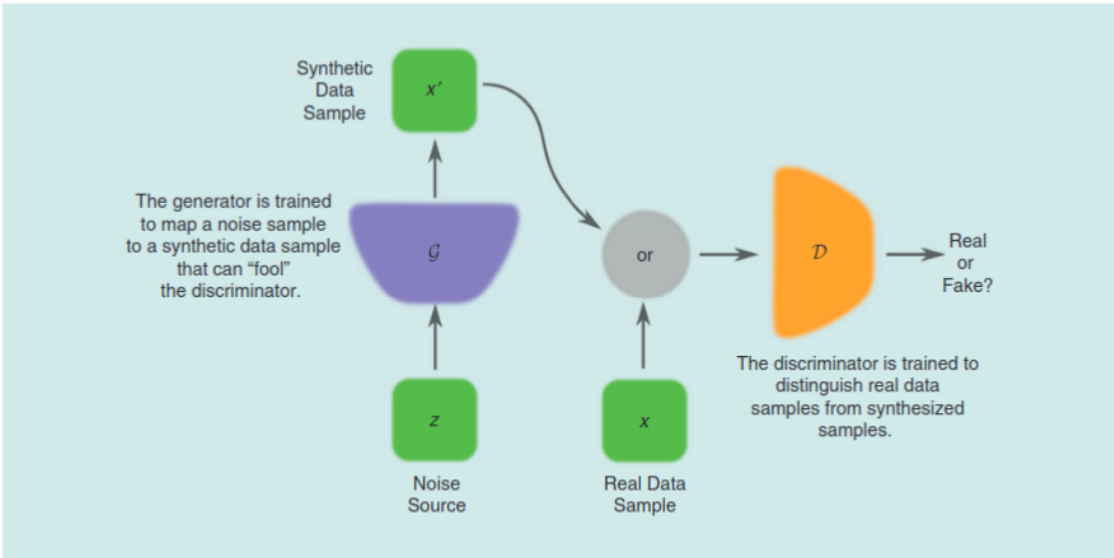


Figure. 1.17. Architecture of a basic GAN. A neural network G converts a synthetic image x' from noise z. Another neural network D is presented with the synthetic output x', or a non-synthetic image x. The cost function backpropagates a high loss through G and a low loss through D when D identifies x' as synthetic & vice versa. Additionally, D's loss is low when it indentifies x as non-synthetic and vice versa. This structure results in a completely unsupervised system where G learns to replicate the features present in x. Figure image from *Generative Adversarial Networks: An Overview* [109].

at which point the adversary will be expected to correctly classify fifty percent of the generator's fakes.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1 - D(G(z)))] \tag{1.9}$$

Equation 1.9 explains the parameters of the game being played by the two networks. The first term denotes the success with which the adversary D (the 'discriminator') is able to correctly label real data x, and the second term its success rate in identifying the generator fakes $G(z)$ as counterfeits. The $\min_G$ and $\max_D$ describe a minimax game, meaning G wins when V is minimised, and D wins when it is maximised.

1.7.4 Deep learning in SMLM

Using deep learning to perform difficult localisation tasks is an intriguing and popular avenue of research in SMLM. These methods promise two main advantages over more traditional algorithms:

1. They deliver a high performance in terms of image accuracy while promising significantly faster processing times, because a fully-trained NN performs nothing more than a large volume of linear transforms on image pixels. This

operation is far less computationally expensive than, for instance, calculating the spatial differential of signal divergence for image sub-pixels in every frame (as in SRRF [77]). This is because the most demanding and time-consuming part of NNs occurs at the developmental, not operational stage. These must be trained on large volumes of data (which can be hard to obtain), while regularly performing high-dimension calculus (section 1.7.1). Nevertheless, once trained, the savings in operational costs of using Deep-STORM compared to, say, 3B, are monumental, taking the order of time taken to complete a full reconstruction task from hours (or even days) to minutes (or seconds, if done on a GPU) [76, 87, 111]. This increase can make NNs more viable for quantitative research, which often demands high sample sizes.

2. Data-driven approaches to analysis like deep learning offer the advantage that they do not rely on a successful modelling of the PSF or blinking dynamics that has to account for multiple sources of stochastic noise and other small variances in the ways that PSFs present when imaged. They can learn complex image transformations, including localisation, without explicit analytical modelling. Rather, they simply rely on a large and broad set of exemplary training data. As a result, they often appear to out-perform alternative high-density localisation algorithms, both in terms of localisation accuracy and precision [71, 86, 87, 111]. A common frustration for researchers and students upon learning about SMLM is that overlapping emitters that lead to artefacts are often quite obvious to the human eye. This kind of complex visual processing that comes naturally to the human brain is exactly the kind of task that NNs have historically found success in. Molecule localisation is thus archetypical of the kinds of task that NNs are typically meant to be highly performant at.

There are several approaches to implementing a deep learning localisation algorithm for SMLM. Instead of approximating the PSF with a function, Deep-STORM makes use of a convolutional NN to recognise PSFs across a range of densities to localise molecules [111]. It can instantly recognise a great number of configurations of overlapping PSFs to tease them out and produce more accurate reconstructions. Two years later, DeepSTORM3D was developed from this model to localise molecules imaged with the tetrapod 3D PSFs [89] (section 1.6), making them usable despite their size and tendency to overlap heavily, while retaining their maximised axial localisation precision [86]. The DeepSTORM3D algorithm was even used to develop a new 3D PSF that is optimised for its own localisation algorithm that is both smaller and better-performing than the tetrapod PSF [86]. More recently,

DECODE (deep context dependent) [87], another high-density deep learning-based localisation algorithm, out-performed all other tested algorithms on benchmarked data [112] by incorporating uncertainty measures for its predictions into its training loop. ANNA-PALM, instead of using high-density data, utilises a sparse 'under-sampled' SMLM reconstruction obtained from a short acquisition of low-density data, and employs deep learning to interpolate the rest of the reconstruction using information from the widefield image [71].

However, these apparent advantages belie several issues with deep learning approaches, particularly in the hard sciences and especially in microscopy. As in all applications of deep learning, performance depends heavily on the quality and quantity of the training data, and the similarity in properties between the training and final experimental data. And as with all approaches to high-density fitting, a fundamental challenge with deep learning methodologies is that it is very hard to tell when the algorithm is operating outside the bounds in which it is accurate. Certainly, the failure modes of deep learning methods is somewhat different from the approaches discussed above – they will not suffer as easily from artificial sharpening. However, when fitting high-density data, they will regularly make the mistake of hypothesising structure that isn't real. Thus, while recent work has demonstrated the enormous power of deep learning approaches, deep learning-based localisation software still suffer from what we might call artefacts, with evaluation and reliability remaining major challenges.

The potential of NNs can be dangerously overestimated, leading to misguided training protocols. Once trained, NNs are 'brittle', that is, the data-driven approach at the core of their development can give them trouble with novel data. These issues have been exhibited in discriminative NNs that are designed to classify images. For instance, trained CNNs have been shown to struggle greatly to classify rotated versions of their training data. This happens because convolutional layers perform feature extraction based on a learned relative positioning of signal in the image, even though this relative positioning is not rotation-equivariant. Additionally, because of their high sensitivity to individual pixel values, some CNNs have also mis-classified images when carefully calibrated noise is added to them – while to a human eye the two images look near-identical [113].

Aside from being brittle – many highly performant algorithms can be classified as such – the NN's inner workings are often described as a 'black box' due to their enormous complexity. This means the above issues cannot be resolved by simply inspecting the individual weights and biases; rather, the network must simply be trained again with new data or a new architecture in the hope that the difficulties

evaporate. Because of their novelty and algorithmic inscrutability, it can also be difficult to predict the ways in which NNs will fail. This is a particularly significant issue given the increasingly widespread development and use of NNs for critical applications, as people rightly feel wary of entrusting important digital infrastructure to inscrutable algorithms. Recent developments in explainable AI (XAI) attempt to overcome this problem by introducing transparency and interpretability to the operations that AIs perform, thus building trust in the results produced by NNs. XAI is a field of study in its own right and increased interpretability is a achieved in a great number of ways; for instance, the impact of certain features on the output in a classification task may be quantified [114]. Thus, while low interpretability remains a significant problem in AI development, it is not intractable and much work is being done to introduce new metrics to assess the functionality of NNs, in part due to the enormous profit potential in increased trust in AIs. Nevertheless, the *extent* of interpretability of NNs is reduced when compared with algorithmic approaches due to their data-centric approach.

Chapter 2

HAWKMAN

2.1 Introduction

Broadly speaking, 'image quality' in SMLM can be broken down into two primary constituents: localisation precision, and localisation accuracy. Localisation precision quantifies the image sharpness in that it determines the width of the LEFs (section 1.4.1) that make up a SMLM image. Localisation accuracy provides the image fidelity, meaning it tells us whether the locations of the LEFs are biased. SMLM is unique among fluorescence microscopy modalities in that *both* measures are required to give a meaningful measurement of an image's resolution. That is, an SMLM image can be full of structures that are mislocalised due to artefacts (section 1.4.3), retroactively rendering other quality metrics like high SNR and image sharpness meaningless. In spite of this, there are presently no quality metrics for SMLM images that can be applied to non-simulated data that incorporate both a) the precision with which we can take measurements and b) the fidelity to the GT.

A clear common thread in currently-existing approaches for evaluating artefact severity is the reliance on a reference image, that does *not* contain artefacts, to evaluate the fidelity of an SMLM reconstruction. For simulated data, as in the Localisation Microscopy Challenge, this reference is the GT of the simulated structure [115]. For SQUIRREL, it is the widefield image, often taken as the sum of the acquisition frames [81]. As we saw in section 1.5.4, these reference images are problematic. In biological imaging experiments, of course, the GT is unavailable. And though SQUIRREL can be used on experimental data, its use of the widefield image as a reference image also engenders several problems – even aside from the fact that there is no super-resolved information present in the widefield to be used for assessment. SQUIRREL's measure of image fidelity employs an optimisation process using intensity information from the widefield image. Its outputs on some

images can be suspect, because they will favour reconstruction algorithms that render images with a more linear intensity distribution (section 1.5.4). This has little to do with whether structures' locations have been biased, and as a result, SQUIRREL sometimes favours algorithms that generate images with more artificial sharpening (section 2.3.4). The severity of background signal sampling, and variances in blinking rate and fluorescence intensity, all contribute to signal non-linearity without necessarily being indicative of bias. For example, the gaussian profile of the illuminating laser, or the decay in signal intensity with distance during TIRF imaging can both cause a non-linear variance in the image intensity distribution (section 1.4.1). These issues in SQUIRREL stem largely from its reliance on the widefield image as a reference image, as the latter presents very differently in terms of structure fineness and signal distribution [82]. SQUIRREL also does not include an integrated methodology for measuring the resolution of the image: though it can be coupled with FRC to achieve a more comprehensive evaluation of image quality, they cannot be used in tandem to measure the 'effective resolution'. Furthermore, FRC uses reduced LEF scatter as a proxy measure for resolution – however, reduced LEF scatter, like linear intensity distributions, is also often indicative of artificial sharpening (section 1.4.2). Thus SQUIRREL and FRC, arguably the leading image assessment technologies for SMLM, have been shown to report many symptoms of bias leading to structural inaccuracies as increased image quality.

Another important factor in measuring bias and resolution in SMLM images is a recognition of the technique's sensitivity to the local structure [67] and the discrete attributes of each LEF in the acquisition. In non-super-resolution microscopy, the diffraction limit for the imaging set-up is also a good approximate measure of the global resolution of the image – the size of imaged PSFs that are equally distanced from the focal plane will be (roughly) the same. Conversely, each LEF in an SMLM reconstruction will be a different size, depending on a range of other factors including the severity of background, the geometry of the local structure being imaged, and the local labelling density. This means that discretely assessing local regions of interest (as opposed to obtaining a global 'score' for the 'image resolution') is essential, because the severity and frequency of artefacts and the sharpness of imaged features vary enormously with local parameters.

HAWK presents a solution to these issues. This is because HAWK'd reconstructions generally do not contain mis-localisation artefacts thanks to the removal of overlapping emitters from the dataset; they can thus be used as the ideal reference image to evaluate localisation algorithms. This is not to say that HAWK'd reconstructions cannot 'lose' truthful structures; on the contrary, a very high emit-

ter density in the input dataset leads to a HAWK'd reconstruction with a lower resolution [78]. This means that even though false structures are removed from the dataset, truthful structures can be obscured by the increased spread of the localisations resulting from HAWK processing. However, the HAWK pre-processing changes the failure mode of the localisation process, meaning disagreement between the HAWK'd and non-HAWK'd images will still successfully indicate local areas where sharpening has occurred.

The HAWK Method for the Assessment of Nanoscopy (HAWKMAN) represents the first solution to the problem of localisation algorithm assessment that is sensitive to the presence of nanoscopic artefacts in novel experimental data. Additionally, it presents a new way of conceptualising SMLM image quality that reconciles structure sharpness with structure fidelity, allowing for a more comprehensive definition of resolution and a more effective method of localisation algorithm evaluation. It is also able to map the scale of artefacts in different regions of the reconstruction, enabling their severity to be reconciled with the resolution of local structures – instead of attempting to quantify the global resolution of the image. To achieve this, HAWKMAN creates error maps in which the structure obtained from a HAWK'd reference image is compared with its HAWK-free counterpart, the 'test' image. These maps have been convolved with multiple PSFs of varying diameter to achieve a progressive 'blurring', which can be used to determine the length scale at which agreement between the test image and its artefact-free (HAWK) counterpart is broken. The smallest 'length scale' at which the scale of local artefacts does not outstrip that of the imaged structure itself is the local resolution. This definition of resolution is uniquely powerful in that it is responsive to both the sharpness of depicted structures *and* the incidence of artefacts: the first length scale at which local disagreement between the HAWK and test image occurs represents the *greatest precision* with which we can *safely* make measurements.

Thus, the goals of HAWKMAN are to:

- locally assess the severity of artefacts in SMLM reconstructions without access to the GT
- locally assess the resolution of SMLM reconstructions – *without* misconstruing structure fineness alone as high resolution, since this is frequently indicative of artefacts
- allow SMLM researchers to easily determine when an SMLM acquisition should be performed again because the severity of artefacts outweigh the capabilities of the localisation algorithm

2.2 Materials & Methods

2.2.1 HAWKMAN sharpening detection

Statement: Dr Richard Marsh at *The Randall Centre for Cell & Molecular Biophysics* developed the HAWKMAN algorithm [82] – many thanks to his efforts and ingenuity. Testing of HAWKMAN was a collaborative effort, and experimental data was acquired, processed and analysed by myself.

HAWKMAN analysis enables the similarity of two SMLM images to be assessed in a way which does not rely on the intensity values in the images being linearly related to the labelling density (or even each other). Instead, it concentrates only on structural differences, enabling us to reliably detect bias in the positions of structures. It should be noted that while this method will detect artefacts caused by errors in the SMLM data analysis, it will not detect artefacts which cause fluorophores to be imaged in a different location to the protein of interest (e.g. labelling artefacts). Therefore, to fully assess the authenticity of a super-resolution image, multiple methods of assessment may be necessary.

Two images are compared: the test image (which is undergoing quality assessment) and the reference image (which is produced using HAWK pre-processing and a fitting algorithm such as ThunderSTORM [40]). We also sometimes used fits of LD data for the reference image as a proxy for the GT when validating HAWKMAN, because the lack of substantial emitter overlap means the LD reconstructions are approximately artefact-free. Very bright points in SMLM images frequently occur as a result of repeated localisations, which can lead to variable thresholding performance. Because of this, we cap the maximum pixel intensity to the 98th percentile of the intensity histogram (not counting any zero-valued pixels in the reconstruction).

A Gaussian blur is then applied to the image. The size of this blur (given by the full width at half-maximum, FWHM) determines the length scale at which the quality of the data is being assessed. By using multiple sizes of blur, the local resolution of any part of the reconstruction can be determined. This size corresponds to the length scale at which information starts to become locally unreliable. The algorithm will assess successive length scales (blur levels) up to a user-specified maximum, with each scale a progressively increasing integer (the scale number, n_{scale}) multiple of the reconstruction pixel size. The images are then normalised to a maximum intensity of one. From this starting point, three different mappings of errors are produced, each likely to highlight errors under different conditions.

The first assessment which we calculate is the sharpening map. An adaptive threshold for each image is calculated using Wellner's method [116], which classifies

pixels based on their value relative to the surrounding pixels within a defined neighbourhood. Here we use a neighbourhood size $(2r+1)$ of the largest odd number of pixels not greater than the current length scale. This produces a map of the mean intensity around each pixel. The images are then binarised based on whether the pixel intensity is above a set proportion (C_b) of the local threshold ($P^{thr}_{x,y}$):

$$P^{thr}_{x,y} = C_b \frac{1}{(2r+1)^2} \sum_{x-r}^{x+r} \sum_{x-r}^{x+r} P_{(x,y)} + \text{ where } r = \text{ceil}(\frac{n_{\text{scale}}}{2}) - 1 \tag{2.1}$$

where ceil(x) raises x to the first integer greater than x. The value of C_b determines what pixels in the image are classed as structure and which are classed as background. A value of $C_b = 0.7$ leads to a thresholded image where the width of linear structures is roughly equal to the FWHM of the intensity profile of a line profile through the structure in the original image. This value is designed to pick up features which have a peak-to-trough intensity ratio of about two or more. Where the image values between parts of a structure does not fall to this level, precise identification of the type of artefacts may require a higher threshold value, however, the default value will still detect that artefacts are present. This can be seen in the clathrin-coated pits image (Figure 2.13e): here, the depth of the central hole is sometimes shallower than a factor of two, so a higher coefficient of 0.9 was used to also detect sharpening in these shallower pits.

At small blur scales (up to $\sim 2\times$ the localisation precision), local variations in labelling density and multiple reappearances of individual fluorophores can be falsely identified as very fine structures. This is particularly common when imaging finer structures, such as microtubules. This can result from rapid increases in the adaptive threshold around these features, breaking the connectivity of the binarised structures. To counter this, we determined experimentally that an additional threshold of $0.1\times$ the calculated threshold at a neighbourhood size of half the PSF helped smooth out this false fine structure, by reducing the rate of spatial variation of the threshold. This stage is not strictly necessary to identify artefacts but gives a smoother, better connected binarised image if this false fine structure is present. Additionally, a small baseline contribution to the threshold ($C_a = 0.4$) eliminated many background localisations and reduced the fixed-pattern noise of some algorithms. With these additions, Equation 2.1 becomes:

$$P^{thr}_{x,y} = C_b \frac{1}{(2r+1)^2} \sum_{x-r}^{x+r} \sum_{x-r}^{x+r} P_{(x,y)} + C_b \frac{0.1}{(2r'+1)^2} \sum_{x-r'}^{x+r'} \sum_{x-r'}^{x+r'} P_{(x,y)} + C_a \tag{2.2}$$

This noise threshold is designed to eliminate the background localisations that are not part of the labelled structure, with the magnitude of these in the reconstruction being assumed to be constant across the image. If it is not, regions of interest in each image can be taken and analysed separately with particularly noisy areas using a higher value for this coefficient. This was not necessary for any of the data analysed here. A colour overlay of the binarised images (the sharpening map) is then produced to reveal sharpening/ bias artefacts. Areas where there are substantial differences between the images are indicative of sharpening in the test image, or loss of precision in the reference; areas where the two images agree can be considered reliable.

This 'sharpening map' is appropriate for detecting structure thinning and biases, but less suited to visualising the collapse of adjacent structures to one. To obtain such a visualisation, the binarisation of the blurred and flattened images is repeated but with a higher threshold coefficient such that $C_b^{\text{skel}} > C_b^{\text{sharp}}$, with default value $C_b^{\text{skel}} = 0.85$ and a baseline threshold of $C_a = 0.02$ being used to detect local maxima. These binarised images are then skeletonised using the skeletonisation function in ImageJ giving a skeletonised interpretation of the structure. Similarly to the sharpening map, a composite image of where the two structures agree and disagree reveals differences in structure collapse.

It should be noted that skeletonisation can lead to very branched structures that are highly dependent on noise in the input images. This problem is greatly reduced by the initial Gaussian blur of the input images, particularly as the length scale increases. The skeletons produced obviously result in single pixel-width structures that may only differ by single pixel displacements. This could possibly lead to a very high level of measured disagreement, even at large length scales – even though there are only differences of much smaller scale present in the images. We actually only want to consider differences that are on the scale of the current length scale. To achieve this, the skeletons are first re-blurred with a Gaussian blur kernel of FWHM equal to the current length scale. This ensures the overlap of the skeletons is substantial *unless* the differences in structure are comparable to, or larger than, the current length scale. The resulting comparison, the *structure map*, gives an indication of where in the sample and structures are biased to a greater degree than the length scale of interest.

The similarity between the test and reference output images is measured using the two-dimensional cross-correlation for both the sharpening and structure images. The structure images are first Gaussian-blurred to the current length scale to ensure that they become progressively tolerant of differences below the length scale under

investigation. A 'confidence map' is produced by summing the normalised, blurred test and reference images. This is then colourised according to the calculated local cross-correlation value, which indicates how likely the reconstruction is to be unbiased. Here, any local correlation above 0.85 in both the sharpening and skeletonised images is deemed as indicating *valid structure*. Areas below this level are colourised according to the level of agreement as measured by this correlation. The confidence level S_{conf} is given by:

$$S_{\text{conf}} = \frac{1}{2}\min(1, \frac{\text{PCC}^{\text{sharp}}}{0.85}) + \frac{1}{2}\min(1, \frac{\text{PCC}^{\text{str}}}{0.85}) \tag{2.3}$$

Where $\text{PCC}^{\text{sharp}}$ and PCC^{str} are the Pearson correlation coefficients for the sharpening and structure comparisons respectively, calculated over a range equal to the current length scale, and $\min(x, y)$ indicates the smaller of x or y. Due to the slightly reduced precision typical of the HAWK reference image and the effect of applying the Gaussian blur to the skeletons, a perfect correlation between test and reference images is unlikely even when the structures are very similar. Thus an 85% correlation threshold for full confidence is incorporated into the score in equation 2.3. This was determined as an appropriate value from all the simulated structures evaluated.

Similar to the Gaussian scale space framework, we consider this comparison at different length scales by convolving the input reconstructions with Gaussian kernels of different sizes. In this case, each Gaussian has an FWHM that is each integer multiple of the reconstruction pixel size, ranging from a single pixel up to the user-specified maximum, which is typically the instrument FWHM. This is suggested as the upper limit because typically this is the smallest separation at which most algorithms can regularly produce unbiased localisations – artefacts generally occur as a result of emitter overlap, which is much less likely above a certain separation. If the acquisitions are of a very high activation density, however, even this scale may contain artefacts and if so, the calculation can be re-performed with a larger upper length scale limit. Gaussian scale space can be used to quantify the amount of information of a specific scale that is contained in an image. Here we are quantifying the difference in the structural information only between the test and reference reconstructions as a function of scale. The above process is repeated for each length scale, producing an output image sequence wherein the degree of local sharpening is mapped as a function of scale. The actual *reliable* resolution obtained locally in the reconstruction can be assessed by observing the length scale at which disagreement between the test and the reference begins to occur.

The three maps described above provide a quantification of the degree of artefacts present at each length scale. If one wished simply to know which parts of an image

are severely affected by artefacts, a summary of the results may be sufficient. Thus, as an optional further step, a map of the smallest length scale at which each part of the image has high confidence can be produced, called the *artefact scale map*. For each pixel in the confidence map, the smallest length scale at which the confidence score is above 90% is recorded. Each pixel of the input image is then colourised according to this length scale with its intensity preserved. This provides a single, two-dimensional map of the length scales of local disagreement. The image can then quickly be assessed to ascertain whether the features of interest are of sufficient authenticity.

As a summarised guide for the non-expert user: the HAWKMAN GUI will suggest default values for $C_b = 0.7$, $C_b^{\text{skel}} = 0.85$, and the microscope PSF. These can all be easily altered to account for different imaging systems and imaged structures. For instance, very small structures such as clathrin-coated pits, whose diameters are well below the diffraction limit, may require higher values for the threshold coefficients C_b and C_b^{skel}. However, it is important that $C_b < C_b^{\text{skel}}$.

2.2.2 Cell preparation for experiments: tubulin & clathrin

Human cervical cancer cells (HeLa) were thawed and cultured in Dulbecco's Modified Eagle Media (DMEM) 1% Penicillin and Streptomycin (PS, Sigma-P0781), 10% foetal bovine serum (FBS, HyClone-SV30160.03) and 1% L-Glutamine (L-Glu, Sigma-59202C). Cells were transferred to a T25 flask (Cellstar-690175) and incubated at 37°C and 5% Carbon dioxide. Passaging cells were trypsinized with 10% trypsin (Sigma-T4174) diluted in phosphate-buffered saline (PBS, Sigma-D8537), for 5 min once cells reached 80% confluence. Complete DMEM was added to neutralise trypsin and centrifuged at 960xg for 3 min. The supernatant was aspirated, and cells resuspended in 5 mL of complete DMEM. Cells were plated on fibronectin (Sigma-FC010) coated 35 mm high glass bottom dishes (ibidi-81158) for imaging. cells were fixed for 15 min in 3.6% formaldehyde (PeqLab-30201) at room temperature (RT) and washed three times with 1×PBS.

Permeabilisation and blocking were undertaken by incubating cells in 'blocking buffer' consisting of 3% BSA (Sigma-10735108001) and 0.5% Triton X-100 (Sigma-X100) in 1×PBS for 10 min. Cells were then incubated with primary antibodies (tubulin from Sigma-T8328 at 1:200, clathrin from Abcam-ab21679 at 1:500) diluted in blocking buffer for 30 min while gently rocked at RT. After three 5 min washes in washing buffer consisting in 0.2% BSA and 0.1% Triton X-100 in 1×PBS, cells were incubated with the Alexa Fluor® 647 secondary antibodies (Invitrogen-A21235 for tubulin, Invitrogen-A21244 for clathrin, both at 1:500 dilution) in blocking buffer

for 30 min, gently rocked at RT. Cells underwent three 5 min washes in 1×PBS before being stored at 4°C in 1×PBS for up to 2 days before imaging.

Cells were imaged in imaging buffer, where 1220 µL buffer was made using 800 µL distilled water, 200 µL 'MEA', 20 µL 'GLOX' and 200 µL dilution buffer. Here, MEA is 1 M cysteamine (Sigma-30070) and 0.25 mM HCl (Sigma-H9892); GLOX is 35 mM glucose oxidase (Sigma-G6125), 13.6μM Catalase (Sigma-C40), 8 mM Tris (Amresco-E199) and 40 mM NaCl; lastly, 'dilution buffer' is 50 mM Tris, 10 mM NaCl (Alfa Aesar-A12313) and 10% w/v glucose (Thermo Fisher-G/0450). Dilution buffer was added just before imaging to minimise cell damage due to pH change.

Objective-based total internal reflection fluorescence (TIRF) was employed to minimise background. Fluorophore bleaching was undertaken with widefield illumination prior to image acquisition to minimise background signal from above the structures of interest. Approximately 5000 of a total 20,000 10 ms frames were acquired with supplementary 405 nm activation to obtain high-density data.

Experimental data of clathrin and tubulin was gathered from fixed HeLa cells on a Nikon motorised inverted microscope ECLIPSE Ti2-E with Perfect Focus System in the King's College Nikon Imaging Centre. It is equipped with a laser bank with 405, 488, 561, and 640nm lasers (LU-NV series), a 160-nm-pixel ORCA-Flash 4. sCMOS (scientific Complementary metal–oxide–semiconductor, Hamamatsu Photonics K.K.) and a CFI SR HP Apochromat TIRF 100XAC oil objective (NA 1.49) with an automatic correction collar. The microscope was controlled using Nikon Elements v5.2. software.

2.2.3 Cell preparation for experiments: mitochondria

Mitochondria were prepared and stained by Fang Huang at the Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA.

COS-7 cells (CRL-1651, ATCC) cultured on 25 mm diameter coverslips (CSHP-No1.5-25, Bioscience Tools) were first fixed with 37 °C prewarmed 3% PFA (15710, Electron Microscopy Sciences) and 0.5% GA (16019, Electron Microscopy Sciences) in 1×PBS (10010023, Gibco) at RT for 15 min. Subsequently, cells were washed twice with 1×PBS and then quenched with freshly prepared 0.1% NaBH4 (452882, Sigma-Aldrich) in 1×PBS for 7 min. After washing three times with 1×PBS, cells were treated with 3% BSA (001-000-162, Jackson ImmunoResearch) and 0.2% Triton X-100 (X100, Sigma-Aldrich) in 1×PBS for 1h, gently rocked at RT. Then cells were incubated with primary antibodies (sc-11415, Santa Cruz Biotechnology) at 4 °C overnight. After washing three times for 5 min each time with wash buffer (0.05% Triton X-100 in 1×PBS), cells were incubated with secondary antibodies

(A21245, Molecular Probes) at RT for 5h. Both primary and secondary antibodies were diluted to 1:500 in 1% BSA and 0.2% Triton X-100 in 1×PBS. After washing three times (5 min each time with wash buffer), cells were postfixed with 4% PFA in 1×PBS for 10 min. Then cells were washed three times with 1×PBS and stored in 1×PBS at 4 °C until imaging.

Immediately before imaging, the coverslip with cells on top of it was placed on a custom-made holder. About 40 mL of imaging buffer (10% (w/v) glucose in 50 mM Tris (JT4109, Avantor), 50 mM NaCl (S271-500, Fisher Chemical), 10 mM MEA (M6500, Sigma-Aldrich), 50 mM BME (M3148, Sigma-Aldrich), 2 mM COT (138924, Sigma-Aldrich), 2.5 mM PCA (37580, Sigma-Aldrich) and 50 nM PCD (P8279, Sigma-Aldrich), pH 8.0) was added on top of the coverslip. Then another coverslip was placed on top of the imaging buffer. This coverslip sandwich was then sealed with two-component silicon dental glue (picodent twinsil speed 22, Dental-Produktions und Vertriebs GmbH).

SMLM imaging was performed on a custom-built setup on an Olympus IX-73 microscope stand (IX-73, Olympus America) equipped with a 100x/1.35-NA silicone-oil-immersion objective lens (FV-U2B714, Olympus America) and a PI-FOC objective positioner (ND72Z2LAQ, Physik Instrumente). Samples were excited by a 642 nm laser (2RU-VFL-P-2000-642-B1R, MPB Communications), which passed through an acousto-optic tuneable filter (AOTFnC-400.650-TN, AA Opto-electronic) for power modulation. The excitation light was focused on the pupil plane of the objective lens after passing through a filter cube holding a quadband dichroic mirror (Di03-R405/488/561/635-t1, Semrock). The fluorescent signal was magnified by relay lenses arranged in a 4f alignment to a final magnification of ∼54 and then split with a 50/50 non-polarising beam splitter (BS016, Thorlabs). The split fluorescent signals were delivered by two mirrors onto a 90 °C specialty mirror (47-005, Edmund Optics), axially separated by 580 nm in the sample plane, and then projected on an sCMOS camera (Orca-Flash4.0v3, Hamamatsu) with an effective pixel size of 120 nm. A bandpass filter (FF01-731/137-25, Semrock) was placed just before the camera.

2.2.4 Reconstruction software

Image stacks were processed using ThunderSTORM, a plug-in for ImageJ chosen for its customisability and ease of use, which couples a multitude of proceeding and post-processing methods with a simple and interactive interface [40]. There are a multitude of filtering software choices on offer – by default, non-salient image features are filtered by using wavelet analysis; we employ the Difference-of-Gaussians

methodology. Many feature-enhancing methods tend to boost high-frequency signal; however, background noise also tends to be of high spatial frequency, meaning the appearance of artefacts is increased as well. The difference-of-gaussians filter, by contrast, is uniquely apt at removing unwanted ranges of high frequency details. When convolved with a gaussian of a discrete standard deviation, information of high corresponding spatial frequency is suppressed from an image. However, if the image is convolved with two normalised gaussian kernels of differing standard deviations, a specific range of spatial-frequencies can be extrapolated in subtracting one convolved image from the other. Thus, information of unwanted spatial frequencies can be removed from an image, and other information made more salient, by choosing appropriate standard deviations for these gaussian functions [117]. For our filtering, we convolved the images with kernels of standard deviations 1.6 pixels and 1.0 pixels before visualising their difference. All acquisitions underwent ThunderSTORM analysis as performed using the SE and ME methods. Settings were as follows: Filter: Difference of Gaussians (min 1.0, max 1.6 pixels), Detector method: local maximum (connectivity 8), PSF model: Integrated Gaussian, Estimator: Maximum likelihood, fitting radius: 3. For ME fitting additional/alternate parameters were fitting radius: 5, max molecules: 5, p-threshold: 0.05.

SOFI analysis was used using the published MATLAB script which performs the calculation to fourth-order. In all cases, the balanced output was used. SRRF analysis was performed using the ImageJ plugin using the default parameters. Compressed sensing was performed using the published MATLAB script. The model PSF was set to match that used in the simulations. DeconSTORM was also performed using MATLAB with input parameters for fluorophore blinking were set to match the simulations. Due to computational limitations, the convolution method was used over the matrix method. Where HAWK preprocessing was used, identical parameters for the subsequent localisation step were used as in the case where HAWK was not used. The only additional processing was the filtering of false positive localisations by fitted width, following the procedure described in the original publication.

Simulated data from the Localisation Microscopy Challenge were analysed similarly. For all analysis algorithms the same methods and settings were used as above, but making allowances for the differences in camera pixel size and gain. DeconSTORM was not used for this dataset, as the blinking properties of the simulated emitters were not known. Only ThunderSTORM was used on the LD simulation, as no significant emitter overlap should be present in this data. Experimental Localisation Microscopy Challenge datasets were also analysed in a similar manner.

For both SE and ME fitting with ThunderSTORM, all parameters were as above, except a magnification factor of 5 was used to render the images (corresponding to a 20 nm reconstruction pixel size). For SRRF with 'default parameters': Ring radius = 0.3, Axes in ring = 6, Temporal analysis = TRPPM. For SRRF using 'optimum parameters': Ring radius = 0.3, Axis in ring = 8, Temporal analysis = TRAC order 2. In both cases the magnification factor was 5.

Experimental sarcomere data and all remaining simulated datasets were also analysed using ThunderSTORM SE fitting using the same parameters as above. Here, the magnification was x10, corresponding to a 10 nm pixel size.

The clathrin-coated pit, mitochondria, and microtubule data were processed in as similar a way as possible, allowing only for differences between experimental setups. These datasets were all analysed using ThunderSTORM SE fitting, using the same parameters as the Localisation Microscopy Challenge data, with the exception of the differing camera pixel sizes these data sets were acquired on (160 nm for the microtubules and clathrin, 120 nm for TOM20). As the mitochondrial data was acquired on a biplane microscope, only a single plane was analysed.

2.2.5 Statement on data selection

To evaluate the performance of algorithms, representative examples were chosen from available data conditions. We chose the three experimental structures used to test HAWKMAN (microtubules, mitochondria, and clathrin-coated pits) because they represent three commonly-found 'categories' of structure shape as seen in two dimensional imaging (fibrillar, two-dimensional, and small pores respectively). These structures cause SMLM algorithms to fail in different, but predictable ways (premature crossover, sharpening, and structure collapse in particular). However, it is important to acknowledge the potential risk of cherry-picking of data during the comparison process. While efforts were made to ensure a diverse and useful representation of different localisation and assessment algorithms operating on a range of structures, the reconstructions examined will of course not capture the full range of possible scenarios. We frequently used data from the Localisation Microscopy Challenge as it is a commonly-employed benchmark for SMLM algorithms, making it an exceptionally valuable tool in the interests of fair comparison.

2.3 Results

2.3.1 The HAWKMAN algorithm

We exploit the accuracy and reliability of HAWK to identify potential artefacts in a localisation microscopy image produced without HAWK, and to indicate where HAWK pre-processing has reduced localisation precision sufficiently that underlying fine structure could have been made unresolvable. This is achieved by quantifying structural differences between the original image and the HAWK-processed reconstruction as produced using the same algorithm. This measure can be used to map out areas which have artefacts, such as artificial sharpening, and to ascribe a confidence level to local regions of the image at the sub-diffraction level. A measure of the local resolution (in the sense defined above of the length scale at which structure is correctly reproduced), can be ascertained by progressively blurring the input images with a Gaussian kernel for longer length scales and repeating the comparison. The length scale at which reasonable agreement (quantified by the degree of local correlation) between the HAWK-preprocessed and non-HAWK-processed output images in the local region is achieved indicates the local resolution achieved. From this, a map of the maximum scale of artefacts in the image can be produced.

Figure 2.1 exhibits HAWK's ability to render a reconstruction artefact-free using pre-processing at multiple densities. Without HAWK pre-processing, we see that the diameter of the mitochondrial pore-like structure being imaged is heavily dependent on the emission density, with lower emission densities corresponding to a greater diameter. Certainly, even with HAWK pre-processing, the structure has a poorer resolution at higher emitter densities. However, the pore never altogether vanishes or even noticeably shrinks in size when HAWK has been applied, thus preserving the fidelity of the structure.

HAWKMAN takes as input data a test (super-resolved reconstructed) image and reference (HAWK-preprocessed) image, and a maximum length scale over which the performance of the algorithm will be evaluated. The input images are intensity-flattened to suppress the influence of isolated outlying high intensity points (due to repeated sampling), which frequently occur in SMLM. The flattened images are then blurred with a Gaussian kernel of width equivalent to the current length scale of interest (ranging in integer multiples of a single pixel up to a user-specified maximum). These blurred images are then binarised according to a length scale-specific adaptive threshold. This produces two images for each input: one (the sharpening map) is produced from a threshold of roughly 50% of the local maximum, the other (the structure map) is produced from a higher threshold (85%) and subsequently

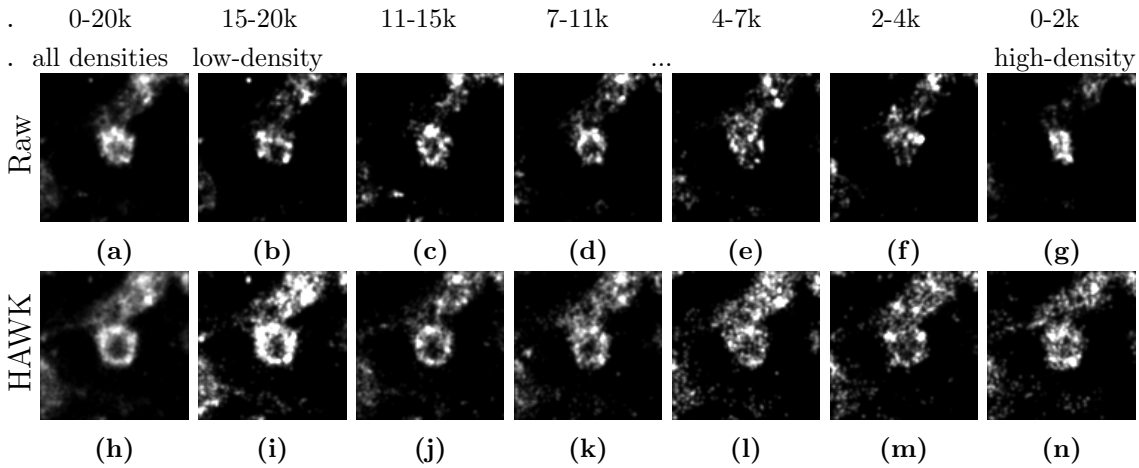


Figure. 2.1. HAWK removes artefacts from SMLM datasets of the mitochondrial epitope TOM20 stained with Alexa Fluor® 568 for a range of emission densities. (a-g) display zoom-ins of a mitochondrion that should form a pore-like structure when imaged. (a) displays the reconstruction from the full acquisition (all 20,000 frames), whilst (b-g) display the reconstruction at different frame intervals. The density starts very high but decreases as the acquisition goes on and more fluorophores are bleached. In the absence of HAWK, the 'radius' of the circle reduces as density increases before collapsing to single structure due to artificial sharpening. With HAWK pre-processing (h-n), the size of the pore remains constant no matter which frame intervals are chosen – only the resolution & overall quality of the image reduce with increasing activation density.

skeletonised and then re-blurred with the same Gaussian kernel. It should be noted that the optimum thresholds will vary slightly depending on the local dimensionality of the sample, but these variations are not critically important. A higher threshold may be required if there is only a small difference in the labelling density of adjoining structure, such as may result from non-specific labelling. The optimum value is only required to resolve these adjacent structures, default parameters will still identify strong sharpening artefacts. Default parameters are used for all data presented here except for the clathrin-coated pit data (Figure 2.13).

Differences highlighted in the sharpening and structure maps indicate areas of artificial sharpening, bias and/or large differences in precision. The local correlation between them is used as a confidence metric for the reliability of that region in the test reconstruction, giving the confidence map (Figure 2.2). The procedure is repeated for increasing length scales up to the maximum of interest (typically the instrument PSF). This allows the accuracy of the localisation algorithm to be assessed across the whole range of length scales. This approach bears some similarities with the Gaussian scale-space technique [118], that has been used in the fields of image processing and computer vision. In this technique, the input image is also examined simultaneously at different scales by convolution with Gaussian kernels of different sizes. Images can then be further processed on structure only of a specific scale.

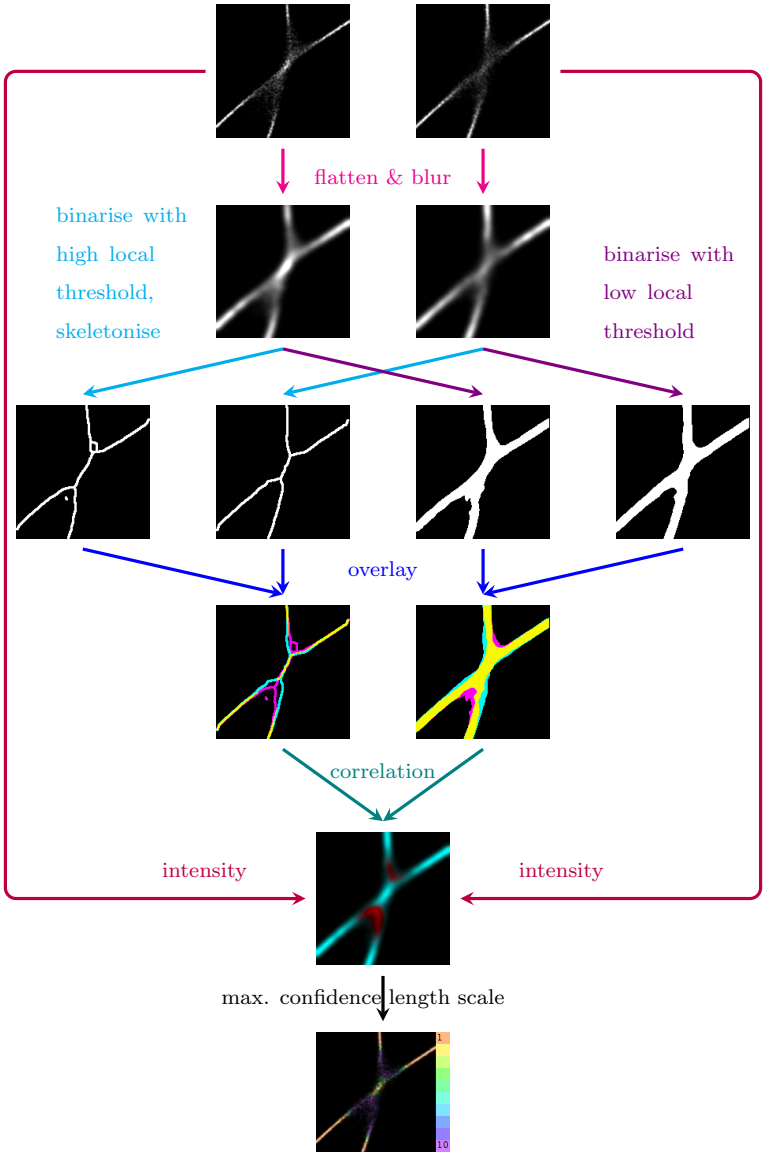


Figure. 2.2. The processing steps of the HAWKMAN algorithm, producing the sharpening map, structure map and confidence map. For the sharpening and structure maps, areas of input image agreement are displayed in yellow, whereas magenta and cyan indicate structure only present in the test and reference images respectively. Therefore, magenta indicates artificial sharpening and cyan indicates missing structure. The confidence map indicates the calculated local correlation score (see Methods for details), ranging from 0 (magenta) – 1 (cyan), whilst brightness denotes the intensity of the original images. Repeating the process at many length scales can indicate how the size of artefacts varies over the structure as indicated in the artefact scale map. Scale bar 500 nm.

Our approach is to compare *only* the structure information in the test and reference input reconstructions at each scale to detect which scales contain disagreement and therefore density-induced artefacts.

As an optional further stage, a map of the largest scale artefact identified at each point in the image can be produced from the confidence map. This, the artefact scale map, consists of the combined input image colourised to reflect the smallest length scale that that part of the image has high reported confidence. This enables a rapid appraisal of which parts of the reconstruction contain acceptably accurate structure and which do not. Although the artefact scale map provides a useful summary of the HAWKMAN output. The other maps better indicate the magnitude of artefacts at each scale and the confidence map gives an indication of how the structure would appear at a resolution known to only contain accurate information.

In summary, a user should interpret the various HAWKMAN maps in the following way:

- The **sharpening map** resembles a colourised image of binarised overlays with relatively low local thresholds. Here, magenta represents strucure that only appears in the non-HAWK'd ('test') image, giving an indication of false structure. Cyan structures appeared only in the HAWK'd ('reference') image, giving an indication of missing structure in the test reconstruction. Yellow structures are present in both images – because the failure modes of the test and reference images are different (section 1.5.3), agreement between the two indicates structure that is resolved and trustworthy. As the name indicates, the sharpening map is best used for detecting artificially 'thinned' structures. These will frequently mislead other SMLM appraisal software and amateur experimenters, since thinner structures have a higher apparent resolution. The sharpening map is generated for a set of user-specified length scales. The user can scroll through each to find the length scale at which agreement (yellow) is reached for a local area of interest in the image to quantify the severity of sharpening.

- The **structure map** resembles a sharpening map to which a higher local threshold has been applied, resulting in a map of 'wiry', skeletonised structures. It uses the same colour scheme as the sharpening map. This more clearly allows the user to visualise structure collapse, as in the sharpening map, the low threshold may cause evidence of this to be blotted out. Instances of structure collapse will resemble magenta skeletons located in the space between cyan and/or yellow skeletons. The structure map is calculated for the same set of length scales as the sharpening map, and the severity of structure collapse can be quantified using the same method of scrolling through them to

find structural agreement (yellow).

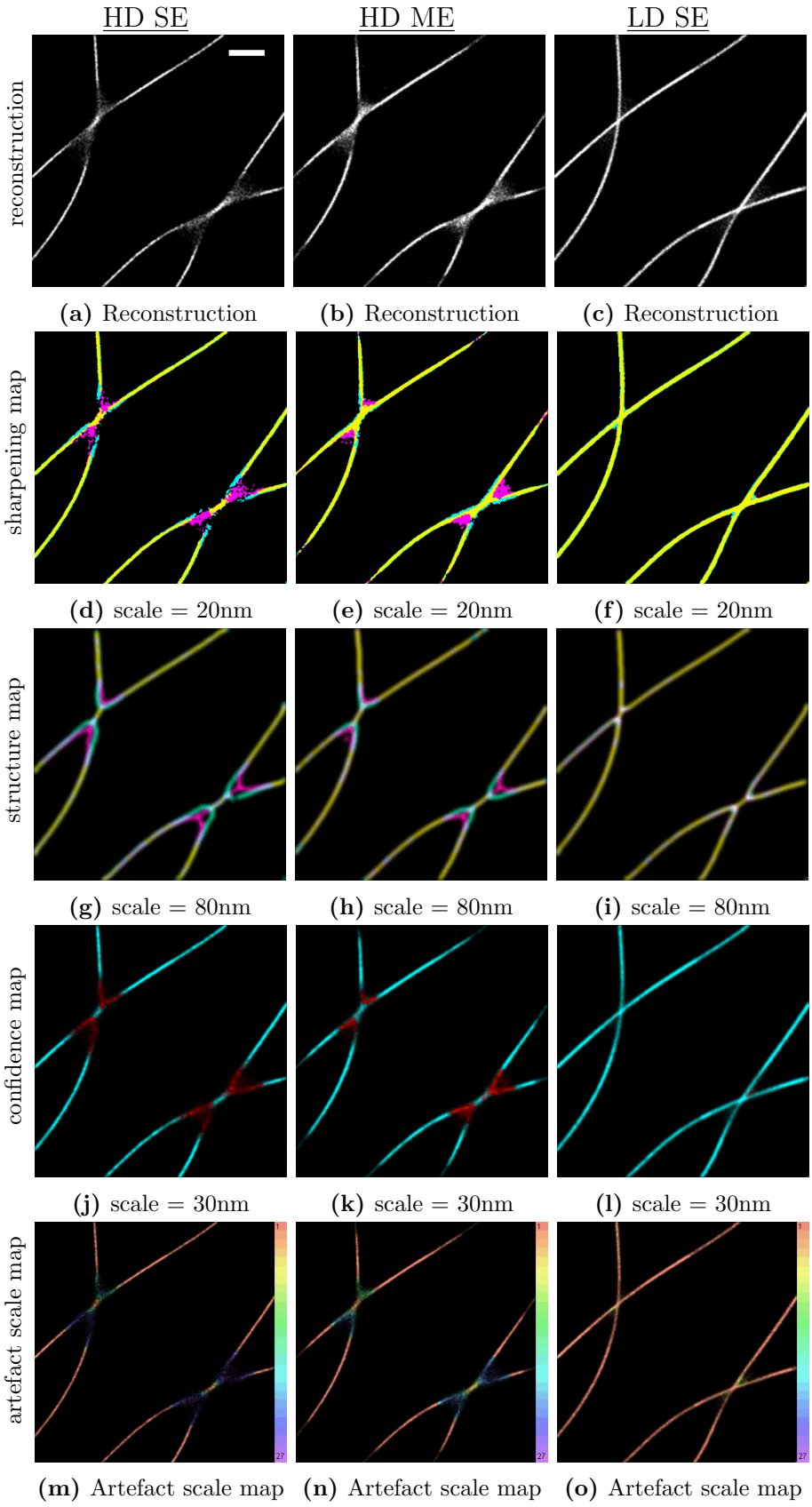
- The **confidence map** is a colourised summing of the two imput images and is calculated using the results from the previous two maps, displaying results for the same length scales. It allows the user to see *where* the test image contains trustworthy structures at each length scale, with cyan structures being most likely trustworthy and red structures being most likely artefactual. This can be used to obtain a measure of 'functional resolution'. That is, for a local region of the image containing structures of interest, the smallest measured length scale at which high confidence (cyan) is measured can be said to be the *reliable resolution* of the structure in that region.

- The **artefact scale map** summarises the results from the confidence map without requiring the user to scroll through the calculated length scales, resulting in an easy-to-assess 2D colourised map of the inputs. It simply colourises each bit of structure according to the *length scale at which it became reliable in the confidence map*. This allows the user to make fast and easy appraisals of their image/localisation algorithm.

2.3.2 Testing with benchmark data

In order to test the performance of HAWKMAN we selected simulated datasets from the Localisation Microscopy Challenge [112, 115] where both low and high emitter density datasets were available. This allowed us to benchmark the performance of HAWKMAN when using a HAWK reference image, against when using a low-density (GT) reference image, which ensured that the location and scale of artefacts in the test image was known. Figure 2.3 shows the results of HAWKMAN analysis on simulations of microtubules for the Localisation Microscopy Challenge for both SE and ME Gaussian fitting of the data. For areas of microtubule crossover (Figure 2.3a-c) the high-density reconstructions (Figure 2.3a, 2.3b) show substantial sharpening artefacts not present in the reconstruction from low-density data (Figure 2.3ac). As expected, the SE reconstruction (Figure 2.3a) is more severely affected than the ME (Figure 2.3a). Figure 2.3a: HAWKMAN analysis of simulated microtubule data from the Localisation Microscopy Challenge.

HAWKMAN analysis was performed on the Microscopy Challenge data sets, with the reference image being produced by the same algorithm on the same data but with prior HAWK pre-processing. Sharpening maps (Figure 2.3d-f), structure maps (Figure 2.3g-i) and confidence maps (Figure 2.3j-2.3l) are shown for the same length scales for each test image. The length scales in this figure (and all subsequent ones)



(a) Reconstruction **(b)** Reconstruction **(c)** Reconstruction

(d) scale = 20nm **(e)** scale = 20nm **(f)** scale = 20nm

(g) scale = 80nm **(h)** scale = 80nm **(i)** scale = 80nm

(j) scale = 30nm **(k)** scale = 30nm **(l)** scale = 30nm

(m) Artefact scale map **(n)** Artefact scale map **(o)** Artefact scale map

Figure. 2.3

Figure. 2.3. A comparison of SE and ME fitting of high emitter density data and equivalent LD data. A section of the reconstructions is shown in (a-c) for HD:SE fitting, HD:ME fitting and LD:SE fitting, respectively, which contain artificial sharpening for both HD methods. The results of HAWKMAN analysis are shown in (d-l) for each of the three test reconstructions with HAWK used as the reference. The sharpening maps (d-f) and structure maps (g-i) are shown at a length scale of 20 nm and 80 nm, respectively. In both the SE and ME cases the magenta areas (test image only) in the sharpening maps (d, e) and the structure maps (g, h) indicate substantial sharpening along with missing structure (cyan), which is more severe (as expected) in the SE case. Both these effects are absent for the LD reconstruction (f, i), which serves as a proxy for the ground truth. These results are reflected in the confidence maps (j-l, 30 nm scale). These show substantially reduced confidence (strong artefacts, highlighted in red) for both HD methods (j, k), whereas for the LD data the confidence (l) is high everywhere. The differences in accuracy of the three reconstructions is further highlighted by the corresponding artefact scale maps (m-o). These show the maximum scale of artefacts is larger for the SE reconstruction than the ME and that the LD is virtually artefact-free. The colour scale ranges from length scales 1 (10 nm red) to 27 (270 nm purple) reconstruction pixel. Scale bar 500 nm.

were selected to give a clear and concise representation of the full HAWKMAN output. A more detailed description of how to use each length scale is given in the methods (section 2.2). Essentially, as the length scale of the comparison is increased, less disagreement between the test and reference images will be observed. The structure and sharpening maps indicate areas of the reconstruction that contain artefacts larger in size than the length scale under investigation. The confidence map quantifies the discrepancies with areas of full confidence (score = 1, coloured cyan), which indicate HAWKMAN detects no artefacts larger in size than the current length scale. By examining all length scales, the user can ascertain the maximum scale of artefacts in every region of the sample. The sharpening and structure maps indicate to what degree these result from sharpening or structural bias.

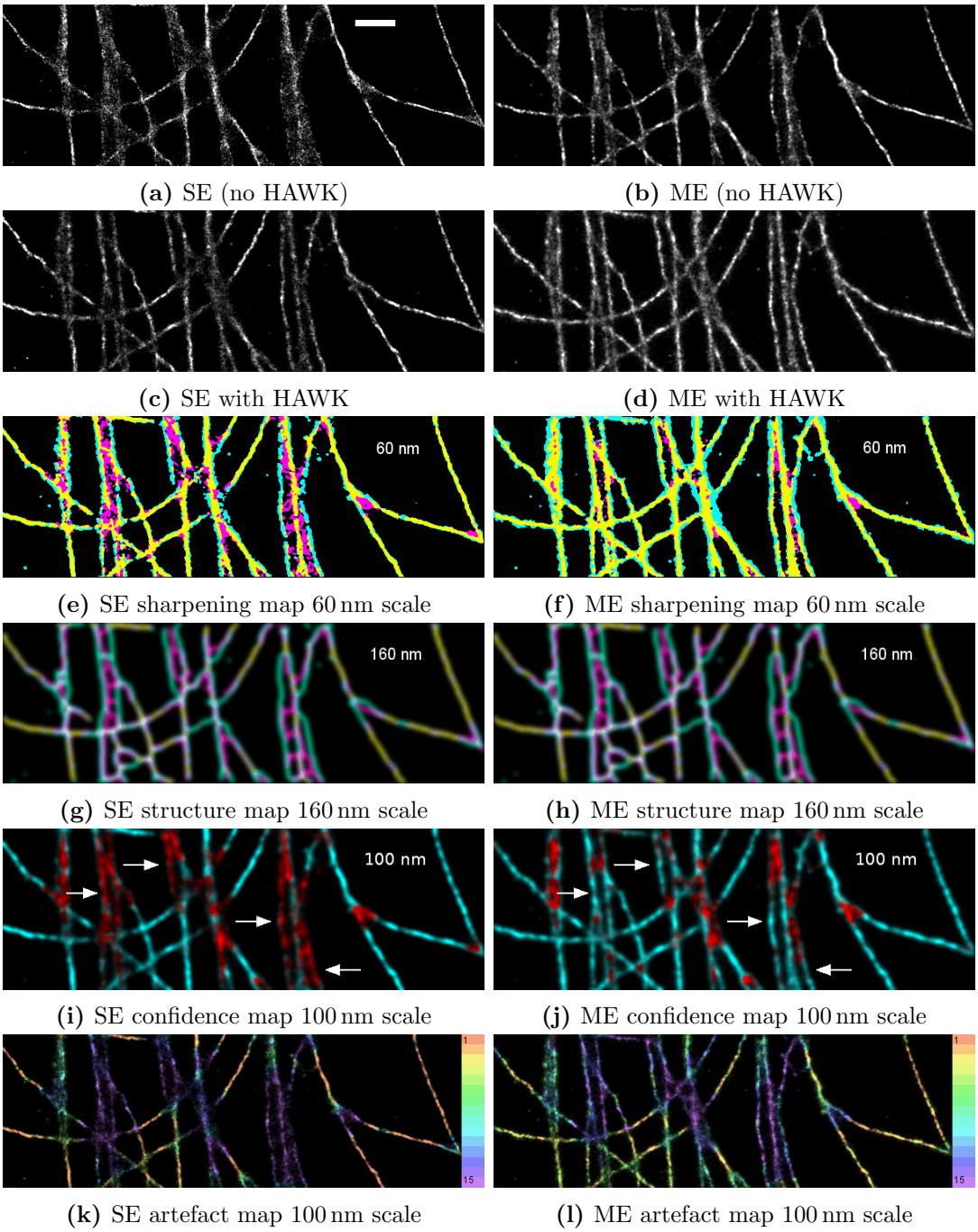
For the HD reconstructions, both the sharpening and structure maps show substantial biases in the microtubule positions, indicated by the magenta (structure that only appears in the test image) where the microtubules cross. In these maps, cyan areas (structure that only appears in the reference image) indicate some missing structure for the HD images not present in the LD reconstruction. The structure maps indicate these persist at even quite large (80 nm) length scales. This 'pinching in' of crossing microtubules is a commonly observed artefact in HD data [67, 78, 81, 119]. The estimated degree of error is quantified in the confidence maps (low confidence marked in red, high confidence in cyan) for both SE (Figure 2.3j) and ME (Figure 2.3k) methods. These indicate ME fitting is more accurate than the SE fitting, but still produces substantial errors. The corresponding confidence map for the same structure simulated at a low emitter density (Figure 2.3l) cor-

rectly indicates no sharpening artefacts are present at this length scale (30 nm). The artefact scale maps (Figure 2.3m–o) summarise the HAWKMAN output by showing the largest artefacts present at each point of the image. These show that both high-density reconstructions contain large artefacts of a scale approaching the PSF near the crossing points whereas even the largest artefacts in the low-density reconstruction are much smaller in scale.

The Localisation Microscopy Challenge datasets also contain experimental data, and we tested the performance of HAWKMAN on HD experimental microtubule datasets [112, 115]. These were analysed using both SE & ME Gaussian fitting [40], giving results with what appear to be common artefacts (Figure 2.4a, b) in the regions of clustering and microtubule crossover including blurring, artificial sharpening and unresolved/collapsed structure. The corresponding reconstruction when HAWK is used (Figure 2.4c, d) show a much more clearly resolved structure in these regions, though some precision is lost. The output of HAWKMAN analysis indicates that these are highly sharpened reconstructions. The sharpening maps (Figure 2.4e, f), structure maps (Figure 2.4g, h) and confidence maps (Figure 2.4i, j) indicate the presence of severe sharpening artefacts even at quite large length scales (60, 160 and 100 nm, respectively).

Examining the reconstruction produced by ThunderSTORM ME fitting [40] (Figure 2.4b), it can easily be seen that this is a superior reconstruction to SE fitting (Figure 2.4a) – but some structures are still not clearly resolved. The confidence map produced by HAWKMAN analysis for this reconstruction is shown in Figure 2.4i (for the same 100 nm scale). The map shows much higher confidence in the overall accuracy of this reconstruction at this scale, yet still highlights some areas where sharpening is present. HAWKMAN clearly demonstrated the presence and scale of artificial sharpening and other artefacts. The maximum size of artefacts at each point in the input image is shown in the artefact scale maps (Figure 2.4k, l) ranging from length scale of 1 pixel (20 nm, red) to 15 pixels (300 nm, purple) showing these are larger in the SE reconstruction.

It should be noted that specialised HD localisation algorithms frequently allow for parameter optimisation, generally at the expense of increased processing times. The output reconstructions can be significantly improved with optimised parameter values [77, 78], and HAWKMAN can be used to clearly assess the magnitude of such improvements. Figure 2.5 demonstrates, with the example of SRRF, how HAWKMAN may aid in the optimisation of input parameters. The relative fidelity of reconstructions using different parameter sets can be assessed by using HAWKMAN on each of them. The results for SRRF with default parameters taken together



(a) SE (no HAWK) **(b)** ME (no HAWK)

(c) SE with HAWK **(d)** ME with HAWK

(e) SE sharpening map 60 nm scale **(f)** ME sharpening map 60 nm scale

(g) SE structure map 160 nm scale **(h)** ME structure map 160 nm scale

(i) SE confidence map 100 nm scale **(j)** ME confidence map 100 nm scale

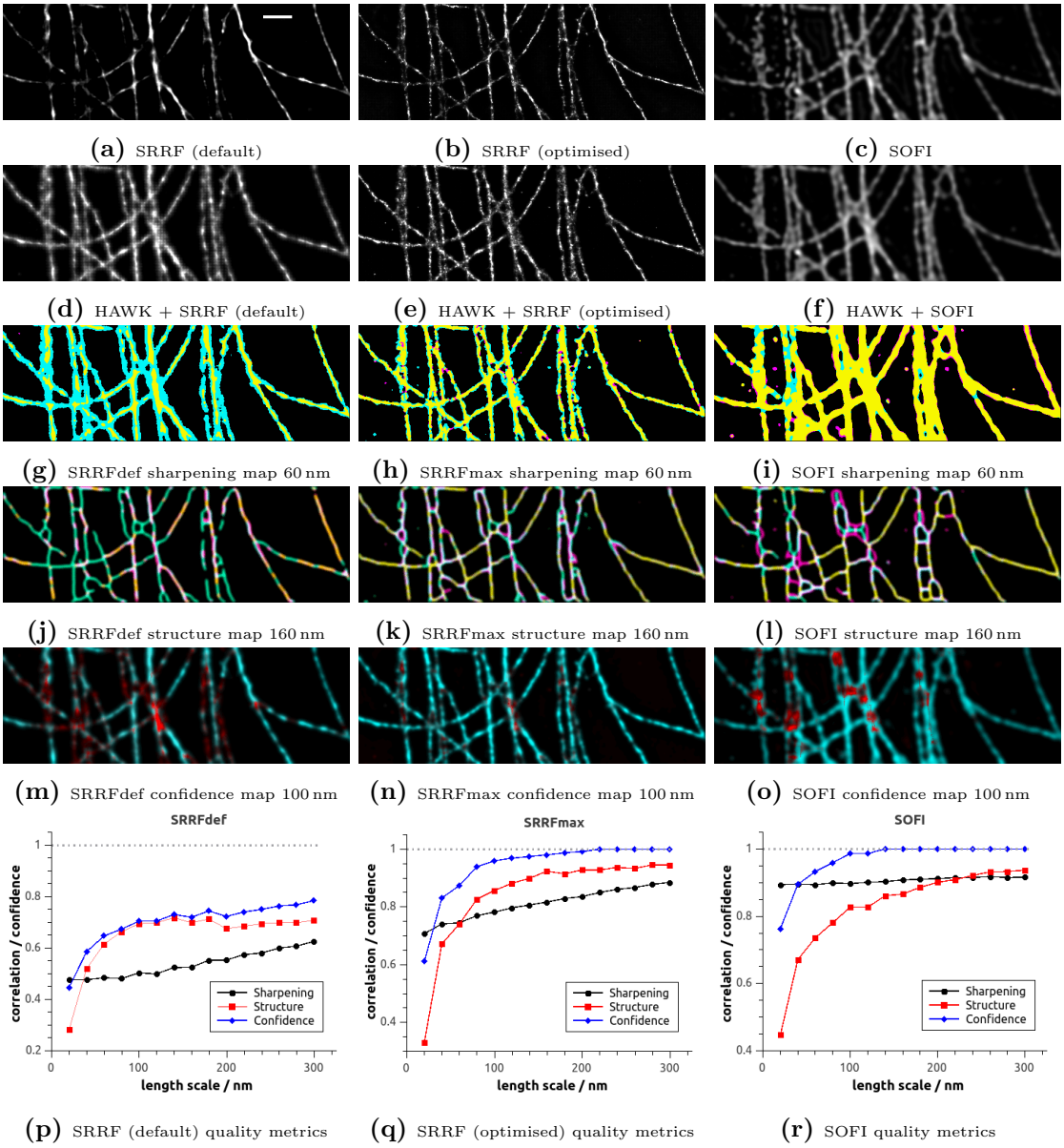
(k) SE artefact map 100 nm scale **(l)** ME artefact map 100 nm scale

Figure. 2.4

Figure. 2.4. Results of HAWKMAN analysis on experimental microtubule data from the Localisation Microscopy Challenge comparing SE and ME fitting. The reconstructions for SE and ME fitting without (a, b) and with (c, d) HAWK. The results of HAWKMAN analysis are shown in the sharpening (e, f), structure (g, h), and confidence (i, j) maps for example length scales (in white). The SE reconstruction (a) shows substantial sharpening artefacts. The differences in structure between this reconstruction and the results of applying HAWK with SE fitting are highlighted by the sharpening map (e) shown at an example scale of 3 reconstruction pixels (i.e. 60 nm for 20 nm pixels). Substantial bias in the SE reconstruction where microtubules are in close proximity/crossing is highlighted in the structure map (g) even at a scale of 60 nm. Reliable regions in the reconstruction where both methods agree are highlighted in the confidence map (i), shown at the 100 nm scale. ME fitting (b) gives a more accurate reconstruction, however, it still contains numerous artefacts (f, h, j). The ME confidence map (100 nm scale) (j) shows a reduction in the severity of artefacts but some remain. The white arrows highlight some structures accurately reported in the ME result (j) that were inaccurate in the SE case (i) at this length scale. A summary of the HAWKMAN results at all length scales is shown in the artefact scale maps (k, l) which indicate the maximum scale of artefacts at each local region (colour indicates length scale, at 20:20:300 nm). Comparison of the SE (k) and ME (l) scale maps shows that for ME fitting the maximum scale of artefacts is reduced in areas of high microtubule density. Scale bar 1 µm.

(Figures 2.5g, j, m, p) show HAWKMAN is indicating the presence of substantial artefacts at all length scales. The sharpening and structure maps (Figures 2.5g, j) suggest that these arise not just from collapse of adjacent structures to one, but also from the highly disjointed & punctate nature of the reconstructed microtubules. This is also caused by the structure in the test being much finer than in the reference image (Figures 2.5a,d). HAWKMAN results for SRRF with optimised parameters (Figures 2.5h, k, n, q) suggest this reconstruction is of a much higher quality. The test image (Figure 2.5b) has much more continuous structure and individual microtubules are better resolved in the typical problem areas of high structural density. In this case the structure map (Figure 2.5k) shows small differences in structure rather than just missing structure, and the sharpening map (Figure 2.5h) shows only slight differences that are restricted to the problem areas. This suggests the precision estimated from the width of the structures would be much more accurate. From the confidence maps and scores (Figures 2.5n, q) HAWKMAN is indicating this reconstruction is fairly accurate at the 100 nm scale. Results for the SOFI reconstruction (Figures 2.5i, l, o, r) show very little sharpening (Figure 2.5i), and only slight structural differences. This suggests the artefacts consist of disjointed rather than collapsed structure, again as a consequence of the highly non-linear relationship with intensity. The whole image confidence score suggests an accurate reconstruction down to the 100 nm scale.

HAWKMAN analysis on microtubule data from the Microscopy Challenge are



(a) SRRF (default) (b) SRRF (optimised) (c) SOFI

(d) HAWK + SRRF (default) (e) HAWK + SRRF (optimised) (f) HAWK + SOFI

(g) SRRFdef sharpening map 60 nm (h) SRRFmax sharpening map 60 nm (i) SOFI sharpening map 60 nm

(j) SRRFdef structure map 160 nm (k) SRRFmax structure map 160 nm (l) SOFI structure map 160 nm

(m) SRRFdef confidence map 100 nm (n) SRRFmax confidence map 100 nm (o) SOFI confidence map 100 nm

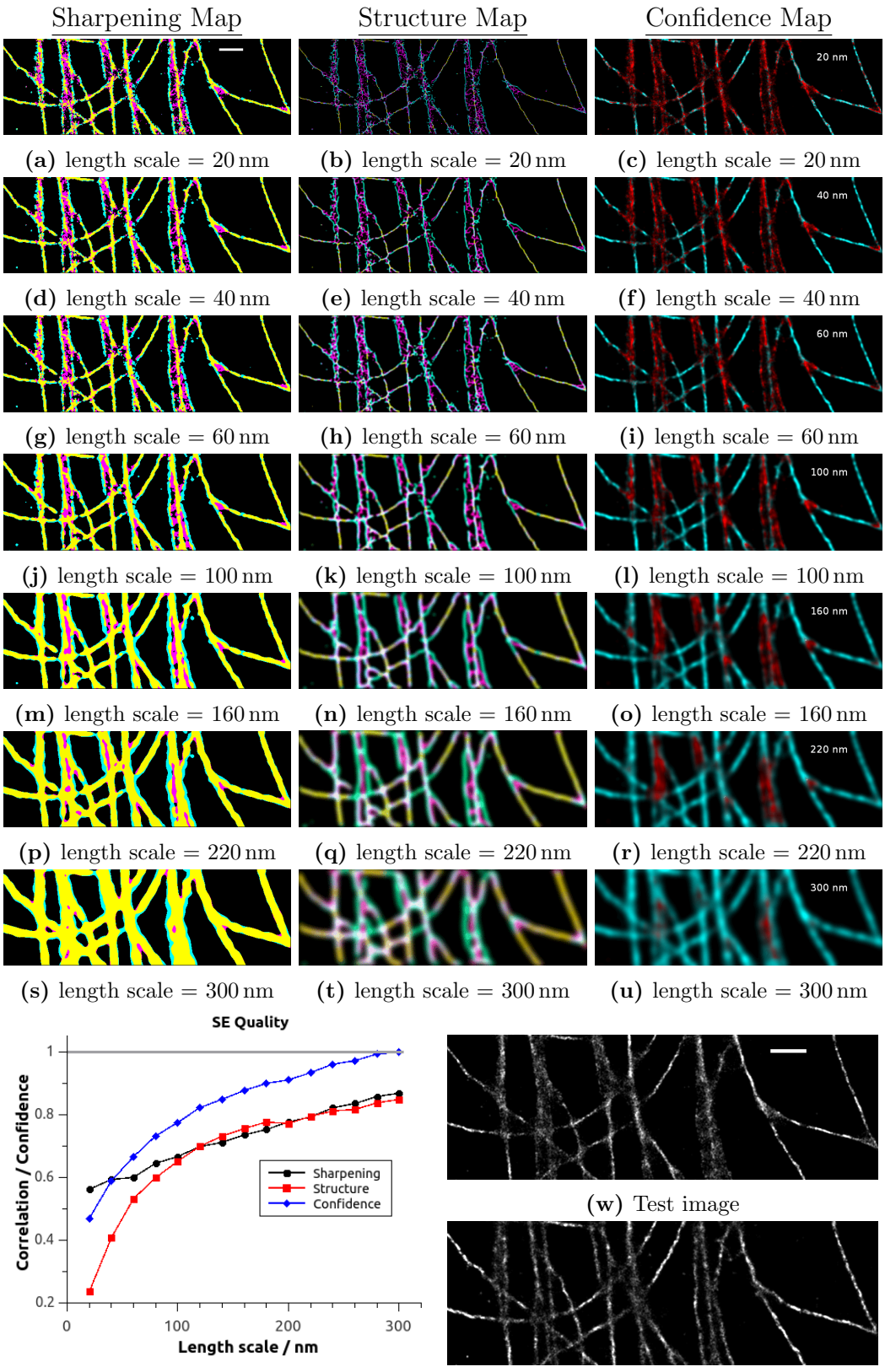
(p) SRRF (default) quality metrics (q) SRRF (optimised) quality metrics (r) SOFI quality metrics

Figure. 2.5

Figure. 2.5. HAWKMAN recognises that SRRF performs better on HD data with optimised parameters. This analysis was performed on microtubule data from the Localization Microscopy Challenge using SOFI and SRRF as the test algorithm. (a-c) Show the un-HAWK'd reconstructions produced by the SRRF and SOFI algorithms, with (a) displaying reconstruction produced by SRRF using the default parameters ('SRRFdef') and (b) showing the SRRF reconstruction using optimised parameters ('SRRFmax') including the Temporal Radiality Auto-Correlations (TRAC) option. Under appropriate conditions, the optimised parameters should yield a higher resolution reconstruction. (c) Reconstruction produced by 4th order SOFI analysis. (d-f) The corresponding reconstructions of each of these methods with prior HAWK processing for SRRFdef, SRRFmax and SOFI respectively. (a-c) were used as the test images and (d-f) as the reference images. (g-o) Selected results for HAWKMAN analysis using these algorithms with and without HAWK for test and reference images respectively. (g-i) show the sharpening maps for length scale of 3 pixels (60 nm), (j-l) the structure maps for length scale 8 pixels (160 nm), and (m-o) the confidence maps for length scale 5 pixels (100 nm). The corresponding quality metrics for the whole image for each algorithm are shown in (p-r). These consist of the correlation between test and reference sharpening maps (black), structure maps (red) and the corresponding confidence score calculated from these values (blue). Scale bar 1 µm

displayed in Figure 2.6. This figure displays HAWKMAN analysis at multiple length scales, exhibiting how a user might assess local regions of an image using the sharpening, structure, and confidence maps. All three maps indicate substantial sharpening artefacts where microtubules are either closely spaced or crossing, as would be expected when the emitter density is too high. The first three scales (up to 60 nm, Figure 2.6a-i) show similar results, all indicating the degree of artificial sharpening in regions of such structural proximity is quite large, but limited elsewhere. At the 160 nm scale (Figure 2.6m-o), better agreement between test and reference is noted for both the sharpening and structure maps in some of the regions that were 'problem areas' at lower length scales. The centre and leftmost microtubule bundles are now considered reliable parts of the reconstruction, as indicated in the corresponding confidence map. It should be noted the HAWKMAN can detect these artefacts though the microtubules are not clearly resolvable in the HAWK reconstruction (Figure 2.6w). The largest analysed length scale (300 nm, Figure 2.6s-u) shows one area where the scale of sharpening in the test images is even greater than the PSF (270 nm). The whole image correlation graph (Figure 2.6v) can give an indication of the quality/reliability of the reconstruction as a whole at different scales. Looking at the correlation graph, we can see that in this case the reconstruction is judged to be reliable only at around 300 nm effective resolution.

This analysis is repeated for an ME fit (Figure 2.7). By comparing the two Figures 2.6 and 2.7, we can see HAWKMAN's ability to quantify the effectiveness of ME fitting with respect to SE fitting. At scales up to 60 nm (Figure 2.7a-i) the maps indicate that the ME reconstruction is only a modest improvement of the



(a) length scale = 20 nm **(b)** length scale = 20 nm **(c)** length scale = 20 nm

(d) length scale = 40 nm **(e)** length scale = 40 nm **(f)** length scale = 40 nm

(g) length scale = 60 nm **(h)** length scale = 60 nm **(i)** length scale = 60 nm

(j) length scale = 100 nm **(k)** length scale = 100 nm **(l)** length scale = 100 nm

(m) length scale = 160 nm **(n)** length scale = 160 nm **(o)** length scale = 160 nm

(p) length scale = 220 nm **(q)** length scale = 220 nm **(r)** length scale = 220 nm

(s) length scale = 300 nm **(t)** length scale = 300 nm **(u)** length scale = 300 nm

(w) Test image

(v) Correlation measure **(x)** Reference image (with HAWK)

Figure. 2.6

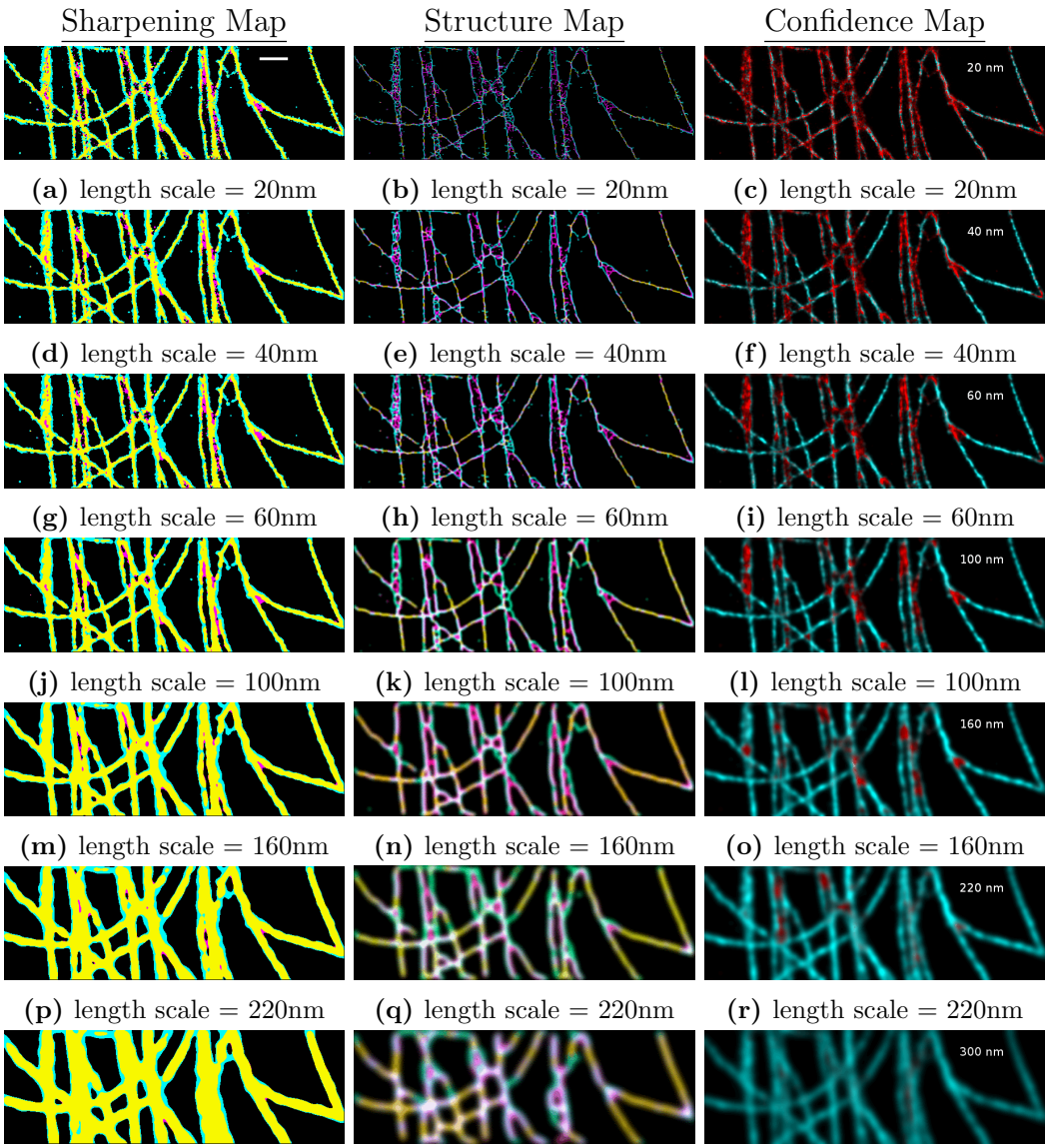
Figure. 2.6. HAWKMAN results for Single-Emitter (SE) fitting of microtubule data at various length-scales. (a-u) Sharpening, structure and confidence maps respectively for experimental microtubule data from the Localization Microscopy Challenge using the single emitter reconstruction as the test image and HAWK + SE as the reference. Results are shown for length scales 1, 2, 3, 5, 8, 11 & 15 pixels (20 nm, 40 nm, 60 nm, 100 nm, 160 nm, 220 nm & 300 nm) from top to bottom. The whole image correlation between the test and reference constituents of the sharpening and structure maps as a function of length-scale is shown in (v) (black & red respectively) along with the confidence metric used to colourise the confidence map (see methods) but calculated from the whole image (blue) rather than just the local correlation. The unprocessed test & reference images are shown in (w) & (x) respectively. Scale bars 1 µm.

SE case. However, at length scales larger than this (100-220 nm, Figure 2.7j-u) the improvement is substantial. The problematic areas of high structural density are indicated to be much more reliable. This can also be seen in the quality metrics (Figure 2.7a). Here, the confidence metric from the SE result from Figure 2.6 is also displayed (cyan), which shows a significantly lower score compared to the ME result (blue) for intermediate length scales. Comparing the ME (Figure 2.7w) and HAWK + ME (Figure 2.7x) reconstructions with their SE counterparts in Figures 2.6w, x shows it is in the regions indicated by HAWKMAN that the ME reconstruction is better resolved.

Tests on other simulated data, also from the Localisation Microscopy Challenge, demonstrate that HAWKMAN is clearly able to detect and even quantify sharpening and other artefacts. This is the case even at length scales substantially below the diffraction limit, right down to the scale of the localisation precision.

The advantage of using a HAWK image as the reference is that it does not contain the same bias common to many other algorithms. Figure 2.8 explores the use of alternative HD reconstruction algorithms to generate the reference image. Disagreement in the structure between the test (SE) and LD reference reconstructions is present either side of where the microtubules cross, indicating strong sharpening artefacts. Because the LD reference represents the most artefact-resistant acquisition, it serves as a representation of a 'perfect' reference image. Using a HAWK'd HD dataset as a reference image provides comparable results to using an LD (i.e. unsharpened/ artefact-free) dataset, in contrast to when other algorithms are used for the reference image. This confirms the suitability of HAWK processing for a qualitative and quantitative assessment.

When the ME fitting reconstruction is used as the reference (Figure 2.8b, f), the skeletonised images are almost identical and no significant artefacts are revealed. This is because both images are equally sharpened at this scale. When SOFI is used as the reference image (Figure 2.8c, g), performance is significantly better –



(a) length scale = 20nm **(b)** length scale = 20nm **(c)** length scale = 20nm

(d) length scale = 40nm **(e)** length scale = 40nm **(f)** length scale = 40nm

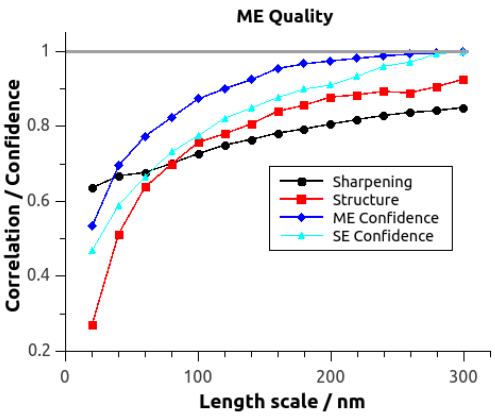
(g) length scale = 60nm **(h)** length scale = 60nm **(i)** length scale = 60nm

(j) length scale = 100nm **(k)** length scale = 100nm **(l)** length scale = 100nm

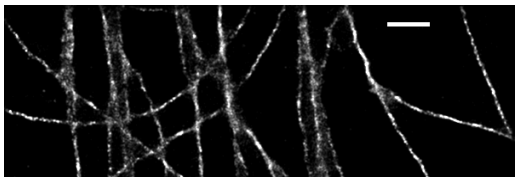
(m) length scale = 160nm **(n)** length scale = 160nm **(o)** length scale = 160nm

(p) length scale = 220nm **(q)** length scale = 220nm **(r)** length scale = 220nm

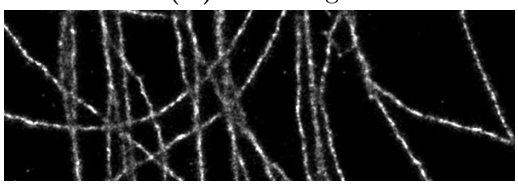
(s) length scale = 300nm **(t)** length scale = 300nm **(u)** length scale = 300nm



(v) Correlation measure



(w) Test image



(x) Reference image (with HAWK)

Figure. 2.7

Figure. 2.7. HAWKMAN results for Multi-Emitter (ME) fitting of microtubule data at the same length scales as Single-Emitter (SE) fitting above. The same analysis as presented in Figure 2.6 but with using the ME reconstruction as the test image and HAWK + ME as the reference. (a-u) The sharpening, structure and confidence maps at scales 20 nm (a-c), 40 nm (d-f), 60 nm (g-i), 100 nm (j-l), 160 nm(m-o), 220 nm (p-r), & 300 nm (s-u). The whole image correlation between the test and reference constituents of the sharpening and structure maps as a function of length-scale is shown in (v) (black & red respectively) along with the confidence metric used to colourise the confidence map (see methods) but calculated from the whole image (blue) rather than just the local correlation. (v) displays whole-image correlation between the test and reference constituents of the sharpening (black) and structure (ref) maps as a function of the length scale, as well as the confidence metric used to colourise the confidence map (blue). This time, it also displays the confidence metric from the SE result in Figure 2.6 in cyan, for comparison. The input test and reference images are displayed in (w) and (x) respectively. Scale bars 1 µm.

reference despite being (in principle) a lower resolution reconstruction. Again, this is is indicative of the importance of having different failure modes for the algorithm being used to generate the reference image and the test image. SOFI detects some of the sharpening present, but not the regions closer to the crossing point. Using the HAWK'd reconstruction for the reference (Figure 2.8d, h) gives results very similar to the LD case, revealing all of the sharpening artefacts present in the test image. We see similar results when the test image is reconstructed with ME fitting (Figure 2.8i-p). When a SRRF (Figure 2.8j, n) or compressed sensing (CS) reconstruction (Figure 2.8k, o) image are used as the reference, the full range of sharpening present is clearly not detected by the structure map. This is clear when comparing these to the LD-referenced structure maps (Figure 2.8i, m), which reveal sharpening much further away from the point of structure crossover. Again, using HAWK:ME for the reference image (Figure 2.8l, p) greatly outperforms other algorithms and gives results very similar to using the LD reconstruction for the reference, indicating detection of all artefacts.

2.3.3 The effect of non-linear transformation SMLM algorithms on image resolution metrics

Methods such as SOFI [75] and SRRF [77] attempt to obtain a super-resolved image from a series of input images in which fluorophore intensity fluctuates. The output of these methods is an image rather than a list of localisations (also a characteristic common to 3B images [76, 78]). The output of these algorithms is usually noticeably sharper than a z-projection of the stack, suggesting that the resolution is being improved by accessing information from fluorophore intensity fluctuations. Images

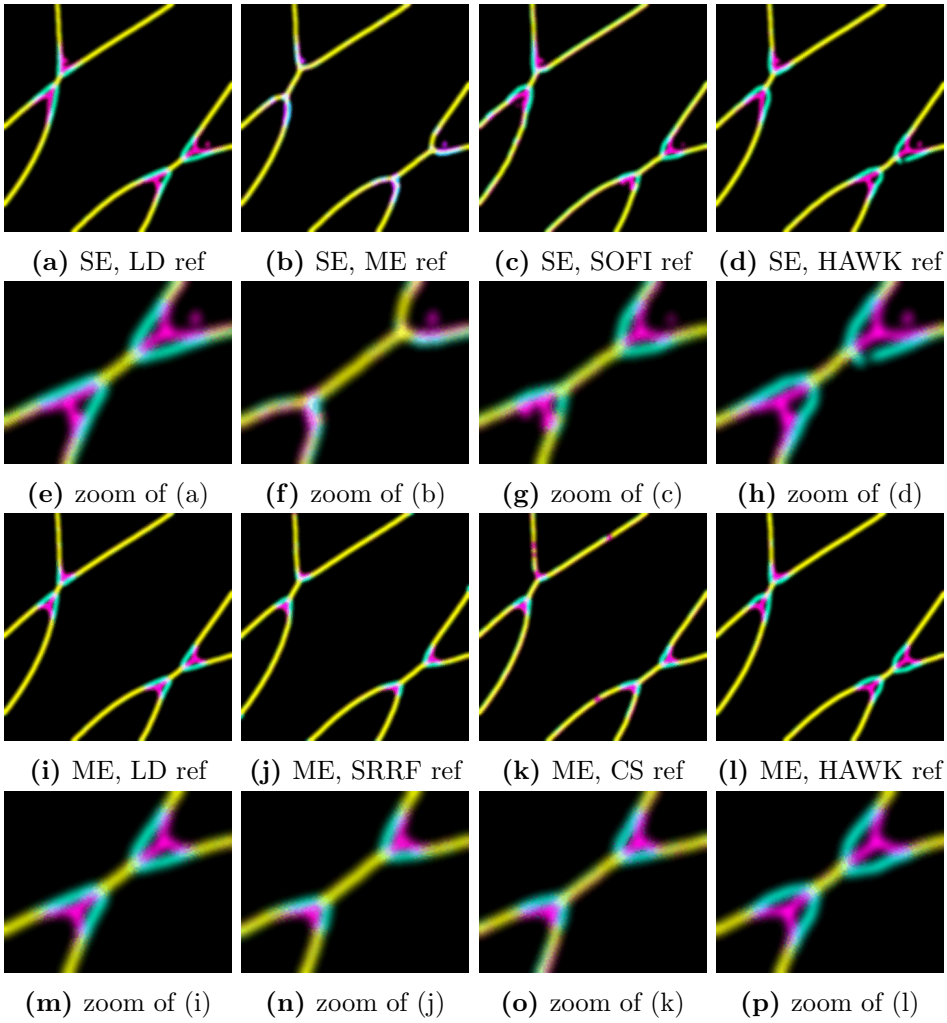
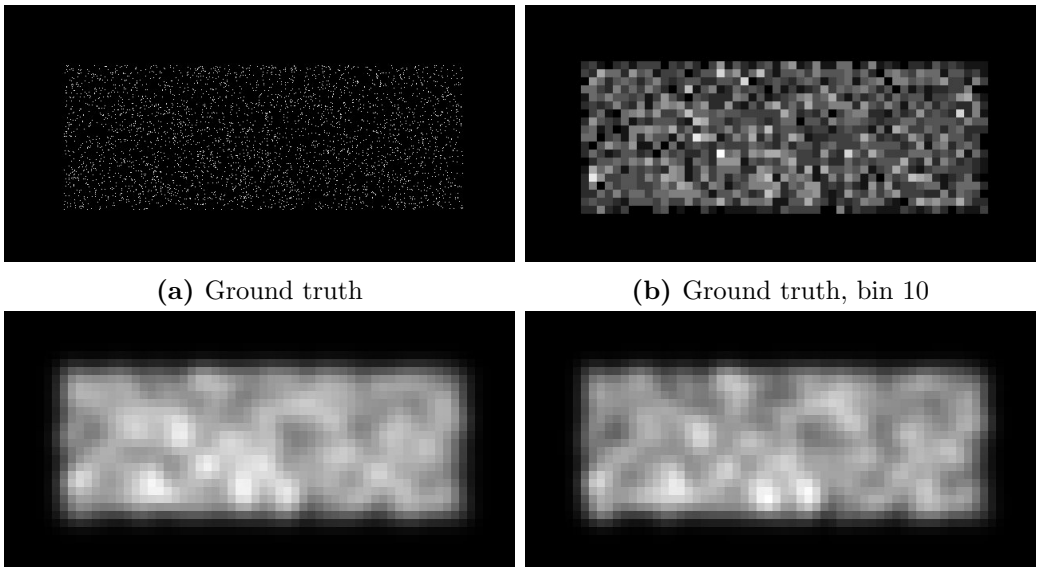


Figure. 2.8. Performance comparison when using alternative HD algorithms to HAWK to create the reference image. Here, the ability of HAWKMAN to identify sharpening artefacts is demonstrated on simulated microtubule-like data from the Microscopy Challenge. (a-d) show the skeletonised output of HAWKMAN tests on the SE fitting reconstruction using ThunderSTORM wherein an alternative set of reconstruction algorithms have been used to generate the reference image (ME fitting (b, f) and SOFI (c, g)). Enlarged regions of (a-d) are shown in (e-h). (i-p) are the same as (a-d), but with ME fitting used for the test image, and a different set of algorithms used for the test image (SRRF (j, n) and Compressed Sensing (k, o)). (a, e) Use of the LD SE reconstruction of the same structures for the reference image provides an accurate, artefact-free image for comparison. This represents the optimum standard of reference image, so any deviation between this image and the test image highlights sharpening artefacts present in the latter. Magenta denotes structure that only appears in the test image, cyan denotes structure that only appears in the reference image, and yellow denotes agreement between both images. Results are displayed for length scale of 8 super-resolution pixels (80 nm). Dilation (ImageJ) has been applied to the images for clarity of display.



(a) Ground truth **(b)** Ground truth, bin 10

(c) Ground truth, blur to PSF **(d)** Widefield (average of frames)

Figure. 2.9. A random distribution of molecules. Shown are the GT of the distribution (a), which has been binned (b), or blurred with a gaussian kernel with σ approximately equal to the diffraction limit (270 nm, c). The widefield is also presented as the sum of frames for the data when presented as a SMLM stack (d).

created with these methods often perform well when evaluated with the FRC metric or SQUIRREL, suggesting substantial resolution improvement. However, good metrics can also be produced even if the image is generated from simulated data too dense to localise accurately and the output exhibits clear artificial sharpening. The reason for this lies in the exact detail of what is being measured by FRC and SQUIRREL, and demonstrates the challenges in coming up with an SMLM image evaluation metric that does not require GT and yet is reliably robust to common artefacts.

A random distribution of fluorophores was imaged, occupying a 6.4×3.2 µm^2 area, including a 0.75 µm-thick pixel border, with an emitter density 400 µm^{-2} and a duty cycle of 1/10 frames resulting in an activation density of 40 µm^2, with no photobleaching. The inclusion of a border ensures that all of the algorithms are able to fit the structure, since some (like ME ThunderSTORM [40]) are unable to fit right up to the edges of a frame (Figures 2.9, 2.10).

The GT image (generated with a 10 nm super-resolution pixel size, Figure 2.9a) has a random fluorophore distribution – it should be noted that this is distinct from a *uniform* fluorophore distribution, which is structured. The random distribution has naturally occurring fluctuations in the density, which are visible if the GT image is binned at the level of one camera pixel (taken to be 100 nm) (Figure 2.9b). These fluctuations in intensity are also visible if the fluorophore distribution is rendered as

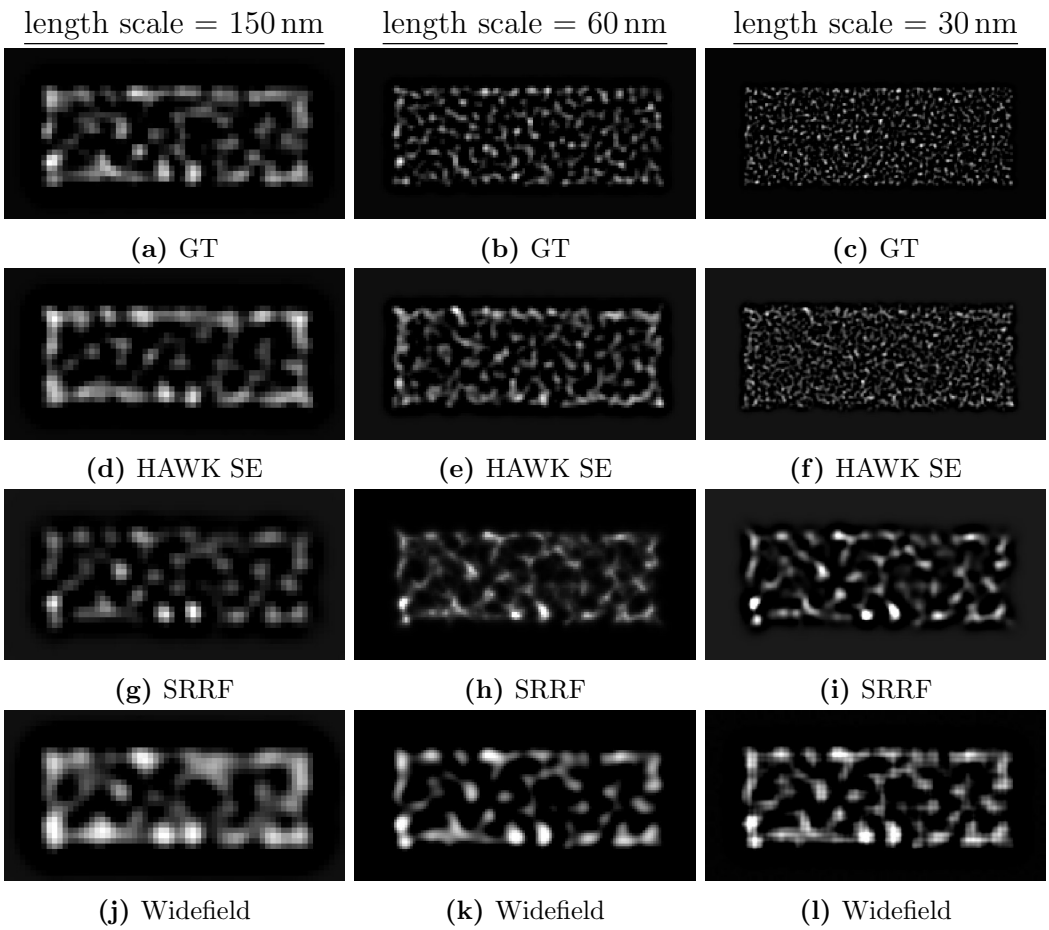


Figure. 2.10. SRRF cannot render high-frequency data in a reconstruction of randomly distributed data. Shown are GT (a-c), HAWK + SE (d-f), SRRF (g-i),and widefield (j-l) images of a random molecular distribution. The distribution is non-uniform and fluctuates in intensity as well as spatial distribution. The image was filtered using a top hat filter in the Fourier domain, carried out in ImageJ with filters 30-35, 12-15 and 6-8, corresponding to a mean Gaussian sigma of 150 nm (a, d, g, j), 60 nm (b, e, h, k) and 30 nm (c, f, i, l) respectively for a pixel size of 10 nm. This enables the structure present in images at particular length scales to be examined.

a widefield image with a Gaussian PSF at the expected size for a high performance fluorescence microscopy (270 nm) (Figure 2.9c). The same features appear if an SMLM image series is simulated with a high activation density and the images are averaged (Figure 2.9d).

If the HD SMLM dataset is analysed, the GT data exhibits different structure at the 150 nm length scale (Figure 2.10a), the 60 nm length scale (Figure 2.10b) and the 30 nm length scale (Figure 2.10c), as expected from a random distribution (Figure 2.10). Similarly, when analysed with HAWK+SE fitting (Figure 2.10d-f), the dense SMLM data produces a reconstructed image with different structure at each of the different length scales. However, since the resolution is degraded with HD data,

the output at the shortest length scale does not resemble the GT. However, it *does* remain unstructured. In contrast, a SRRF analysis of the same dataset (Figure 2.10g-i) produces structure which *always* (no matter the length scale) resembles the GT structure at the 150 nm length scale (Figure 2.10a). An insight into why this might be can be gained by considering the operating mode of SRRF. This algorithm uses a measure of radial symmetry of sub-pixels in the acquisition to provide a more accurate prediction of the emitter location, such that the sub-pixels with the greatest radial symmetry will be deemed most likely to designate the emitter's position [77]. However, this result invites us to consider the output SRRF can create when the activation density is too high for a radial transformation to produce an accurate position of the fluorophore. In this case, the most intense fluorophores will be measured as the 'centres' of high radial symmetry, leading to an output which is a non-linear transformation of the input, with an upsampling of the number of pixels. This is unsurprising and would normally be considered a fairly benign failure mode, as the algorithm is failing in such a way as to render an image closely related to the widefield.

However, it is important to note that SQUIRREL quantifies the quality of the image by comparing correspondence of the reconstructed image with an upsampled version of the widefield image. While this is effective at identifying non-linearities in super-resolution imaging, it means that the best-scoring image will be the one which most closely resembles the upsampled widefield image. When we upsample the simulated widefield image and consider the information present at different length scales (Figure 2.10j-l) we can see that it much more closely resembles the output SRRF image that it does the GT. This means that both FRC and SQUIRREL will report this type of artificial sharpening as being both high resolution, and highly accurate. Because of this, it was essential that HAWKMAN not be susceptible to the same failure mode, and which could warn users if the results they were seeing were the result of artificial sharpening.

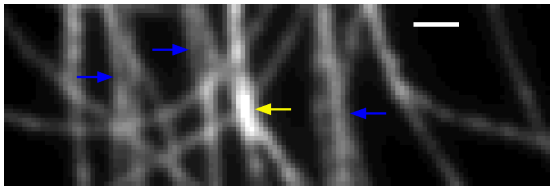
2.3.4 Performance comparison to SQUIRREL

The problem of using an intensity-based comparison, as SQUIRREL does (between a measured widefield and the sampling-dependent labelling density in the reconstruction), is clearly demonstrated by comparing it with HAWKMAN. Even at the PSF scale, HAWKMAN is able to correctly identify the relative fidelity of reconstructions differently affected by sharpening and other artefacts in situations where SQUIRREL cannot (Figure 2.11). Here, the assessment of experimental microtubule data from the Localisation Microscopy Challenge is performed using both HAWK-

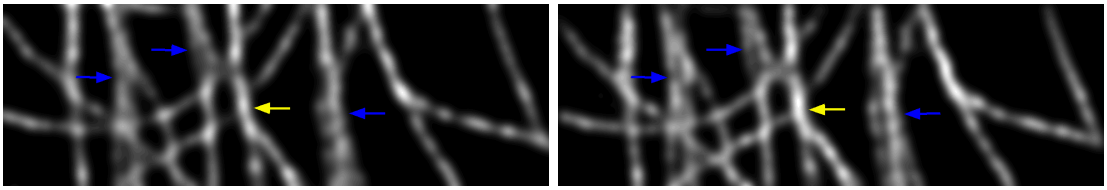
MAN and SQUIRREL for comparison. A widefield image produced by summing image frames (Figure 2.11a) shows areas where microtubules are poorly resolved, *even* when compared to the widefield (blue arrows) and a high intensity area (yellow arrow). The reconstruction produced by SE and ME fitting are shown in Figure 2.11b, where a Gaussian blur of a scale equivalent to the PSF has been applied (the scale at which SQUIRREL makes its comparison). Comparing these with the widefield image shows significant differences, particularly in the intensity (a proxy for labelling density). The error maps produced by SQUIRREL are displayed below these (Figure 2.11c). They show the largest errors are reported where there are large intensity differences between the reconstructions and the widefield reference. For easier comparison with the HAWKMAN results the relative error in the SQUIRREL maps was converted into a confidence score (see Methods) and used to colourise the blurred reconstructions (Figure 2.11d) in the same manner as for HAWKMAN. Comparison with the HAWKMAN results at the PSF length scale (Figure 2.11e) highlights how SQUIRREL has detected the intensity differences (yellow arrow), but largely ignored the differences in structure (blue arrows). The converse is true for HAWKMAN, which indicates low confidence in areas of structural dissimilarity, and high confidence in the areas that differ only in intensity. HAWKMAN also reports the ME result as a much more authentic reconstruction than that from SE fitting, something that can be confirmed by inspection of the areas marked by blue arrows. Conversely, SQUIRREL rates these similarly, as any errors arising from structural differences are swamped by the much larger intensity error common to both reconstructions. This undesirable behaviour arises from the assumption that the density of localisations identified by the localisation algorithm and the pixel intensity contained in the widefield image are both linearly related to the local labelling density. This is seldom the case however due to repeat appearances of some emitters and non-appearance of others. It is particularly untrue with high-density imaging, as localisation algorithms generally progressively miss individual emitters as the density increases. Binarising the images and using the same localisation algorithm and reconstruction method for both images circumvents this problem.

2.3.5 HAWKMAN can indicate the presence of unresolved structure

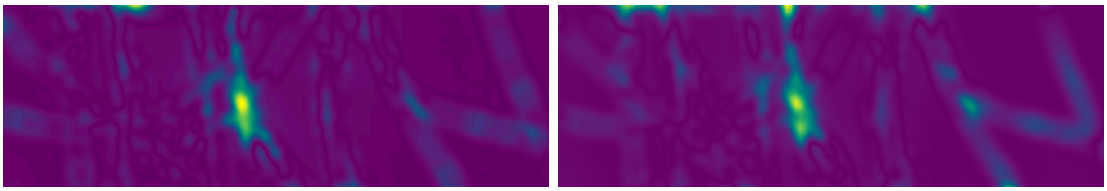
Using HAWK with many algorithms does introduce a small decrease in localisation precision compared to when that algorithm is used alone [78]. The magnitude of this decrease typically increases with the severity of bias present in the reconstruction without HAWK. For any given localisation algorithm, as the emitter density



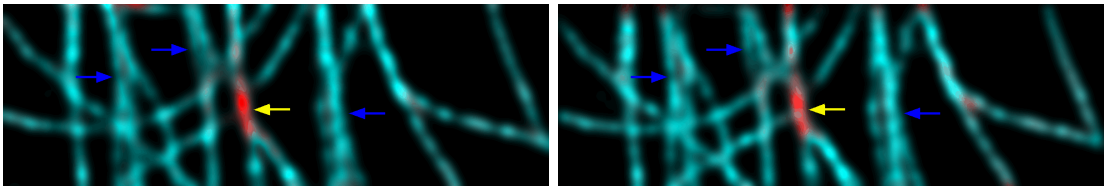
(a) widefield image (summed frames)



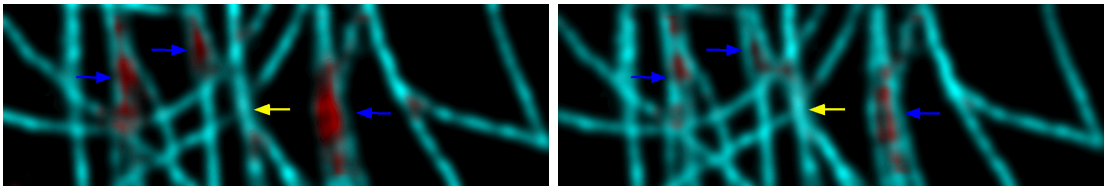
(b) reconstructions blurred to PSF scale



(c) SQUIRREL error maps



(d) confidence maps from SQUIRREL errors



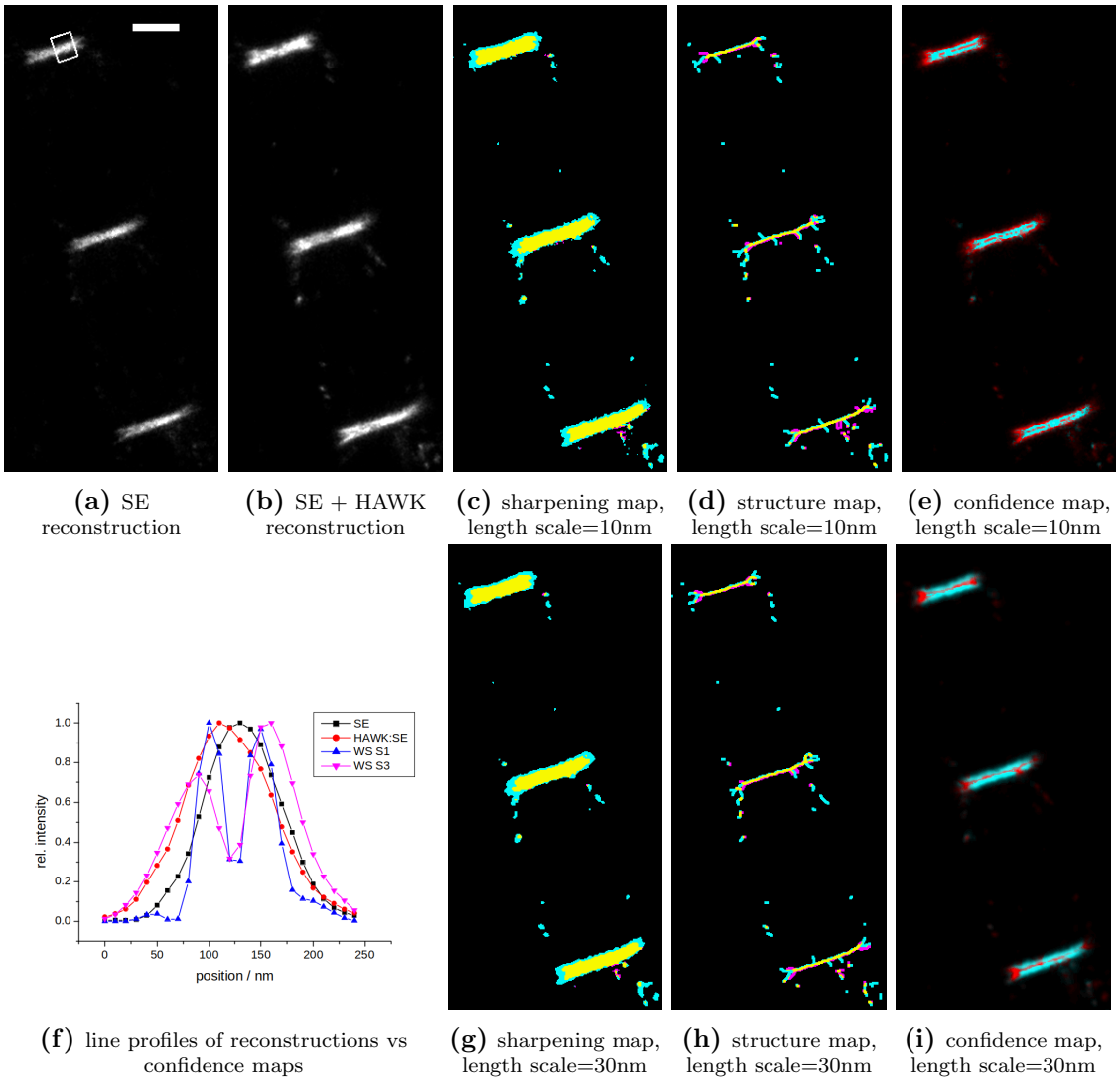
(e) HAWKMAN confidence map

Figure. 2.11. A comparison of SQUIRREL and HAWKMAN error map sensitivity. (a) A pseudo-widefield image produced by summing image frames. (b) SE and (c) ME reconstructions, blurred to the scale of the PSF. Blue arrows indicate areas of obvious sharpening compared to the widefield, the yellow arrow denotes an area of high intensity. (d, e) Error maps produced by SQUIRREL analysis show the largest errors (green/yellow) are dominated by the high intensity region in (a). The relative error is used instead of the confidence score to produce a HAWKMAN-style confidence map (f, g) from the SQUIRREL data. When compared to the actual HAWKMAN results (h, i), it shows SQUIRREL principally detects the intensity differences and does not recognise the superior structural details (blue arrows) of the ME reconstruction. Scale bar 1 µm.

increases, the precision of the HAWK reconstruction will be reduced, but remain unsharpened. Although this leads to a reduction in the resolution attained in the HAWK image, its fidelity when compared with the non-HAWK'd image will be increased. This means that using the HAWK reconstruction as the reference image slightly exaggerates differences with the input image when compared with an image of the GT (for instance, an LD reconstruction). This means that use of the HAWK'd image as a reference may lead to a slight overestimation of errors in some instances.

Additionally, large differences in precision between the HAWK'd and non-HAWK'd images may indicate the emitter density has exceeded the capabilities of the algorithm, even when used in combination with HAWK, resulting in reduced resolution. That is, the emitter density can be so high that the resulting distorted structures cannot even be resolved with HAWK-pre-processing due to the resolution loss incurred by HAWK. However, in such cases, the HAWKMAN confidence map may still indicate a low confidence in the offending region at certain length scales. This means that HAWKMAN can indicate where in a reconstruction, and at what length scale, the precision of the HAWK reference is sufficiently reduced that it may conceal unresolved finer structure. The experimenter is thus warned of the possibility of an unresolvable structure, and may decide that the experiment bares repeating, to attain a lower emitter density. This is useful, as for some emitter densities, the structure can be unresolvable but comparable in scale to the resolution in the HAWK image. Alternatively, a higher performance, more computationally intensive localisation algorithm may be required, such as switching from SE to ME fitting.

Thus, even though the structure is not visible in the HAWK'd or non-HAWK'd images, HAWKMAN may be able to give a first-order approximation to the underlying structure. This effect is demonstrated on and on the Z1Z2 domain in sarcomeres imaged at high density in Figure 2.12. Here the structure is not visible in either input image but is present in the HAWKMAN confidence map at a 50-70 nm length scale, which is in line with the expected observation for this structure. Figure 2.12 exhibits HAWKMAN's ability to reveal the possibility of unresolved structure in the HAWK'd reference image. This, combined with other possible effects (structural variation, antibody labelling, SNR and drift) mean the expected line separation is not resolved in either the test (Figure 2.12a) or HAWK'd (Figure 2.12b) reconstructions. HAWKMAN analysis (Figures 2.12c-2.12e) at the 10 nm scale suggests the possibility of missing structure. The sharpening map (Figure 2.12c) indicates some loss of precision in the reference image (cyan areas), indicating that the activation density is very high and concealed structure as a result of HAWK's precision



(a) SE reconstruction **(b)** SE + HAWK reconstruction **(c)** sharpening map, length scale=10nm **(d)** structure map, length scale=10nm **(e)** confidence map, length scale=10nm

(f) line profiles of reconstructions vs confidence maps **(g)** sharpening map, length scale=30nm **(h)** structure map, length scale=30nm **(i)** confidence map, length scale=30nm

Figure. 2.12. HAWKMAN analysis of experimental data of Z1Z2 antibody labelling of muscle sarcomeres demonstrates HAWKMAN's ability to reveal the possibility of structure unresolved in both inputs. Data was acquired at a sufficient activation density to produce significant artificial sharpening when using SE fitting. The input (a) and reference (b) were both reconstructed with SE fitting and processed with HAWKMAN. The output sharpening, structure, and confidence maps are displayed in (c-e) at length scale 10 nm and in (g-i) at length scale 30 nm. In (f) there is a graphed comparison of the line profiles across the boxed region in (a) for the test image, reference image, and the confidence map. The data presented above are from a single experiment selected to have sufficiently reduced resolution, such that the structure is not resolvable in the standard reconstruction and only partially resolvable in the HAWK reconstruction. The results are similar for all 3 sarcomeres in the image. Scale bar 500 nm.

loss may be possible. The confidence map (Figure 2.12e) suggests some sharpening artefacts in the test image and the possible presence of unresolved structure in the HAWK reference due to precision loss, as a thin area of low confidence (red) is located between the sarcomeric doublet. The unprocessed test (black) and HAWK'd reference (red) intensity profiles show only moderate differences, indicating sharpening artefacts in the test and possible loss of precision in the reference. However, the intensity-weighted correlation score shows a greatly reduced confidence in the central part of the structure for both the 10 nm (blue) and 30 nm (magenta) length scales. These two profiles both show a bimodal distribution similar to the simulated data above with a peak spacing of ∼50 nm (10 nm length scale) and ∼70 nm (30 nm length scale). Although these distances compare favourably with the expected separation of a Z1Z2 doublet (60-70 nm), they differ significantly from each other over a short range of length scales. This does *not* mean that this should be interpreted as a reliable measure of the underlying structure, but simply as an indication that underlying structure of some form may well exist and further experiment optimisation is warranted.

2.3.6 Common biological structures & two-colour data

To further assess the performance of HAWKMAN on experimental data we selected three well-known biological structures of differing geometry and scale to demonstrate the performance of HAWKMAN on experimental data: clathrin-coated pits, microtubules and mitochondria. This allowed us to test the performance of the system on extended linear structures (Figure 2.13a-d), small structures a few hundred nanometres in size (Figure 2.13e-h), and structures with an extended fluorophore distribution (Figure 2.13i-l). Differences between SE-fitted data processed without (Figure 2.13a, e, i and with HAWK (Figure 2.13b, f, j) are visually clear and are reflected by the HAWKMAN sharpening and confidence maps. These are displayed at representative length scales from the full HAWKMAN set for each dataset (Figure 2.13) that show the presence of artefacts at a scale both a few and several times the localisation precision for each structure. As mentioned previously, a modified value of the threshold coefficient may be required for optimum performance when analysing low-contrast structures; this was the case for the clathrin data (see section 2.2.1 for details). For this data the reconstruction intensity of the central region relative to their parameters range from ∼50–75%, requiring the threshold coefficient to be raised to 0.9 in order to separate these structures. This change is only required to identify the artefact as being the disappearance of the 'hole' in the pits but is not necessary for HAWKMAN to detect them as artefactual in the first place, as

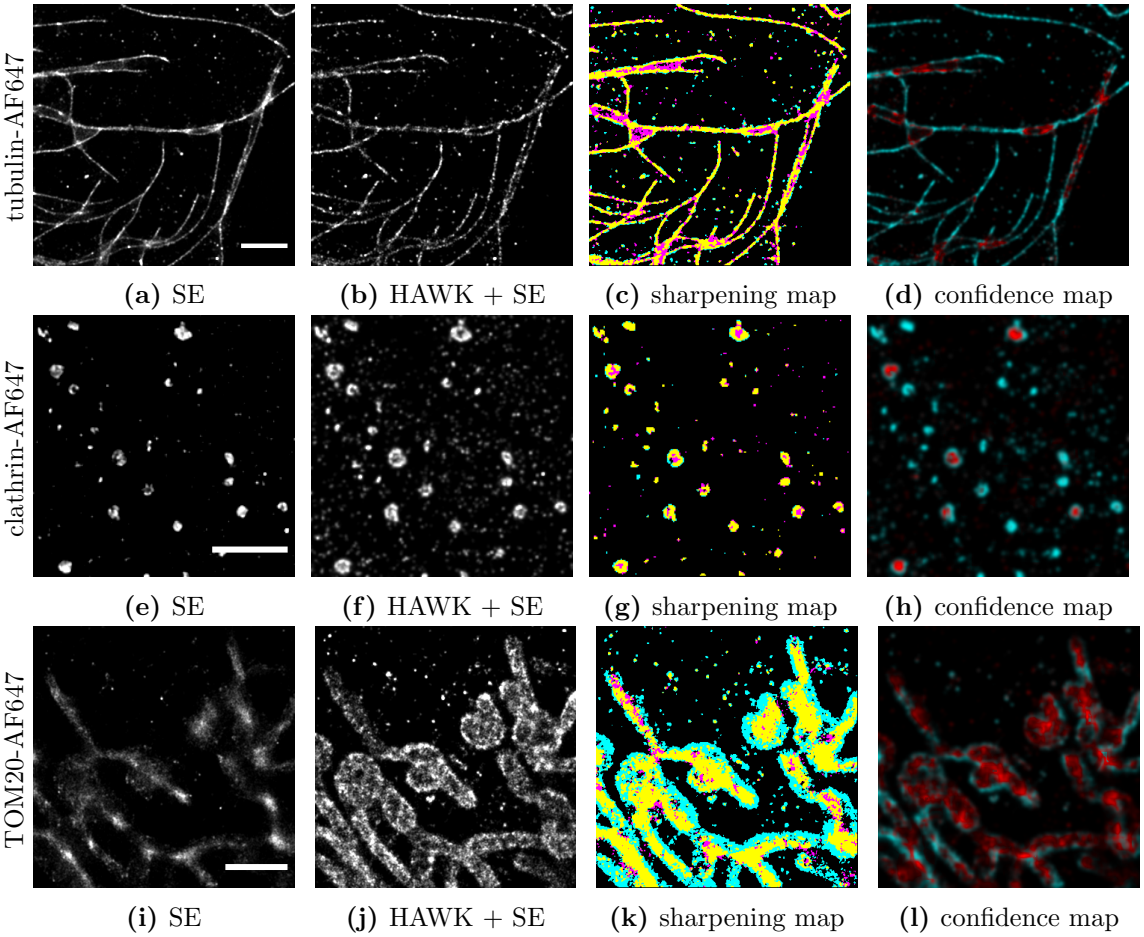
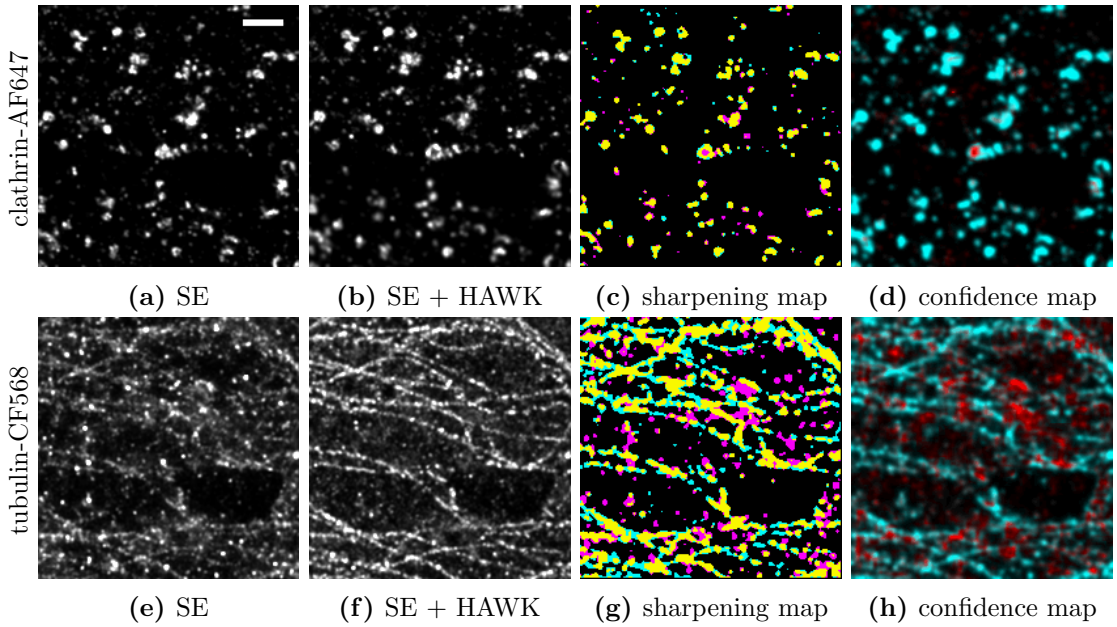


Figure. 2.13. HAWKMAN can assess reconstructions from a variety of different structures. Here, tubulin (a-d), clathrin (e-h) and the mitochondria epitope TOM20 (i-l) have been imaged at medium-to-high density and evaluated with HAWKMAN. The test images were reconstructed using SE ThunderSTORM (a, e, i), and analysed with a reference image similarly reconstructed from the HAWK pre-processed dataset (b, f, j), and analysed using the sharpening map (c, g, k) and the confidence map (d, h, l). Data were displayed at a length scale of 32 nm (c, g) and 24 nm (k) for the sharpening maps, and at 96 nm (d, h) and 120 nm (l) for the confidence maps. The sharpening map threshold coefficients were 0.9 (c) and 0.7 (g, k). The confidence maps (d, h, l) highlight the substantial degrees of sharpening, particularly the TOM20 reconstructions (i-l). Colouring as in Figure 2.2, scale bars 2 µm.



(a) SE **(b)** SE + HAWK **(c)** sharpening map **(d)** confidence map

(e) SE **(f)** SE + HAWK **(g)** sharpening map **(h)** confidence map

Figure. 2.14. HAWKMAN is sensitive to artefacts likely to be encountered in two colour imaging. Here, the fidelity of SE reconstructions for clathrin-coated pits stained with Alexa Fluor® 647 (a, b; lower emitter density) are evaluated using HAWKMAN (c, d). A similar investigation was carried out on tubulin stained with CF568 (e-h; higher emitter density). HAWKMAN sharpening maps (c, g) were generated using a structure-appropriate threshold coefficient of 0.9 for clathrin and the default value of 0.7 for tubulin. Confidence maps (d, h) were produced using a correlation measure of the blurred sharpening maps and the blurred structure maps (not displayed), the latter of which were used a threshold coefficient of 0.95 in the clathrin-AF647 case (a-d) and a default value of 0.85 in the tubulin-CF568 case (e-h). Due to high background, the global offset coefficient was also raised from its default value of 0.04 to 0.1 in the tubulin-CF568 case (e-h). All maps are rendered for analysis at a 96 nm length scale (3 pixels at 5x magnified reconstruction pixel size of data taken on a 160 nm camera). Scale bar is 1 µm.

the sharpening map will still indicate the different sizes of the sharpened structures in the test and reference images.

Experimentally, localisation artefacts such as artificial sharpening are particularly likely to be experienced in multicolour dSTORM measurements due to the variable performance of dyes. While Alexa Fluor® 647, with its high-intensity, LD emissions is exceptionally well-suited to dSTORM acquisitions, most other dyes are far less so. Researchers looking to perform multicolour dSTORM analysis must therefore contend with the fact that dyes whose emission spectra peak in other laser channels blink with much lower brightness and at much higher density. As when imaging living cells, this makes multicolour stains highly prone to the generation of sharpened/biased data. HAWKMAN can identify the regions in which artificial sharpening is occurring in typical two-colour dSTORM data (Figure 2.14). Very little sharpening is present in the LD AF647-clathrin case (Figure 2.14a-d) as evi-

denced by the prevalence of cyan colouring in Figure 2.14d and yellow dominating all non-background signal in Figure 2.14c, except for a single clathrin-coated pit whose 'pit' is unresolvable in the input image (Figure 2.14a) that is detected by both HAWKMAN maps (Figure 2.14c, d). This demonstrates the value of HAWKMAN even in cases where a low density of emissions would often be assumed to result in little sharpening. In the HD case (Figure 2.14e-h), the sharpening map (Figure 2.14g) shows the source of the severely punctate nature of the test image (Figure 2.14e) to be a conflation of emissions towards their mutual centre both between (Figure 2.14f, magenta) and along (Figure 2.14f, cyan) the microtubules, with magenta colouring suggesting an introduction of false structure between proximal fibres, and cyan suggesting an absence of data due to emitter mis-localisation. Two-colour experiments are an example of where the blinking statistics may be quite far from optimal, resulting in artefacts. Fluorescent proteins and DNA paint also have blinking properties that can be quite different to ideal STORM data.

2.4 Discussion

The basis of HAWKMAN's proficiency is that it exploits the fact that algorithms fail in different ways when used with and without HAWK. At extreme activation densities, most algorithms become significantly biased, whereas with HAWK there is no bias, but a decrease in localisation precision. Therefore, in a scenario where both images were in complete agreement, one could be confident that the activation density was appropriate for that algorithm, and that the image contains the most accurate and precise localisations it can produce (i.e. comparable to an LD acquisition). That is to say, in this case the HAWK (and test) reconstruction will not only be unbiased but also have a minimal precision loss.

It is worth noting that, aside from error maps, HAWKMAN outputs graphs measuring the global correlation between the test and reference images, yet these are often omitted from the figures in this section. It could be argued that part of the philosophy behind HAWKMAN is that global image quality metrics are generally not a practical way of measuring image quality in SMLM, since it is commonly the case that some parts of a reconstruction contain structures that have not been severely warped by artefacts while others are completely unusable. Additionally, of course, the correlation between the input images follows a not-dissimilar pattern regardless of the imaged structures.

HAWK's ability to remove mis-localisation artefacts arising from emitter overlap from SMLM datasets with even extreme activation densities begs the question: 'if

HAWK'd images contain no artefacts, then why use HAWKMAN at all?'. This can be answered by looking at data in which the activation density of the dataset has exceeded the capacity of HAWK. HAWK'd images, though unbiased, suffer from precision loss due to temporal filtering, which grows more severe with increased activation density. This means that simply using HAWK – though it *can* ensure that the image will not contain false structure – cannot guarantee that the image is informing of all potentially resolvable structures in the image. Because algorithms fail differently when HAWK is used, disagreement between HAWK'd and unHAWK'd images can reveal the presence of further structure. This is a tremendously useful feature for experiment optimisation.

Image artefacts are a critical issue in localisation microscopy. They can be very difficult to detect and quantify, even for experienced users, as they can involve subtle structure changes. In addition, they may only be present in certain regions of the image, with many parts of the image appearing to be of good quality. Therefore, providing tools which can detect artefacts below the resolution limit is critical if SMLM is to provide reliable and trustworthy results. The results here demonstrate that HAWKMAN can locally map and quantify the severity artefacts generated from HD data even by the some of the most reputable and sophisticated HD SMLM reconstruction algorithms. This enables, for the first time, a measurement of sub-diffraction limit *functional resolution* in SMLM images. They also show that HAWKMAN can indicate the possible presence of unresolved structure in the reconstruction. Taken together, these capabilities make for an image analysis step that allows researchers to know if an SMLM experiment should be performed again. HAWKMAN is a straightforward, fast (less than 1 min on a desktop PC for the challenge data presented in Figure 2.4) and reliable test, available as an ImageJ plugin [79], which users can apply to verify the quality of their data.

Chapter 3

HAWK in 3D

3.1 Introduction

The acquisition of 3D data is an increasingly essential tool in bioimaging. Certainly, 2D imaging has a long and successful history in the field, and is responsible for many important results. However, biological processes are themselves three-dimensional, meaning that the factual accuracy of results presented by 2D data can be limited. This is firstly an issue in experimental design – for instance, experiments whose aim is to determine the behaviour of cellular processes are frequently performed on cells on coverslips. In these scenarios, the cells is 'flattened' and extended, reducing the amount of information lost in a 2D imaging experiment. However, though such experiments can produce useful results, they often do not accurately capture the physiological environment in which these cells are operating: cells *in vivo* do not spend their time navigating flat, hard substrates, meaning that by fixing cells to coverslips, experimenters are introducing a 'non-nativeness' to imaged cellular structures. In fact, the breadth of functions of the extra-cellular matrix means that cell behaviours associated with cellular migration are complex and multi-valent, resulting from cellular sensitivity to minute inhomogeneities in the medium around it; these inhomogeneities are of course, not present for cells whose behaviours are so impeded by being spread out over coverslips [120–123]. Additionally for cells seeded into 3D matrices, 2D imaging can produce ambiguous data, particularly when imaging living cells. For instance, it might be unclear whether elongated cellular components are growing/ shrinking, or entering/ leaving the 2D imaging plane. This is partily because the cell is so much more extended in z, meaning that less of it can be captured without resulting in heavy background.

To create 3D images using SMLM, researchers must break the 'symmetry ambiguity' (wherein molecules equally separated positively or negatively from the focal

plane are indistinguishable) by engineering the PSF or modifying the way it is imaged. Once the PSFs have been individuated and 'engineered', their geometries can be used to localise them axially. However, acquiring and processing high-quality 3D fluorecence microscopy data is generally significantly more difficult than doing so for 2D data, and in this regard, SMLM is no exception. There are several reasons for this:

1. Engineering PSFs to break the symmetry ambiguity tends to make them larger, as the axial location becomes encoded in the shape of the PSF. In fact, 3D PSFs that span a larger range of configurations/ sizes can make the function that maps PSF shape to z-position less 'coarse', increasing the attainable z-precision. However, this will reduce the xy-localisation precision by distributing the signal over a wider area. It also increases the probability of emitter overlap and thus the prevalence of artefacts.

2. In 2D SMLM, specialised sectioning illumination techniques like TIRF enable thin volumes of the sample to be illuminated. This ensures a higher SNR by reducing background fluorecence and out-of-focus signal, which contribute to reduced localisation precision and an increase in artefacts. However, in 3D SMLM, illuminating a larger volume becomes important, as it increases the amount of usable data – there isn't much to be gained from acquiring 3D data over a 100 nm-thick volume. Of course, imaging larger volumes reduces SNR (lowering the localisation precision) and increases emission density (creating more emitter overlap and hence more artefacts) Thus, gathering 3D data nullifies one of our primary methods for ensuring a high quality of data in SMLM.

3. Emitter overlap can radically alter the shape of the PSF as perceived by the localisation algorithm. In other words, 3D SMLM datasets can produce a much greater variety of artefacts, which can be even harder to diagnose and avoid than those typically present in 2D SMLM reconstructions.

4. In most cases, HD reconstruction software usually cannot localise molecules that have been imaged with 3D SMLM modalities, as few of them are easily compatible with an analysis of PSF shape. This can make all of the above issues nigh-impossible to overcome.

It is worth noting that a common denominator each of these issues is *artefacts*. 3D SMLM leads to more artefacts, and they are more varied, more severe, and harder to remove. The increased incidence of artefacts is difficult to tackle because these processes often do not generally preserve the legibility of the dataset to most HD

Figure. 3.1. An illustration of the symmetry ambiguity. Despite the differing axial positioning in the three molecules, the leftmost molecule (which is located above the focal plane) and the rightmost molecule (which is located below it) are indistinguishable.

SMLM reconstruction software. This difficulty is significantly exacerbated when we consider the *live-cell* 3D scenario, which necessitates yet another dramatic decrease in the requisite dataset quality to shorten acquisition times.

There have been some attempts to overcome this problem, with varying success. Some improvements in 3D localisation have been attained on HD data by coupling DAOSTORM [74] with astigmatic imaging [83] (section 3.3.2) to perform HD 3D SMLM on fixed, 100 nm-sectioned samples [124]. Additionally, HD 3D localisation has been attained by combining compressed sensing with biplane imaging [125] (biplane imaging in section 3.3.5) and the double-helix PSF [126] (double-helix PSF in section 3.3.4). The problem with these solutions is twofold. Firstly, the base HD localisation algorithms – DAOSTORM and compressed sensing – whilst being able to handle a higher density of fluorophores than standard SE fitting, still fail by producing artefacts at a sufficiently high emitter density [78]. Secondly, these algorithms were only shown to work on thin sections of fixed samples with organic dyes, which typically have a much higher photon count than fluorescent proteins which are normally needed in the live-cell case. A combination of astigmatic and biplane imaging with the Fast Localization algorithm based on a Continuous-space formulation (FALCON) tackles the former issue [127], but FALCON's failure mode is still ultimately expected to produce artefacts at sufficiently high densities.

However, HAWK offers a solution to these issues. It redistributes the PSFs temporally without dramatically changing their shape, or demanding that they be localised in a different way. HAWK has already been shown to eliminate all but the most severe (and thus, easily spotted) artefacts from 2D data [78]. It is also unique in its failure mode in that extreme emitter densities reduce the resolution of the final reconstruction instead of producing artefacts. This means it certainly bears investigating whether the same improvements can be achieved by applying HAWK filtering to 3D datasets, and whether the approach needs to be adapted for specific 3D methods. Here, we demonstrate that HAWK pre-processing enables HD 3D SMLM to a high precision without artefacts. HAWK separates the signal contributions to each pixel into artificial frames using a temporal filter of each pixel's intensity/time trace. This process splits overlapping emitters into separate frames, removing artefacts from the reconstruction (section 1.5.2). Importantly, this process is expected to preserve PSF geometry – we therefore expect HAWK analysis to be compatible with 3D SMLM variants, which typically use PSF geometry to measure axial location. We exhibit HAWK's ability to remove artificial sharpening from 3D datasets obtained using the most widely used methods: the astigmatic [83] and double-helix PSFs [84], and biplane microscopy [85].

It should be noted that there are several other 3D SMLM methods that are fairly widely used, but are not as popular as the three we investigated. This is likely due to their relative novelty, or difficulty of implementation, or both. Direct Optical Nanoscopy with Axially Localized Detection (DONALD) and Supercritical Illumination Microscopy Photometric z-Localization with Enhanced Resolution (SIMPLER) image the near-field emissions, which are effervescent and thus decrease quickly with distance from the emitting fluorophore. This means their imaged intensity can be used to determine the distance from the coverslip, giving a measure of the z-position [128, 129].

Thus, the goals of this research into the use of HAWK on 3D datasets are to:

- ensure that HAWK can successfully remove artefacts arising from emitter overlap in 3D SMLM datasets, wherein individual PSFs have been engineered to have a diversity of shapes and sizes
- diagnose the unique kinds of artefacts that are present in different kinds of 3D SMLM data and ensure that HAWK can successfully remove them
- ensure that HAWK temporal filtering does not sufficiently modify the PSF profiles that novel artefacts are introduced

3.2 Materials & Methods

3.2.1 Astigmatic sample preparation & imaging protocol: nuclear pore complex

Nuclear pore complex data analysed here was published in [130] with sample preparation details below.

Wild-type U-2 OS and genome-edited U-2 OS cells that expressed Nup107–SNAP37 were cultured under adherent conditions in DMEM (high-glucose, without phenol red) supplemented with 10% (v/v) FBS, 2 mM L-glutamine, non-essential amino acids, and ZellShield at 37 °C, 5% CO2, 100% humidity. All incubations were carried out at room temperature. For nuclear pore staining, the coverslips were rinsed twice with PBS and prefixed with 2.4% (w/v) FA in PBS for 30 s. Cells were permeabilized with 0.4% (v/v) Triton X-100 in PBS for 3 min and then fixed with 2.4% (w/v) FA in PBS for 30 min. Subsequently, the fixation reaction was quenched by incubation in 100 mM NH4Cl in PBS for 5 min. After being washed twice with PBS, the samples were blocked with Image-iT FX signal enhancer (Thermo Fisher Scientific, Waltham, MA, USA) for 30 min. The coverslips were incubated in staining solution (1 µM benzylguanine Alexa Fluor 647 (S9136S; NEB, Ipswich, MA, USA), 1 mM DTT, 1% (w/v) BSA in PBS) for 50 min in the dark. After being rinsed three times with PBS and washed three times with PBS for 5 min, the sample was mounted for imaging.

SMLM images were acquired on a commercial Leica SR GSD 3D microscope. The setup is equipped with a 500 mW 642 nm laser (MPBC Inc., Montreal, Quebec, Canada), a 30 mW 405 nm diode laser (Coherent Inc., Santa Clara, CA, USA), a DBP 405/10 642/10 excitation filter, an LP649 dichroic mirror, a BP 710/100 emission filter, and a high-NA oil-immersion objective (160×/1.43-NA; Leica). Astigmatism was induced with a modified tube lens, and the image was recorded on an iXon3 897 EMCCD camera (Andor, Belfast, Northern Ireland).

3.2.2 Astigmatic sample preparation & imaging protocol: tubulin-AF568

Human cervical cancer cells (HeLa) were thawed and cultured in Dulbecco's Modified Eagle Media (DMEM) 1% Penicillin and Streptomycin (PS, Sigma-P0781), 10% foetal bovine serum (FBS, HyClone-SV30160.03) and 1% L-Glutamine (L-Glu, Sigma-59202C). Cells were transferred to a T25 flask (Cellstar-690175) and incubated at 37°C and 5% Carbon dioxide. Passaging cells were trypsinized with 10%

trypsin (Sigma-T4174) diluted in phosphate-buffered saline (PBS, Sigma-D8537), for 5 min once cells reached 80% confluence. Complete DMEM was added to neutralise trypsin and centrifuged at 960xg for 3 min. The supernatant was aspirated, and cells resuspended in 5 mL of complete DMEM. Cells were plated on fibronectin (Sigma-FC010) coated 35 mm high glass bottom dishes (ibidi-81158) for imaging. cells were fixed for 15 min in 3.6% formaldehyde (PeqLab-30201) at room temperature (RT) and washed three times with 1×PBS.

Permeabilisation and blocking were undertaken by incubating cells in 'blocking buffer' consisting of 3% BSA (Sigma-10735108001) and 0.5% Triton X-100 (Sigma-X100) in 1×PBS for 10 min. Cells were then incubated with primary antibodies (tubulin from Sigma-T8328 at 1:200,) diluted in blocking buffer for 30 min while gently rocked at RT. After three 5 min washes in washing buffer consisting in 0.2% BSA and 0.1% Triton X-100 in 1×PBS, cells were incubated with the Alexa Fluor® 568 secondary antibodies (ThermoFisher Scientific-A10037, at 1:500 dilution) in blocking buffer for 30 min, gently rocked at RT. Cells underwent three 5 min washes in 1×PBS before being stored at 4 °C in 1×PBS for up to 2 days before imaging.

Cells were imaged in imaging buffer, where 1220 µL buffer was made using 800 µL distilled water, 200 µL 'MEA', 20 µL 'GLOX' and 200 µL dilution buffer. Here, MEA is 1 M cysteamine (Sigma-30070) and 0.25 mM HCl (Sigma-H9892); GLOX is 35 mM glucose oxidase (Sigma-G6125), 13.6 µM Catalase (Sigma-C40), 8 mM Tris (Amresco-E199) and 40 mM NaCl; lastly, 'dilution buffer' is 50 mM Tris, 10 mM NaCl (Alfa Aesar-A12313) and 10% w/v glucose (Thermo Fisher-G/0450). Dilution buffer was added just before imaging to minimise cell damage due to pH change.

Objective-based total internal reflection fluorescence (TIRF) was employed to minimise background. Fluorophore bleaching was undertaken with widefield illumination prior to image acquisition to minimise background signal from above the structures of interest. Experimental data of tubulin was gathered from fixed HeLa cells on a Nikon motorised inverted microscope ECLIPSE Ti2-E with Perfect Focus System in the King's College Nikon Imaging Centre. The data was acquired using a cylindrical lens to obtain astigmatic 3D datasets that had been pre-calibrated on a sample of fluorescent beads. It is equipped with a laser bank with 405, 488, 561, and 640 nm lasers (LU-NV series), a 160 nm-pixel ORCA-Flash 4.0 sCMOS (scientific Complementary metal–oxide–semiconductor, Hamamatsu Photonics K.K.) and a CFI SR HP Apochromat TIRF 100XAC oil objective (NA 1.49) with an automatic correction collar. The microscope was controlled using Nikon Elements v5.2. software.

3.2.3 Simulations of resolving 3D structure

Simulations were carried out in Matlab. For all simulations pixels sizes were 100 nm, with a PSF of FWHM 270 nm (σ of 117 nm). The mean brightness of each emitter was 2500 photons per frame, and background was 25 photons per pixel. The different densities simulated were designated by HD1, HD2, HD3, HD4, and HD5, which correspond to notional activation densities in the ratio 1:3:10:30:100, with the starting density being defined in each case such that fluorophores are very well-separated and results would exhibit no artefacts if that density were used to image a 2D structure. However, it should be noted that due to the vertical structure present in some simulations and the larger size of the PSF used for 3D imaging, artefacts could occur even at HD1. In these cases to achieve artefact-free images a lower density, HD0, was defined with 0.3 times the density of HD1. Before pixels were rendered, Poisson noise was added, as is expected from the imaging process.

3.2.4 Astigmatism simulations

The impact of astigmatism was simulated using a polynomial model centred around the focal point as used in [131]. The two minima were separated by around 600 nm, a figure chosen as it is approximately the separation observed in experimental data taken on the N-STORM system. The range of the fit was typically restricted to 3 µm to obtain smooth fitting over the central region, and to avoid distortion by PSFs that are no longer approximately Gaussian.

3.2.5 Statement on the choice of ThunderSTORM as fitting algorithm

ThunderSTORM [40] is an open-source plugin for ImageJ that is used throughout this research for localising molecules from SMLM image stacks. Different localisation software use different visualisation algorithms, offer different localisation algorithms, enable different parameters to be modified, and some different filters to be applied to output molecule location tables. Free and open-source software allows for unparalleled accessibility and flexibility, which is exeptionally useful when data from different research groups has been used, as it has in this chapter. Additionally, and perhaps most importantly, ThunderSTORM is capable of localising molecules using both SE and ME fitting, in conjunction with 2D PSFs, astigmatic PSFs, and biplane PSFs. This allowed the various methods investigated to be compared on as equal footing as possible, allowing us to hold some variables relating to the localisation

method and parameters constant (with the exception of DHPs, which required the use of *Easy-DHPSF* [132]).

3.2.6 Astigmatic data analysis

All acquisitions underwent ThunderSTORM analysis as performed using the SE and ME methods. We used a Difference of Gaussians filter to identify initial areas of interest, generated using Gaussians of $\sigma_{min} = 1.0$ and $\sigma_{max} = 1.6$. Areas of contiguous components were detected with a connectivity of 8. The PSF was modelled using an integrated gaussian, using a maximum likelihood estimator to detect the PDF centroids, with a fitting radius of 3 pixels for SE and 5 pixels for ME. The maximum number of molecules to be detected within an area of contiguous signal during ME fitting was 5, with a p-threshold of 0.05. ME fitting, though it offers increased fidelity by accounting for an increased number of potential molecules in each connected component, significantly slows down the reconstruction speed due to the increased ambiguity associated with fitting. The impact of astigmatism was simulated using a polynomial model as described above

3.2.7 DHP sample preparation & imaging protocol: lamin

Data was kindly provided by Leonard Möckl, then of the Moerner group, Stanford University, now at the Max Plank Institute for the Science of Light. Images were of HeLa cell plated onto plasma-etched coverslips (Fisher Premium Cover Glass, no. 1.5) and subsequently cultured for 24h in high-glucose, phenol-red-free DMEM (HyClone) supplemented with 10% FBS.

For lamin B1 labeling, HeLa cells (CCL-2, ATCC) were washed three times in PBS, fixed in chilled 4% paraformaldehyde (PFA) (Electron Microscopy Sciences) in PBS for 20 min, washed once in PBS, and incubated with 10 mM NH4Cl (Sigma-Aldrich) in PBS for 10 min. Next, the cells were permeabilized with three washing steps with 0.2% (v/v) Triton X-100 (Sigma-Aldrich) in PBS with 5 min incubation between each wash, and blocked with 3% (w/v) BSA (Sigma-Aldrich) in PBS for 1h. The cells were then labeled with primary rabbit anti-lamin B1 (ab16048, Abcam) using a 1:1000 dilution in 1% (w/v) BSA in PBS for 2h, washed three times with 0.1% (v/v) Triton X-100 in PBS with 3 min incubation between each wash, and labeled with secondary donkey anti-rabbit conjugated with Alexa Fluor®647 (ab150067, Abcam) using a 1:1000 dilution in 1% (w/v) BSA in PBS for 1h. The cells were finally washed five times with 0.1% (v/v) Triton X-100 in PBS with 1 min incubation between each wash.

For single-molecule SR imaging, the PBS was replaced by a reducing and oxygen-scavenging buffer comprising 100 mM Tris-HCl (Invitrogen), 10% (w/v) glucose (BD DiFco), 2 µL mL^{-1} catalase (Sigma-Aldrich), 560 µg mL^{-1} glucose oxidase (Sigma-Aldrich), and cysteamine (Sigma-Aldrich) with a concentration of 10 mM (Alexa Fluor®647 immunolabeled lamin B1 (2D)), 20 mM (Alexa Fluor®647 immunolabeled mitochondria), or 40 mM (Alexa Fluor®647 immunolabeled lamin B1 (3D)). At the beginning of the measurement, a large fraction of the Alexa Fluor®647 molecules were converted into a dark state using 647 nm epi-illumination at 5 kW cm^{-2}. For all single-molecule measurements, an exposure time of 50 ms and a calibrated EM gain of 186 was used. For each imaging experiment, at least 300 dark frames were acquired when the shutter of the camera was closed. The mean of these dark frames was subtracted from the images before analysis. When using 3D PSFs, calibration of each PSF was carried out by axial scanning of the fiducial beads in the sample over the full PSF range using the piezoelectric objective scanner. A transmissive dielectric phase mask was used to create the double helix PSF [133].

3.2.8 DHP simulations

For the DHP the separation between the spots was set to be three times the width of the PSF (810 nm) and each of the two spots which comprised the PSF were enlarged by a factor of 1.7 (469 nm), values which were selected to match the PSFs in our experimental data. The two lobes of the DHP rotated around 180° for a 1200 nm change in z. These parameters are fairly typical for DHP experiments as slower rotation, which enables larger depths to be imaged, conversely leads to lower z-resolution. As in most applications of 3D SMLM the aim is to achieve somewhere close to isotropic resolution, the PSF almost always rotates through a 600-2000 nm range [134].

For both the biplane and the double helix simulations calibration, data of a few simulated beads over the whole z-range were generated. To aid the calibration and minimise the data generation required, Poisson noise was not generated. Calibration was done on this data in ThunderSTORM before fitting the generated data was fitted. For experimental data new calibrations were taken from the corresponding experimental calibration data.

3.2.9 DHP data analysis

Double-Helix data was analysed using the *Easy-DHPSF* reconstruction software [132]. Analysis required a z-calibration scan, which was either simulated (for simulated data analysis) or experimental, in both simulated and experimental calibration data images were of beads (as described above). The dark offset, the expected camera value with the laser off, was set to zero for simulations and was obtained from a calibration image for the experimental data.

3.2.10 Biplane simulations

For biplane simulations two output images were created, with the separation between focal planes set to 600 nm. The defocus was calculated using a polynomial equation model used in ThunderSTORM [131].

3.2.11 Biplane sample preparation & imaging protocol: mitochondria

Mitochondria were prepared and stained by Fang Huang at the Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA.

COS-7 cells (CRL-1651, ATCC) cultured on 25 mm diameter coverslips (CSHP-No1.5-25, Bioscience Tools) were first fixed with 37 °C prewarmed 3% PFA (15710, Electron Microscopy Sciences) and 0.5% GA (16019, Electron Microscopy Sciences) in 1×PBS (10010023, Gibco) at RT for 15 min. Subsequently, cells were washed twice with 1×PBS and then quenched with freshly prepared 0.1% NaBH4 (452882, Sigma-Aldrich) in 1×PBS for 7 min. After washing three times with 1×PBS, cells were treated with 3% BSA (001-000-162, Jackson ImmunoResearch) and 0.2% Triton X-100 (X100, Sigma-Aldrich) in 1×PBS for 1h, gently rocked at RT. Then cells were incubated with primary antibodies (sc-11415, Santa Cruz Biotechnology) at 4 °C overnight. After washing three times for 5 min each time with wash buffer (0.05% Triton X-100 in 1×PBS), cells were incubated with secondary antibodies (A21245, Molecular Probes) at RT for 5h. Both primary and secondary antibodies were diluted to 1:500 in 1% BSA and 0.2% Triton X-100 in 1×PBS. After washing three times (5 min each time with wash buffer), cells were postfixed with 4% PFA in 1×PBS for 10 min. Then cells were washed three times with 1×PBS and stored in 1×PBS at 4 until imaging.

Immediately before imaging, the coverslip with cells on top of it was placed on a custom-made holder. About 40 mL of imaging buffer (10% (w/v) glucose in 50 mM Tris (JT4109, Avantor), 50 mM NaCl (S271-500, Fisher Chemical), 10 mM

MEA (M6500, Sigma-Aldrich), 50 mM BME (M3148, Sigma-Aldrich), 2 mM COT (138924, Sigma-Aldrich), 2.5 mM PCA (37580, Sigma-Aldrich) and 50 nM PCD (P8279, Sigma-Aldrich), pH 8.0) was added on top of the coverslip. Then another coverslip was placed on top of the imaging buffer. This coverslip sandwich was then sealed with two-component silicon dental glue (picodent twinsil speed 22, Dental-Produktions und Vertriebs GmbH).

SMLM imaging was performed on a custom-built setup on an Olympus IX-73 microscope stand (IX-73, Olympus America) equipped with a 100x/1.35-NA silicone-oil-immersion objective lens (FV-U2B714, Olympus America) and a PI-FOC objective positioner (ND72Z2LAQ, Physik Instrumente). Samples were excited by a 642 nm laser (2RU-VFL-P-2000-642-B1R, MPB Communications), which passed through an acousto-optic tuneable filter (AOTFnC-400.650-TN, AA Opto-electronic) for power modulation. The excitation light was focused on the pupil plane of the objective lens after passing through a filter cube holding a quadband dichroic mirror (Di03-R405/488/561/635-t1, Semrock). The fluorescent signal was magnified by relay lenses arranged in a 4f alignment to a final magnification of ∼54 and then split with a 50/50 non-polarising beam splitter (BS016, Thorlabs). The split fluorescent signals were delivered by two mirrors onto a 90° specialty mirror (47-005, Edmund Optics), axially separated by 580 nm in the sample plane, and then projected on an sCMOS camera (Orca-Flash4.0v3, Hamamatsu) with an effective pixel size of 120 nm. A bandpass filter (FF01-731/137-25, Semrock) was placed just before the camera.

3.3 Results

3.3.1 Note on assessment of 3D images

Assessing the success or failure of 3D SMLM reconstruction algorithms is highly challenging. As we saw in chapter 2, developing a viable quantitative method of assessment even for 2D SMLM images that can assess image quality comprehensively (measuring image accuracy and resolution locally) proved a complex undertaking. 3D SMLM images are significantly more complicated to assess than 2D ones, because of the volume of data and the increased variety and severity of artefacts. With this in mind, assessment of the 3D SMLM data presented in this chapter was primarily conducted on a basis of visual assessment (with the exception of some DHP data, in which the number of rendered localisations could serve as a proxy measure for the quality of the [pre-processed] dataset). We thus endeavoured to render our data in such a way that made changes in the quantity and severity of artefacts, and any

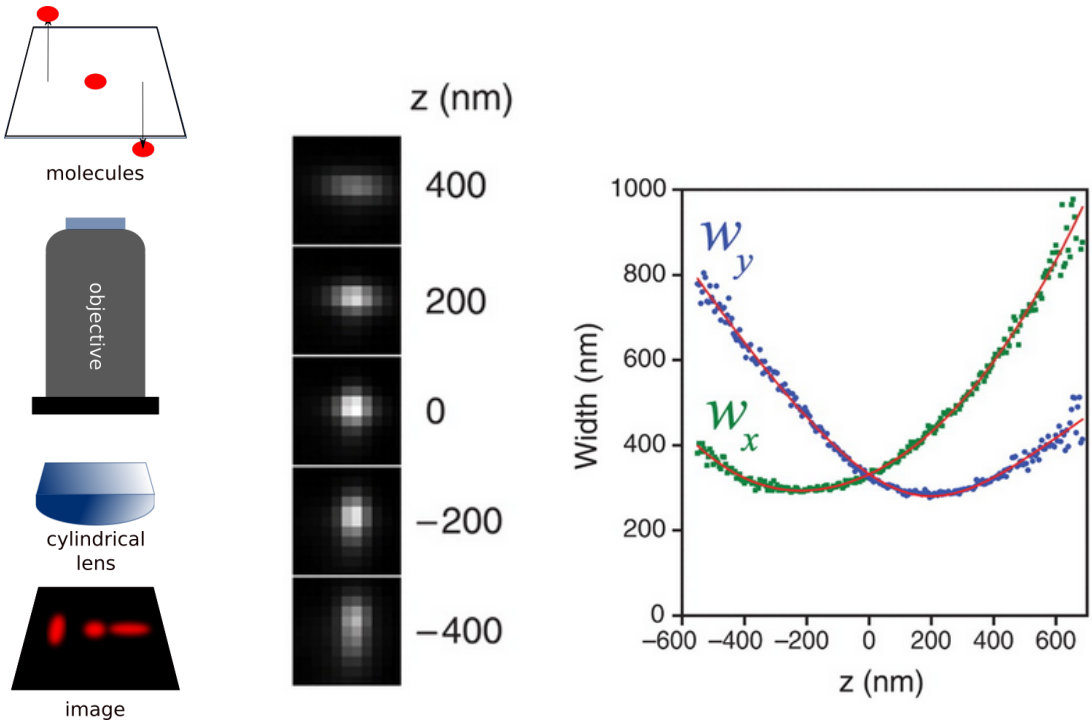
changes to the resolution, visually obvious. Methods for conducting a quantitative assessment of 3D images will be explored in more detail in chapter 4.

3.3.2 The Astigmatic PSF

The central requirement of 3D SMLM is the disambiguation of molecular displacement. The astigmatic PSF can break the symmetry ambiguity while requiring only the addition of a cylindrical lens into the imaging set-up to acquire 3D SMLM datasets. The cylindrical lens causes out-of-focus PSFs to be imaged as ellipses, with the ellipticity of the PSF serving as a measure of its z-position, and the PSFs of molecules below the focal plane being perpendicular to those above it. This requires a calibration curve to map ellipticity to focal displacement, which is generated by moving a single emitter (or multiple, so long as they are easily resolvable from one another) steadily through z. Here, the z-position is obtained from the stage piezo, and the ellipticity is measured using the semi-major axis with a compatible astigmatic localisation software like ThunderSTORM [40, 83].

The functional range and axial localisation precision of the astigmatic method can vary somewhat depending on the physical parameters of the cylindrical lens (or SLM, if this is used to engineer the elliptical PSF instead). If the system is designed to make the calibration curve sharper, the axial precision will be greater, since smaller molecular displacements in z will be required to further 'ellipsify' the PSF. However, this modification will also reduce the axial range of the method because the ellipse's minor axis will quickly become so reduced that it is no longer legible to software, making the PSF irresolvable. Relatively speaking, the astigmatic PSF does not have an outstanding functional range ($\sim$600 nm) or localisation precision (20 nm in the focal plane, to around 60 nm when 300 nm away from it [135]), meaning such adjustments do not offer tremendous benefits. Its main asset is thus its unique ease of implementation, with cylindrical lenses being frequently implemented into commercial STORM systems, and astigmatic calibration and reconstruction features being fully implemented into ThunderSTORM [40]. The accessibility of astigmatic 3D SMLM also makes it the ideal candidate for developing hybrid 3D localisation modalities such as Fast Localization algorithm based on a Continuous-space formulation in 3D (FALCON3D) when combined with biplane imaging and a HD localisation technology [127] or Dual-view Astigmatic Imaging with SAF Yield (DAISY) when combined with DONALD [128, 135].

For the reasons discussed in section 3.1, 3D PSFs are more prone to artefacts. However, the astigmatic PSF does not stand out as especially problematic due to the fact that it is minimally enlarged to achieve 3D localisation – at least when



(a) schema **(b)** PSFs **(c)** calibration curve

Figure. 3.2. A simplified schematic of the astigmatic 3D localisation approach. Molecular z-coordinates can be obtained from the ellipticity of the PSF if a cylindrical lens is introduced into the imaging path (a). This forces an PSF extension along perpendicular axes depending on whether the molecule is positively or negatively displaced from the focal plane. This can be seen more clearly in the astigmatic PSF images in (b), for which the axial positions are indicated. Performing astigmatic SMLM requires that the reconstruction software be calibrated in accordance with the hardware. This is achieved by acquiring a calibration curve as in (c), by using the microscope piezo to steadily move a sample of point emitters (like fluorescent beads) through z while acquiring images. By obtaining the axial location from the piezo, it can be graphed against the major and minor axes of the PSFs acquired from the calibration sample, generating calibration curves to which the researcher can fit experimentally acquired data. (b, c) from *Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy* [83].

compared with some of its most well-known peers [83, 84, 89]. This being said, it still stuffers from unique and troubling artefact-related difficulties. In particular, artefacts induced by emitter overlap do not only bias localisations towards the mutual centre of the overlapping emitters – they also result in severe mis-localisations in z. When two emitters are mis-localised as one in 2D SMLM, the result is often easily removed by data filtration, since the standard deviation of the fitting funciton (usually a Gaussian) will be abnormally large. However, in astigmatic 3D SMLM, large PSFs with strong ellipticity are no longer necessarily an indication that something has gone wrong – it could be that they are simply heavily displaced in z. Thus, they cannot be easily filtered out without also removing localisations that are accurate. In addition, two PSFs that cannot be resolved may simply present to a localisation algorithm as a single, highly ovular PSF. This means that situations that would ordinarily only lead to artificial sharpening now also lead to heavily displaced z-locations. These axial displacements can present much more dramatically than artificial sharpening, whilst often being more difficult to diagnose and remove.

Some of the most rudimentary cases of unique sharpening issues encountered in astigmatic images are presented in Figure 3.3. Here, simulated pairs of adjacent emitters are localised as single, elliptical PSFs located in their mutual centre, causing even more severe mis-localisations in z. Figure 3.3 exhibits a range of relative positioning scenarios, wherein the emitters are in the focal plane, or far removed from it; they may have the same axial position, or be heavily separated. In every case, there is a fundamental ambiguity when sharpening occurs between the two emitters: the resulting artefactual emitter in their mutual centre is heavily displaced in z. This artefact mimics perfectly a 'real' (i.e. credible) single emitter at practically those exact 3D coordinates, which is sometimes impossible to diagnose even visually, let alone algorithmically. The relative positioning of the two emitters in z ultimately determines the severity of the artefacts and the ease with which they can be diagnosed. Figure 3.3a-d is an especially common and problematic case. Individualy, both emitters produce circular PSFs, but when emitting together produce a single elliptical PSF. As with 2D SMLM, this artefactual PSF produces a single localisation between the two emitters in the x/y-plane. However, their shared artefact has an ovular shape that fits well to the astigmatic calibration curve (Figure 3.3c) but is *not* located between the two emitters – instead, it is shifted by a huge amount to a position almost 600 nm below them, beyond the typical functional range of the technique. This produces an almost perfect mimicry of an elliptical PSF at that location (Figure 3.3d), that is, the artefact (Figure 3.3c) is indistinguishable from a single emitter at this location. The position/scatter in localising this combined

PSF is also much lower than for each single emitter, as the increased signal is more significant than any small deviations in shape from the ideal astigmatic calibration curve. Thus, the error does not result from poor fitting, but from a fundamental ambiguity in the relation between emitter positions and PSF shape.

Artefacts generated by molecules that are very displaced from the focal plane to begin with can be similarly severe *and* hard to detect if the semi-major axis of each emitter's PSF is parallel to their separation (Figure 3.3e-h) because there is such a high degree of emitter overlap between the molecules and the output mimics a 'very ovular' PSF so well. If the semi-major axes of the PSFs are perpendicular to their separation, the severity of mis-localisation in z will be even more severe, displacing the molecule by up to a full micrometre (Figure 3.3i-l). However, because the degree of overlap between the PSFs is less pronounced in these kinds of scenarios, their artefacts may, in some cases, be easily removed, as the quality of the fit will be poorer due to high SNR. The resulting artefact can be somewhat easier to notice due to its 'rounded', non-elliptical nature – but this gives rise to a bimodal fitting of the artefact, since a large but circular PSF could be located above *or* below the focal plane – hence the large standard deviation (Figure 3.3k, l) A variety of alternative configurations for the z-locations of each molecule are explored in Figures 3.3m-x. In each case the combined PSF is very similar to a single emitter PSF from the (incorrect) fitted z-position. In spite of the range of relative positionings for the two emitters, *none* of the examples display a case in which the artefactual emitter is axially localised in their mutual centre. This simple situation of just two emitters 300 nm apart in x leads to a variety of complex artefacts when localised in a 3D astigmatic context. For a real structure with emitters separated by various distances in both x and y, the number and nature of artefacts becomes large and complex. The reader should note that in Figure 3.3, the artefacts always displace the molecule to a far-negative z-position – however, the opposite will occur in cases where the molecules are aligned vertically (in y as opposed to x), forcing the molecule to the 'top' of the imaging range.

We wished to investigate whether HAWK still successfully removes 'lateral' artefacts from astigmatic datasets, and whether it can do this at edges of the function z-range of the method. This last point is important because astigmatic reconstructions deteriorate rapidly towards the edges of the functional range (± 300 nm) – the axial and lateral localisation precision decrease, and the tendency towards artefact generation increases. Axial precision loss occurs because fitting the shape of the molecule to the calibration curve becomes more uncertain. Lateral precision loss happens along the axis of ellipticity because the 'ellipsification' of the PSF causes

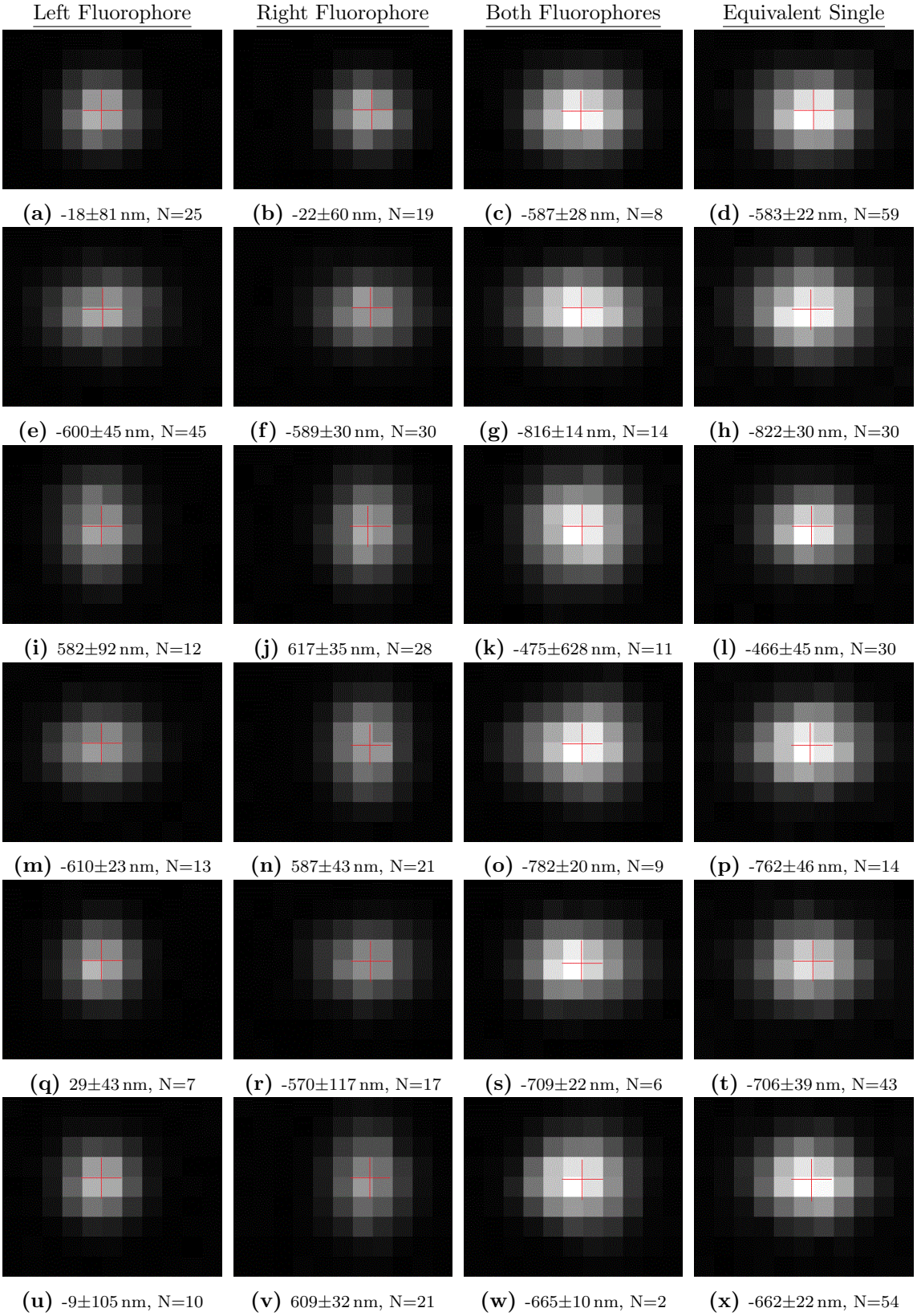

Left Fluorophore
Right Fluorophore
Both Fluorophores
Equivalent Single
(a) -18±81 nm, N=25
(b) -22±60 nm, N=19
(c) -587±28 nm, N=8
(d) -583±22 nm, N=59
(e) -600±45 nm, N=45
(f) -589±30 nm, N=30
(g) -816±14 nm, N=14
(h) -822±30 nm, N=30
(i) 582±92 nm, N=12
(j) 617±35 nm, N=28
(k) -475±628 nm, N=11
(l) -466±45 nm, N=30
(m) -610±23 nm, N=13
(n) 587±43 nm, N=21
(o) -782±20 nm, N=9
(p) -762±46 nm, N=14
(q) 29±43 nm, N=7
(r) -570±117 nm, N=17
(s) -709±22 nm, N=6
(t) -706±39 nm, N=43
(u) -9±105 nm, N=10
(v) 609±32 nm, N=21
(w) -665±10 nm, N=2
(x) -662±22 nm, N=54

Figure. 3.3

Figure. 3.3. Demonstration of common artefacts arising from the mis-localisation of overlapping emitters in astigmatic 3D SMLM. The images show simulated emitter PSFs, and the red crosses indicate the positions of the molecules as determined by the localisation fit. Beneath are the means and standard deviations of the fitted z-positions for each fit. The first column shows a single emitter, and the second column another emitter positioned 300 nm to its right. The images in the third column show images where both emitters are emitting simultaneously, resulting in an artificially ovular PSF. Each image in the forth column diplays a *single* emitter 'equivalent' to those in the third column, positioned at the mean localisation position found from the ThunderSTORM fit of the two-emitter image in column 3. In the first row (a-d), the two emitters are positioned in the centre of the focal range (z=0 nm). The second and third rows show the case for when both emitters are positioned below (e-h) or above (i-l) the focal plane, respectively. In (e-h), the semi-major axis of each emitter's PSF is parallel to their separation and the emitter overlap is large. In (i-l), the semi-major axes are perpendicular to their separation, so there is less emitter overlap. The remaining rows (m-x) show various combinations of z-positions for the two emitters. In each case the resultant localisation when both flurophores are emitting is a single localisation positioned centrally in x but with a large, non-intuitive z-error.

the signal to be distributed over a larger area, reducing the curvature of the fitting function along this axis. Artefacts become more severe as the progressive enlargement of the PSFs results in greater emitter overlap; again, this is especially true along the axis of ellipticity. Figures 3.4 and 3.6 demonstrate HAWK's ability to remove artificial sharpening from this type of data. Figure 3.4 displays 100 nm z-slices of zoom-ins of simulated microtubules from the 3D Localization Microscopy Challenge, processed with and without HAWK. This data represents a classic manifestation of artificial sharpening in the reconstruction of 1D structures, which we might call 'premature fibrillar crossover'. That is, MD and HD datasets of fibres approaching each other often give rise to false structure in which the two fibres prematurely blend into one. HAWK pre-processing clearly enables two fibrils to be clearly resolved at a much tighter proximity in the reconstruction than in the reconstruction without HAWK. In addition, this behaviour was observed for a range of z-positions including one 300 nm above the focal plane.

We also wanted to investigate whether HAWK resolved axial sharpening artefacts. These occur when multiple PSFs are so severely overlapping that they are localised as a single emitter in their mutual centre, forming a PSF with extreme ovularity that will be heavily mis-localised in z. We studied the effects on both microtubules (Figure 3.5) and the nuclear pore complex (Figure 3.6). The microtubule study was conducted using the Alexa Fluor®568 dye, which has poor blinking mechanics that lead to extremely high activation densities in the acquisition. The non-HAWK'd data (Figure 3.5a) exhibits such severe lateral artificial sharpening artefacts that the underlying structure is all but lost. While the HAWK'd data

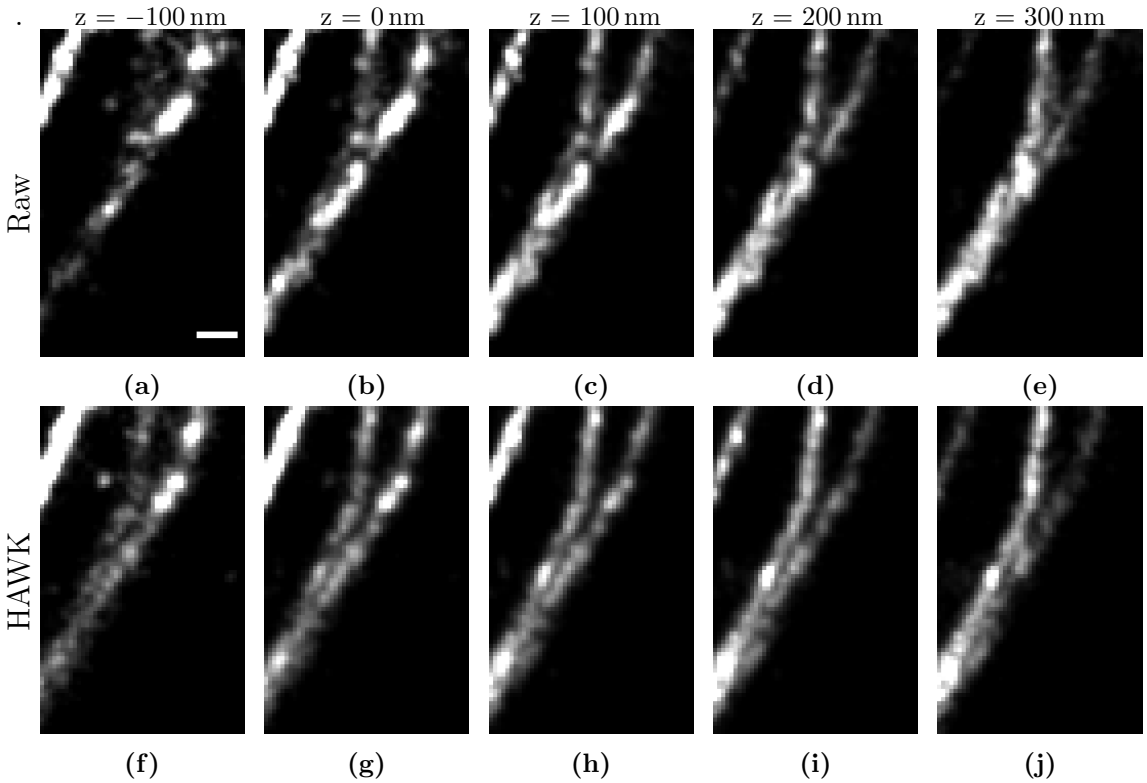
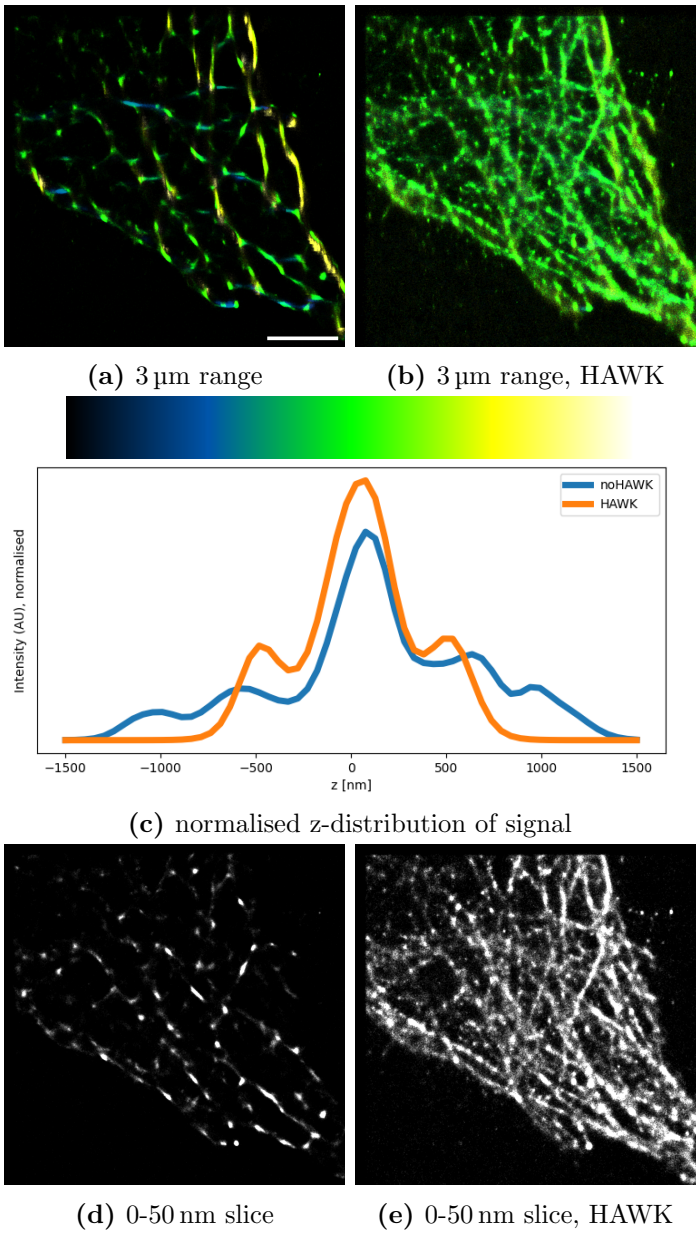


Figure. 3.4. HAWK removes artefacts from HD astigmatic 3D SMLM image crops of tubulin from the Microscopy Challenge. Shown are five 100 nm z-slices of two microtubules without (a-e) and with (f-j) HAWK pre-processing. Without HAWK, the two fibres cannot be discretely resolved in much of each image due to artificial sharpening. HAWK pre-processing removes this sharpening across the full z-range. Scale bar 500 nm.

(Figure 3.5b) shows a significantly reduced resolution due to the effects of temporal filtering, the underlying structure of continuous, fibrous structures has been largely retrieved. Additionally, based on the colouring of the image, the astigmatic localisation algorithm (ThunderSTORM) has clearly placed many of the localisations in Figure 3.5a at an extreme distance outside of the functional z-range, which ordinarily spans approximately 600 nm. Conversely, the HAWK'd image is almost completely green, representing the middle 1000 nm axial range. This discrepancy in the z-distributions of the molecules is reflected in the normalised axial intensity distributions in Figure 3.5c. Here, we see that the non-HAWK'd distribution spans a much wider z-range, from -1500-1500 nm, while the HAWK'd distribution spans less than half that range. The central peak of the non-HAWK'd data is slightly thinner, which points to the precision loss induced by HAWK's temporal filtering. This has the expected effect of increasing the apparent size of the PSFs and thus slightly affects their perceived ovularity, increasing their axial distribution. Here, we note that the majority of structures from the full range of the HAWK'd dataset in (Figure 3.5b) are clearly resolved in the focal plane (Figure 3.5e) and the continuity of the microtubules has not been broken by astigmatic artefacts. However, in the non-HAWK'd data, what little continuity between fibres that present in the full axial range (Figure 3.5a) has been broken, pointing to the severity of astigmatic artefacts. This is also clear from the erratic distribution of colouring in Figure 3.5a, which clearly suffers from discontinuities.

Figure 3.6 shows 100 nm z-slices of data to exhibit HAWK's ability to successfully remove 2D sharpening from astigmatic datasets, but this time, this is displayed using experimental data of the nuclear pore complex. Here, we note that in the non-HAWK case there is significant undersampling of data in one of the pores when compared with the HAWK'd reconstruction. This suggests that many of the localisations have been either filtered, or thrown out to far-off z-positions (as with the simulations in Figure 3.3). Both diagnoses suggest the same thing: that the localisation algorithm conflated multiple PSFs, localising huge PSFs with an excessive diameter or ellipticity. Both pores (the holes in the middle) can be resolved over the entire examined z-range in the HAWK'd reconstruction; this is evidently untrue in the raw reconstruction.

We also needed to interrogate HAWK'd reconstructions from astigmatic data for any abnormalities, particularly when compared with LD data (which would be expected to have very few artefacts). Whilst we already knew that the lateral resolution of HAWK'd reconstructions is somewhat reduced by the filtration process to a degree correlating with the emission density of the acquisition, we did not know



(a) 3 µm range **(b)** 3 µm range, HAWK

(c) normalised z-distribution of signal

(d) 0-50 nm slice **(e)** 0-50 nm slice, HAWK

Figure. 3.5. HAWK removes artefacts in all three dimensions from ultra-high density astigmatic datasets of microtubules stained with Alexa Fluor®568. Images were acquired using ThunderSTORM SE 3D astigmatic fitting. The images are not filtered and reconstructed over a large z-range, from 1500 nm below the focal plane to 1500 nm above it (a, b). This is significantly larger than the functional z-range of the technique, but serves to illustrate the severity of the artefacts encountered in astigmatic data. The colour code (indicated with a scale bar that lines up with the normalised axial signal distribution, (c)) indicates the localised z-positions of the molecules. The central 50 nm slice of the dataset is depicted for both the HAWK'd and non-HAWK'd data in (d, e). Scale bar 5 µm.

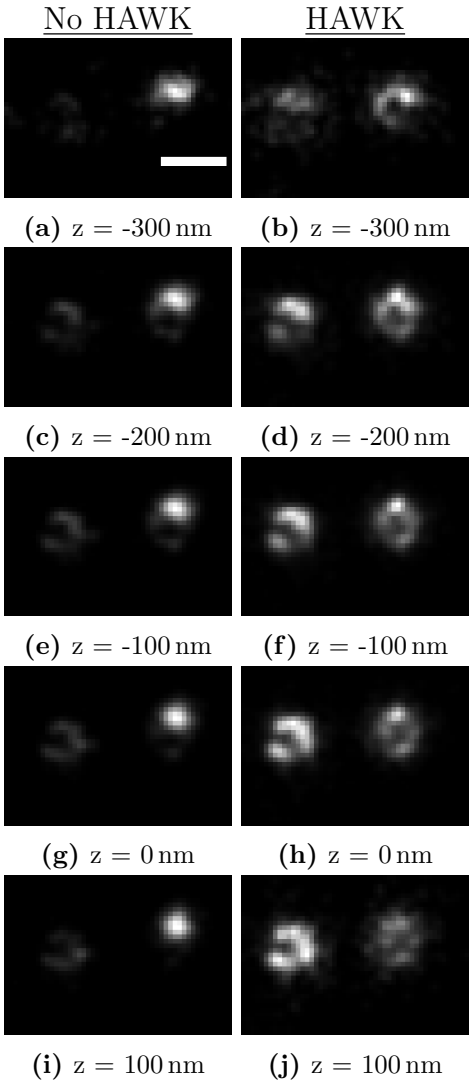


Figure. 3.6. HAWK removes artefacts from high-density astigmatic 3D SMLM image crops of the nuclear pore complex from the microscopy challenge. Shown are 100 nm z-slices for 2 pores both without (left column) and with (right column) HAWK. Each row is a 100 nm z-slice centered on z=−300 nm (a, b), z=−200 nm (c, d), z=−100 nm (e, f), z=0 nm (g, h) and z=100 nm (i, j). Of particular note is the fact that HAWK can resolve the right pore over a much wider z-range, wheras in the raw reconstruction many of the localisations are sharpened to the top-right edge of the pore. Scale bar 200 nm.

that the same thing would happen to the axial resolution when dealing with 3D datasets. In particular we noticed this happening in astigmatic datasets, and sought to diagnose and resolve the issue.

3.3.3 Loss of z-precision in HAWK-processed astigmatic reconstructions

It is well-established that pre-processing HD datasets with HAWK slightly reduces the overall xy-resolution of the reconstruction [78]. This was happens largely because the filtration removes a proportion of the signal (meaning it will remove more total signal given a higher density of emissions), reducing the curvature of the fitting function and thus the localisation precision. Experimenting with HAWK on astigmatic 3D SMLM datasets also revealed a slight resolution loss along the z-axis, as can be seen by the wider central peak in Figure 3.5. The mechanism of localisation in xy (location of PSF centroids) is radically different to that in z (PSF shape), which intuitively suggests a different causal mechanism for this precision loss. We theorised that a possible cause would relate to HAWK's generation, and subsequent handling, of negative signal.

HAWK-processed datasets display the acquisition blinking events rather than the signal itself. Molecules 'turning on' are represented by positive signal, and molecules 'turning off' are represented by negative signal. There are two standard ways for HAWK to handle these negative signal peaks, and thus two ways in which they might affect astigmatic molecule fitting:

1. The absolute value of the Haar-convolved intensity trace $|H(t')|$ is passed to the reconstruction software. This method works well when using ME fitting [73] as it leaves proximal switch-on and switch-off events in the same frames. However, this makes the method more vulnerable to sharpening as localisations are biased towards the mutual centre of these events where their signal has accumulated. In addition, the increase of signal present between the molecules would theoretically increase their ellipticity along the axis connecting them, slightly displacing the localisation of the PSF in z.

2. The Haar-convolved intensity trace $H(t')$ is separated into two haar sequences, $H^-(t')$, representing only the negative signal, and $H^+(t')$, representing the positive signal. The entirety of $|H^-(t')|$ is then appended to $H^+(t')$, keeping proximal switch-on and switch-off events in separate frames and making the reconstruction maximally resistant to sharpening. However, these events will bias the localisations *away* from their mutual centre, where the magnitude of

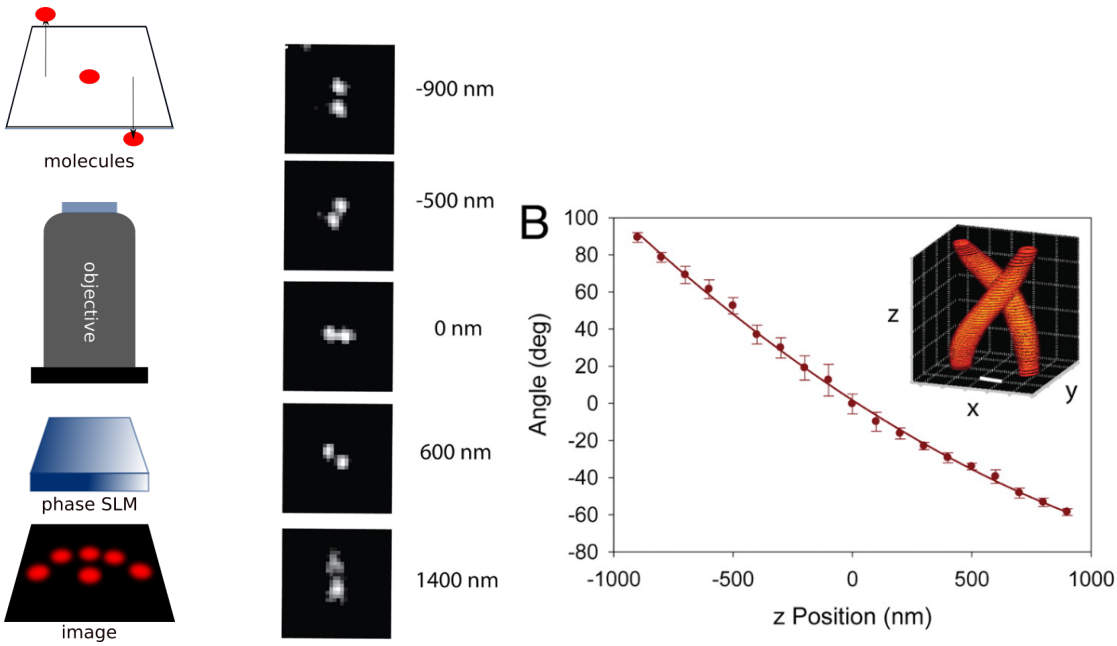
the negative peak has a deleterous effect on the magnitude of its neighbour. This increases localisation spread, reducing xy-resolution. Furthermore, the reduction of signal in-between the molecule pair would theoretically serve to increase their ellipticity along the angle perpendicular to the axis of their separation, causing a slight displacement in z.

In both cases, the subtle warping of the PSFs due to the adjacency of a negative and positive signal peak in the HAWK'd dataset could cause an even more significant displacement in z than in xy when handling astigmatic 3D datasets. Minimising the possibility of bias incurred by this configuration of blinking events would require a third way of handling negative peaks, that is, *not* handling them. The warping of PSF intensity profiles is necessary because standard localisation software cannot fit a negative signal peak, but it leads to problems of localisation bias. In other words, we thought that if the localisation software was edited to fit the negative signal peaks and localised using ME fitting, then the positional bias incurred by HAWK would be minimised. However, implementing negative fitting into ThunderSTORM did not, in fact, have the expected impact on 3D datasets, instead only showing a marginal decrease in HAWK-induced lateral precision loss. Further investigations into z-precision loss when HAWK is used on astigmatic datasets will need to be conducted to ascertain its cause.

3.3.4 The Double-Helix PSF

The DHP is another relatively popular PSF-engineering method [84]. It benefits from a relatively simple implementation process, requiring the addition of a $4f$ imaging system, and a phase-only SLM at the Fourier plane whose phase mask ins the inverse Fourier transform of the DHP, to the imaging set-up to disambiguate z-displacement symmetry [136]. The SLM splits the PSF into two lobes whose mid-point serves as the xy-location of the molecules, and the mutual angular rotation measures its z-displacement. This approach was developed to attain a high and uniform Fischer information (section 1.4.1), preserving a higher z-localisation precision throughout its functional range. As such, it boasts a significantly increased z-localisation precision when compared with the astigmatic PSF [136] with localisation precisions as low as 10 nm laterally and 20 nm axially over axial ranges of up to 3 µm [84].

Like other methods that aim to maximise the 3D Fischer information for the molecule (such as the tetrapod PSF [89], section 1.6), the DHP can suffer experimentally from its large diameter, as it comprises two well-resolved PSF lobes. Emitter overlap (or perhaps more accurately in this case, 'PSF ambiguity') occurs more



(a) schema **(b)** PSFs **(c)** calibration curve

Figure. 3.7. Diagram demonstrating how a double-helix imaging set-up localises molecules in z. Not unlike the astigmatic schema, the DHP is generated by altering the PSF shape before imaging. To generate a two-lobed PSF, the acquired signal is passed through a phase-only SLM (a). This forces the lobes' mutual rotation about their mid-point to increase/ decrease as the molecule is diplaced positively/ negatively from the focal plane. The effect can be seen more clearly in the DHP images in (b), for which the axial positions are indicated. As with the astigmatic PSF, performing 3D SMLM with a DHP requires that the reconstruction software be calibrated in accordance with the hardware. This is achieved by acquiring a calibration curve as in (c), by using the microscope piezo to steadily move a sample of point emitters through z while acquiring images. By obtaining the axial location from the piezo, it can be graphed against the mutual roation of the PSF lobes obtained by imaging the calibration sample, generating calibration curves to which the researcher can fit experimentally acquired data. (b, c) from *Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function* [84].

easily when utilising the DHP than it does for other methods since the PSF enlargement is so severe. Excepting the ultra-low emission density cases, the localisation software required to generate a reconstruction from these datasets can often struggle to confidently discern *which lobe belongs to which PSF*. This is a particular problem for the DHP, as it does not necessarily even require that the PSFs be overlapping – they need only be close enough for a high uncertainty in PSF 'ownership'. An exploration of some the types of ambiguity that can arise from only a single pair of molecules is explored in Figure 3.8.

It should be noted that because the localisation mechanism for this PSF is so different, it requires the use of specialist localisation software of which there is a much more limited selection than for 2D SMLM or for the astigmatic method. These software have differing methods for handling PSF ambiguity, but often DHPs comprising lobes whose ownership is too ambiguious are simply discarded. We used one such software, *Easy-DHPSF*, in our experiments. It constitutes one of the most recent reconstruction software, developed in collaboration with W. E. Moerner who created the original DHP technology for SMLM [84, 132]. It was our preference to use a software that handled ambiguious fits in this way, as it meant that increasing emitter density, rather than producing strange and complex artefacts, generally resulted in a progressive undersampling, which is a highly salient visual cue that the emitter density has exceeded the capacity of the algorithm. The effects of undersampling with increasing emitter density are displayed in Figure 3.9. Here, we were able to simulate multiple activation densities to exhibit the rapid deterioration in reconstruction quality when using the DHP. At the highest emitter density displayed (Figure 3.9e, f), almost nothing remains of the non-HAWK'd image (Figure 3.9e), while the HAWK'd image has 20 times more localisations (Figure 3.9f). This figure shows that HAWK's temporal filtration disambiguates the lobe pairs, restoring data to the image which, in some cases, is otherwise so undersampled as to be unusable. In fact, the contrast in the images of this figure was heavily reduced in order to ensure that the image rendered with visible localisations.

We also conducted this test on experimental data of nuclear lamin (Figure 3.10). This was an important control to verify whether use of HAWK might introduce novel issues by changing the z-coordinates of a fixed cell dataset. Visualising the reconstruction in MATLAB allowed us to see the localisations and their coordinates more clearly (Figure 3.10a-d) Even on an experiment on a fixed cell, the HAWK'd data has roughly five times as many localisations present than the non-HAWK'd data . We also generated histograms displaying the distribution of the molecules in z (Figure 3.10e, f). The results indicate that HAWK successfully increased the

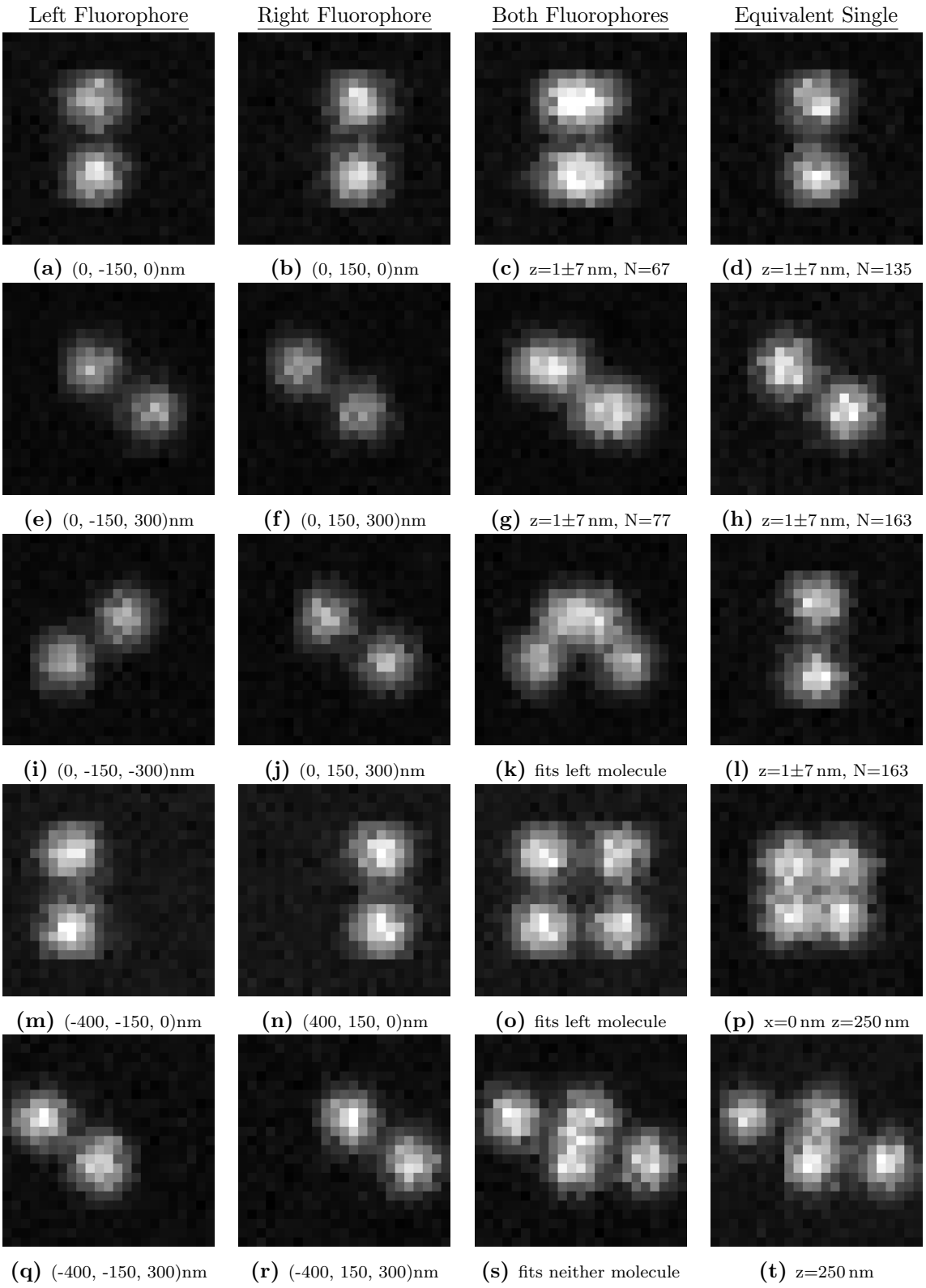
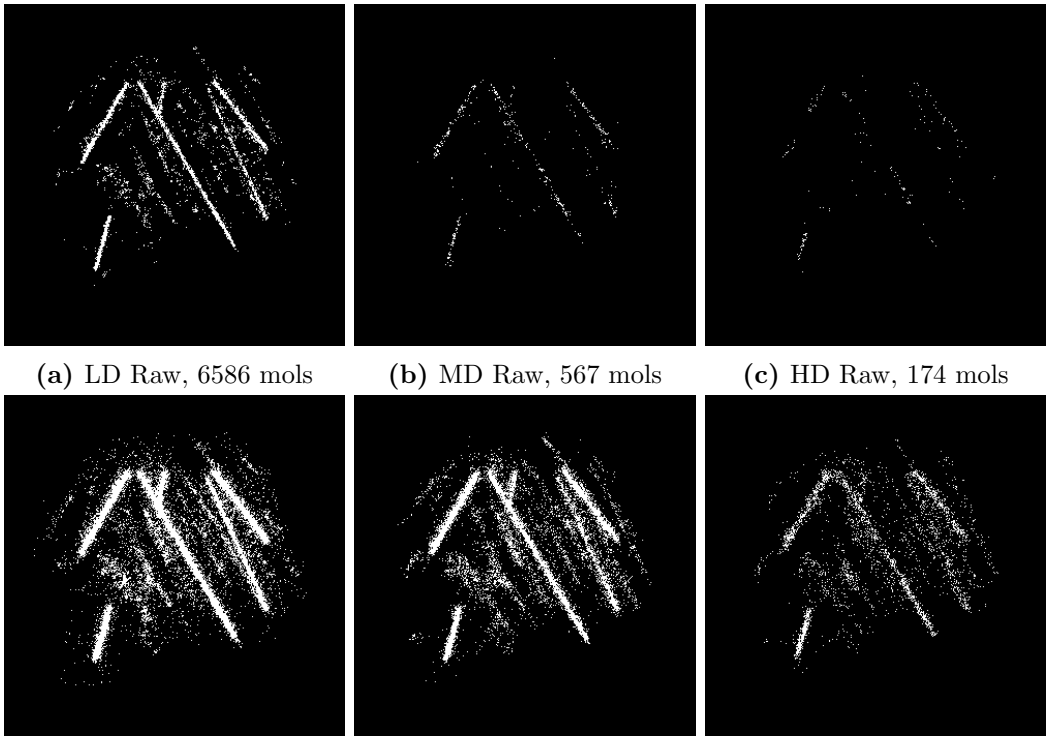

Left Fluorophore
Right Fluorophore
Both Fluorophores
Equivalent Single
(a) (0, -150, 0)nm
(b) (0, 150, 0)nm
(c) z=1±7 nm, N=67
(d) z=1±7 nm, N=135
(e) (0, -150, 300)nm
(f) (0, 150, 300)nm
(g) z=1±7 nm, N=77
(h) z=1±7 nm, N=163
(i) (0, -150, -300)nm
(j) (0, 150, 300)nm
(k) fits left molecule
(l) z=1±7 nm, N=163
(m) (-400, -150, 0)nm
(n) (400, 150, 0)nm
(o) fits left molecule
(p) x=0 nm z=250 nm
(q) (-400, -150, 300)nm
(r) (-400, 150, 300)nm
(s) fits neither molecule
(t) z=250 nm

Figure. 3.8

Figure. 3.8. Example ambiguities and fitting failures produced in DHP SMLM datasets. Shown are simulated DHPs from both single and overlapping emitters. The first column shows a single emitter located left of centre in x, and the second column shows an emitter right of centre. The third column shows a combined PSF from both emitters fluorescing simultaneously, and the final column shows an example of DH-PSF(s) that are roughly equivalent those in the combined case. Each row shows different values for the separation of the emitters in x and z. The statistics in the caption show the results from repeated fitting of these simulated PSFs. The first row (a-d) shows emitters separated by 300 nm in x but both positioned at the same z-height (0 nm). PSFs resulting from each emitter emitting individually are shown in (a) & (b). An example combined PSF when both are emitting simultaneously in shown in (c). Fitting of these PSFs only gives the z-positions shown in the captions. (d) shows a single emitter PSF from a molecule located centrally between the two emitters. Fits to these PSFs give very similar results to the overlapping case (c) indicating an ambiguity in the PSFs. The second row (e-h) shows two emitters at the same separation, but a z-height of 300 nm. Again, a single emitter placed centrally between the two (h) gives similar fitting results to the overlapping case (g) but now there is a small bias in the fitted z position of the overlapping emitters as well as an averaging of their x positions. In (i-l) the emitters are separated by 600 nm in z as well as 300 nm in x. In this case the combined PSF (k) could not be fitted reliably by the software with these PSF either being ignored of fitting just the left component of the two. In this case the combined PSF does not resemble a single emitter at the central location (l) and there is no ambiguity, only missing localisations. In (m-p) there are two emitters at the same z height (0 nm), but a larger separation in x (800 nm). In this case the combined PSF (o) shows a visually obvious ambiguity with two molecules separated in y by 800 nm at a z height 600 nm higher or lower. The fitting algorithm, however again principally fits just the left emitter, presumably because of the pattern matching feature of the *EasyDHPSF* software. Shown in (p) is another potential ambiguity from two emitter separated in z, but not in x, that gives a somewhat similar combined PSF. Here the individual PSFs cross each other, rather than in being aligned parallel. (q-t) has the z-position of both emitters at 300 nm. Again, the complex combined PSF (s) is largely ignored by the fitter, as it does not match one of the pre-selected patterns. However (t) shows another potential ambiguity with two other emitters at different x,y & z positions. Here the individual emitter PSF corresponds to the upper and lower pair of lobes rather than the left and right pair in (s).



(a) LD Raw, 6586 mols **(b)** MD Raw, 567 mols **(c)** HD Raw, 174 mols

(d) LD HAWK, 32359 mols **(e)** MD HAWK, 14592 mols **(f)** HD HAWK, 3587 mols

Figure. 3.9. HAWK removes artefacts from simulated DHP SMLM datasets reconstructed with *Easy-DHPSF* at three different densities. The datasets have been projected in x to visualise them 'from the side'. (a-c) display the non-HAWK'd versions at three increasing activation densities, while (d-f) show the same dataset that has been pre-processed with HAWK. The MD activation density is 3 times higher than the LD activation density, while the HD activation density is 10 times higher. We also show the number of molecules successfully localised in each scenario. All simulations are labelled with a sparsity of 10 nm.

number of localisations by approximately five-fold, without creating any novel artefacts that might alter the z-coordinates of the localisations, since the profiles of these histograms are superimposable . Again, although *Easy-DHPSF* ensures a low incidence of artefacts, HAWK is still required to optimise the reconstruction as it restores localisations that are lost to the filtration process, and this experiment demonstrates that it does so without introducing any new problems.

3.3.5 Multifocal Plane SMLM

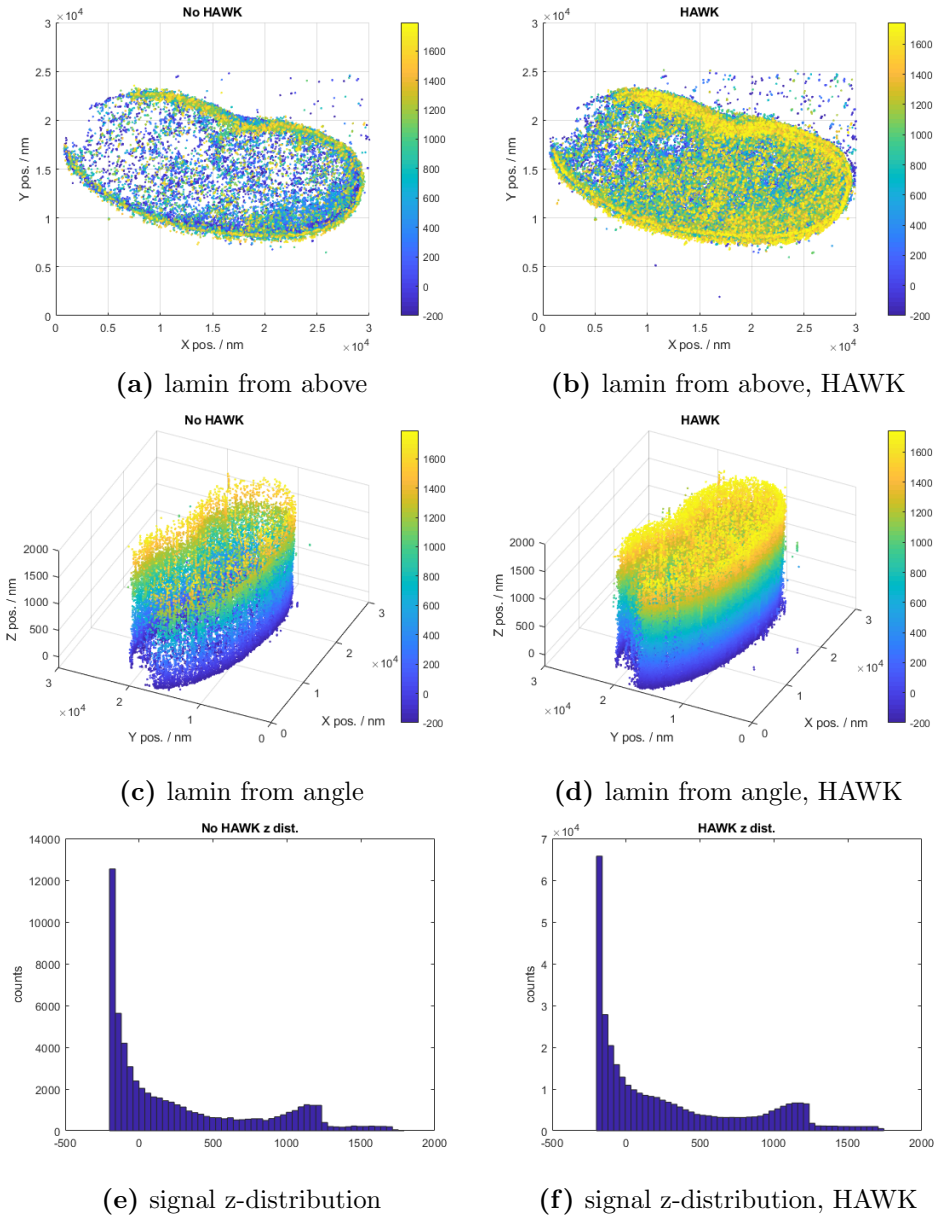
3D SMLM can also be achieved without PSF engineering, as the symmetry ambiguity can also be broken by imaging two (or more) planes at once. That is, the diametric ratio between the two images of a PSF (as imaged in each plane) will correspond to a specific z-location. This approach comes with a major advantage: in the absence of PSF engineering, which makes the PSF larger, emitter overlap and ensuing artefacts are less likely. In addition, the axial range/ precision trade-off can be calibrated by changing the separation of the two planes with a greater separation corresponding to a greater range but a poorer z-precision and vice versa. It does, however, come with its own difficulties.

1. Biplane microscopy can be difficult to implement – it requires two cameras with a perfectly calibrated difference in the imaging light path length, to image both planes at once.

2. Splitting the signal between two cameras, predictably, lowers the amount of signal being imaged by each, reducing the fitting function curvature and thus the localisation precision. This lowers the resolution of the final image by increasing emitter spread. In other words, biplane microscopy is a less sensitive technique than the previously discussed methods, which can make it more difficult to detect faint or low-contrast features in a sample. This can be avoided by imaging each plane with a separate objective, allowing the 'full' amount of light to be collected from each PSF – but this can be prohibitively difficult to implement correctly for many experimenters.

3. Aligning the two planes, such that the perceived location of both images of the emitter is identical, can be tricky.

As there is no PSF engineering involved when performing biplane microscopy, it does not suffer from many of the issues present in those methods. For this reason we did not always test biplane microscopy using the same simulations and types of



(a) lamin from above

(b) lamin from above, HAWK

(c) lamin from angle

(d) lamin from angle, HAWK

(e) signal z-distribution

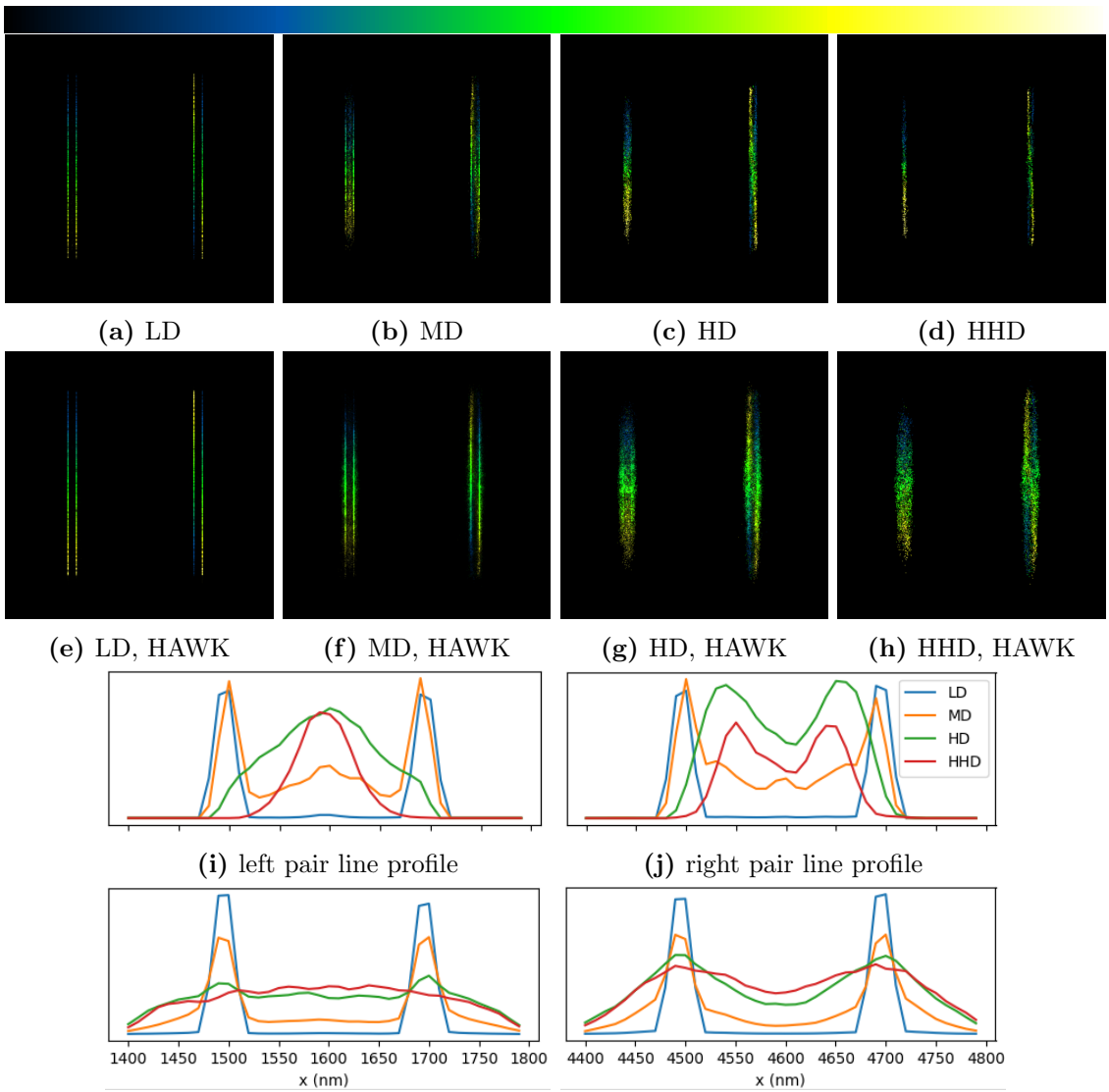
(f) signal z-distribution, HAWK

Figure. 3.10. HAWK restores undersampled data to a lamin DHP dataset, stained with AF647. (a-d) display the localisations plotted on a graph generated in MATLAB, while (e, f) display the distributions of the molecules in z. Here the localisations are displayed in MATLAB for clarity because image file formats suffer from low visibility due to the relatively small number of localisations. The z-distribution histograms (e, f) were also generated to ensure that HAWK pre-processing did not create any novel artefacts that might re-distribute the molecules in z.

data as with e.g. astigmatic microscopy. For example, we did not use a simple two-emitter simulation to test the biplane system as in Figure 3.3, as elliptical PSFs with extremely large diameters are erroneous in a biplane dataset and it is most prudent to simply filter them out. Instead, we were primarily curious to see how biplane microscopy handled more complete structures across an experimentally relevant z-range, including line pairs (Figure 3.11) and mitochondria-like two-dimensional 'pill' structures (Figure 3.12).

Figure 3.11 displays non-intersecting line pairs propagating in three dimensions. In the LD reconstructions (Figure 3.11a, b), all line pairs are equally well-resolved whether HAWK is used or not. This functions as a control to demonstrate that HAWK does not introduce any novel issues on LD biplane data. In the MD reconstructions (Figure 3.11c, d), we begin to see artefactual localisations in the non-HAWK'd reconstructions between the two lines. These are still notably absent in the HAWK'd data. The highest-density reconstructions (Figure 3.11e-h) show the line becoming almost completely condensed into a single structure when HAWK is not used (Figure 3.11e, g), reducing the apparent distance between them. The HAWK'd versions (Figure 3.11f, h) display a reduced precision due to HAWK's temporal filtering, but the apparent separation of both pairs of lines remains unchanged when compared with the LD reconstructions. The signal profiles of these reconstructions, taken horizontally across the images (Figure 3.11i, l) reflect HAWK's ability to remove artefacts from this simulation, regardless of density. It should be noted that the HHD+HAWK reconstruction of the parallel lines profiled in Figure 3.11j exhibits such severe precision loss due to extreme activation density that the peaks representing the two lines are not resolvable. Importantly, however, this is still not a problem of sharpening, only reduced resolution. Thus, HAWK does not mislead the experimenter with a biased representation of the line separation – instead, it is clear that the experiment should be tried again, with different acquisition parameters or fluorophores.

Figure 3.12 displays simulations of mitochondria-like pill-shaped structures located 300 nm below the focal plane. Here we see that issues introduced into biplane SMLM reconstructions by an increased activation density in the acquisition dataset are remedied by HAWK pre-processing. For instance, substantial distortion of the tubular structure is observed in the MD reconstruction in Figure 3.12b. The hollow nature of the structure visible in the LD reconstruction (Figure 3.12a) is equally visible in the HAWK pre-processed case (Figure 3.12c). The LD z-profile of the reconstruction (Figure 3.12d) shows a slight shoulder indicating there are still some mis-localisations in z, even at this low density. However, the MD z-profile (Figure



(a) LD **(b)** MD **(c)** HD **(d)** HHD

(e) LD, HAWK **(f)** MD, HAWK **(g)** HD, HAWK **(h)** HHD, HAWK

(i) left pair line profile **(j)** right pair line profile

(k) left pair line profile, HAWK **(l)** right pair line profile, HAWK

Figure. 3.11. HAWK removes artefacts from HD biplane SMLM simulations of straight line pairs with a 200 nm separation, at multiple emitter densities. The lines on the left are parallel to one another (both coming 'out of the page' from the top to the bottom of the image), while those on the right are perpendicular (left going 'into the page', right coming 'out of the page'). The relative z-positions range from -1500 nm (dark blue) to 1500 nm (yellow/white) through the spectrum indicated above the images, with the 'focal plane' at 0 nm highlighted in green. The z-stack was rendered in increments of 100 nm. The first row of data depicts the non-HAWK'd reconstructions (a-d) and the second row depicts the HAWK'd reconstructions (e-h). The lowest-density reconstructions (a, e) are reconstructed from data emitting 10 times less densely than the medium-density reconstructions (b, f), 30 times less densely than the high-density reconstructions (c, g), and 100 times less densely than the ultra-high-density reconstructions (d, h). The signal profiles of these reconstructions (i-l), taken horizontally across the images, are generated from a vertical projection of the signal.

3.12e) shows very large distortions when compared with the LD profile. There is a second major peak at ∼850 nm and a third smaller one at ∼-250 nm. These localisations have substantial bias in z, to the point where they are outside the range of the structure which ranges from 0-500 nm in this case. The z-profile when HAWK is used (Figure 3.12f) is superior even to the low density case, with a smaller shoulder for localisations at extreme z-positions. Here the hollowness of the structure is cearly visible even in the MD case, although the gap between the walls is narrower than either the LD or MD HAWK pre-processed cases (which reflect the 500 nm width of the structure). The relative clarity of the MD slice (Figure 3.12h) may not be intuitive given how obviously sharpened the 2D projection is (Figure 3.12b). This is because the localisations that would be located in the central positions have very large z-bias, and therefore are not present in the z-range of this slice (100 nm). These images look very similar in all three cases, and represent the typical 'artificial sharpening' artefacts expected in 2D and observed in the MD 2D-projection (Figure 3.12b) What is dramatically different in the MD cases (Figure 3.12k, n) is the much larger proportion of localisations in these z-slices (where these is no actual structure). With HAWK pre-processing the proportion of these mislocalisations is dramatically reduced to a similar number as the LD data.

Figure 3.13 exhibits artefact removal with HAWK from a biplane dataset of mitochondria using a stain of the epitope TOM20. Two regions have been highlighted for further analysis using plots of the axial signal distribution (Figure 3.13c, d). These indicate that the HAWK+HD reconstruction much more closely resembles the LD reconstruction than the HD reconstruction without HAWK, indicating that HAWK processing has corrected bias in the axial locations of the molecules. It can be seen that there is substantial difference in the z-distribution of localisations from the HD (12000-2000 frames) and LD (<4000 frames) parts of the acquisition. For both regions, 1 and 2 there are substantial peak shifts in the molecule z-distributions (200 nm and 400 nm, respectively) that are caused by emitter overlap. The shift is greatly reduced, even eliminated, with HAWK pre-processing. This is made even more apparent when we look at the z-slices of region 1 that contain significant peak differences between the LD and HD reconstructions (Figure 3.13e-m). The HAWK reconstructions clearly match the LD images much more closely than the HD reconstructions. It is particularly noteworthy that in the −450 nm HD reconstruction (Figure 3.13i) the mitochondrial 'hole' is still resolvable despite the apparent excess density. This is because the overlapping emitters that would normally lead to collapse of this structure form an oval PSF that is localised to a different z-position. These incorrect localisations are thus removed from this z-plane (Figure 3.13h) and

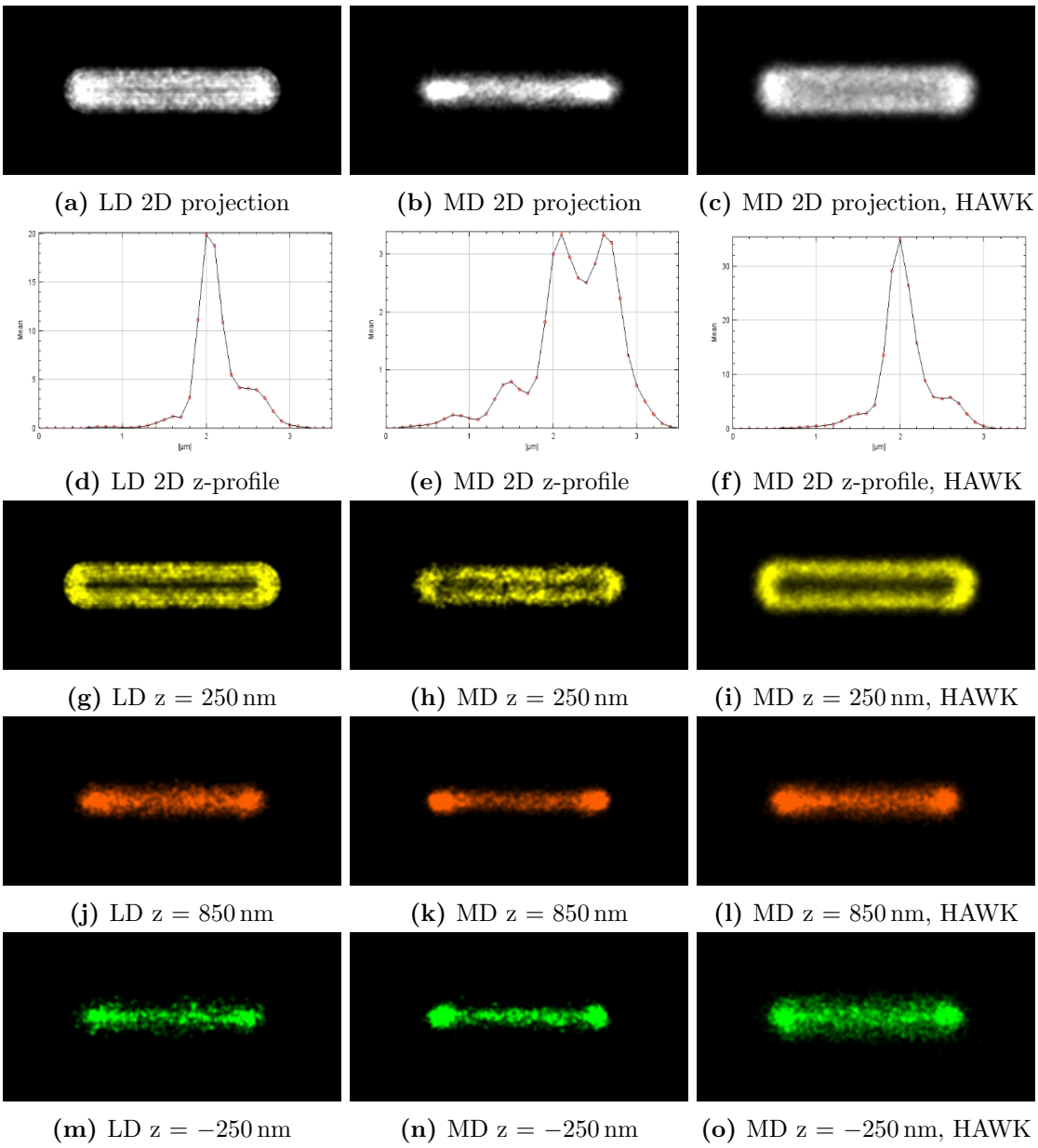
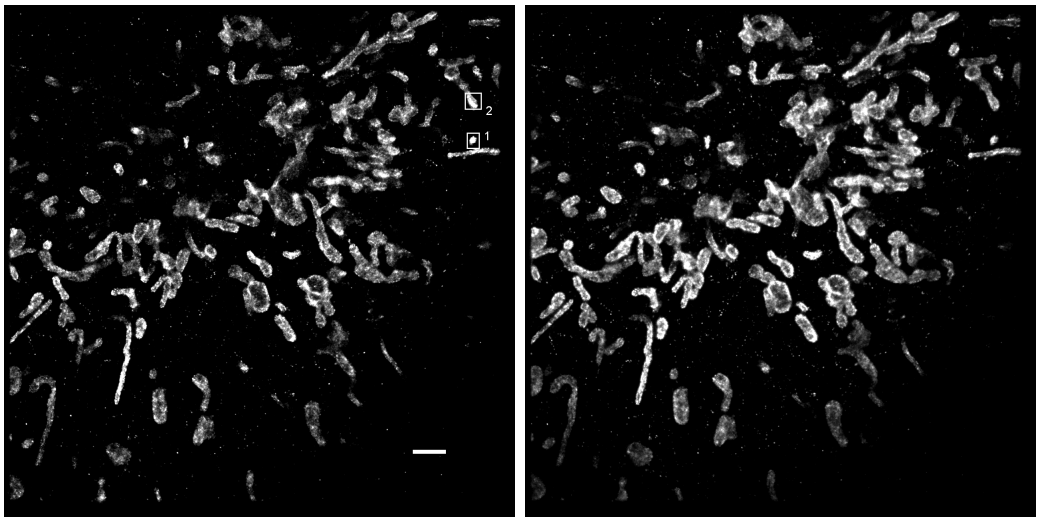
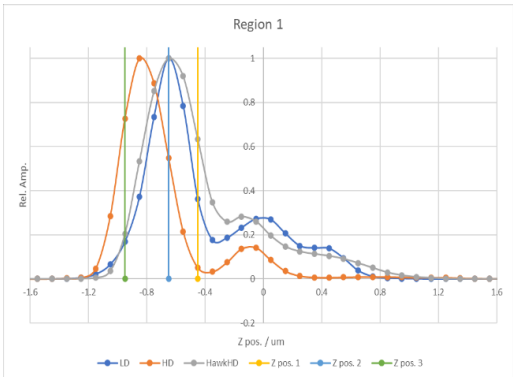


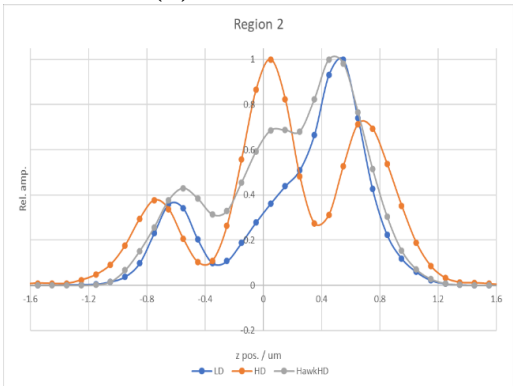
Figure. 3.12. Simulated mitochondria data at low and medium density show HAWK greatly reduces mislocalisations in 3D biplane datasets. Results are shown for ThunderSTORM biplane fitting of low-density data (left column), mid-density data (centre) and HAWK pre-processed mid-density data (right). The top row shows 2D projections of the reconstructions for low-density (a) and mid-density (b) data compared to mid-density data with HAWK pre-processing (c). The second row (d-f) indicates the distribution of localisations in z in each case. The third row (g-i) shows the central 100 nm z-slice of the simulated microtubule structure. The bottom two rows show z-slices centred around the second and third peaks of the mid-density z-distribution, at 850 nm (j-l) and −250 nm (m-o).



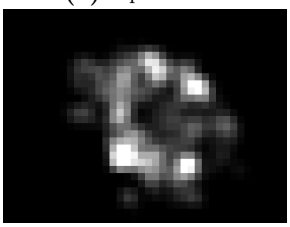
(a) TOM20 **(b)** TOM20, HAWK



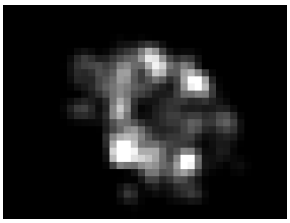
(c) z-profiles of ROI 1



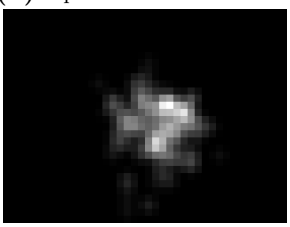
(d) z-profiles of ROI 2



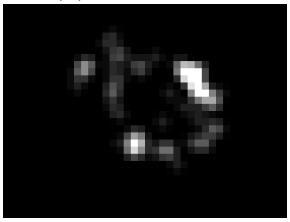
(e) LD z=-450nm



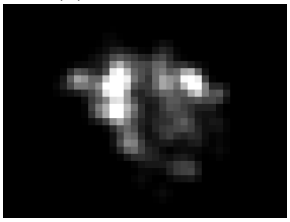
(f) LD z=-650nm



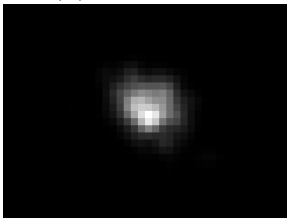
(g) LD z=-950nm



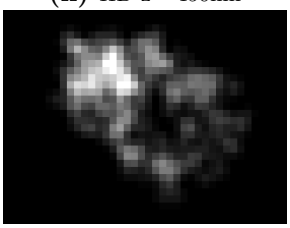
(h) HD z=-450nm



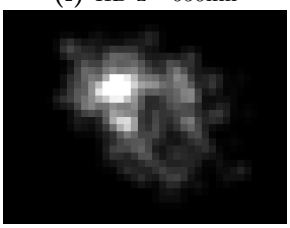
(i) HD z=-650nm



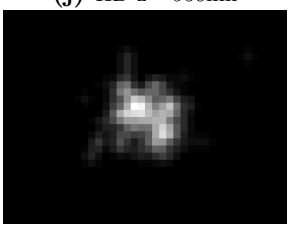
(j) HD z=-950nm



(k) HD z=-450nm, HAWK



(l) HD z=-650nm, HAWK



(m) HD z=-950nm, HAWK

Figure. 3.13

Figure. 3.13. Reconstructions of 3D mitochondria data using a TOM20 stain show the distortions in z-position arising from excess emitter density, and how they are greatly reduced by HAWK. (a, b) show 2D projections of 3D localisation data using ThunderSTORM (a) and HAWK + ThunderSTORM (b). Evidence of artificial sharpening artefacts are present in the raw reconstruction (a) compared to when HAWK is used (b), of which two examples are indicated in boxes 1 and 2. Here the non-HAWK'd distributions is taken from two subsets of the localisation data. In each case the red line (high-density) refers to localisations from the first 4000 frames of the acquisition where the emitter density is sufficient to cause PSF overlap and therefore localisation errors. The blue line (low-density) is obtained using localisations from 12000-20000 frames, where the emitter density is lower with fewer artefacts. The gray curve (high-density, HAWK) represents localisations from the HAWK-processed data coresponding to the first 4000 frames, as with the high-density data. Region 1 is examined in more detail for 3 slices of the z-stack in the subsequent images (e-m). Three 100 nm z-slices of the reconstructions (−450 nm, −650 nm and −950 nm) are shown. These correspond to significant positions in the z-distributions, as indicated by the vertical lines in plot (c) (yellow, light blue and green, respectively). Slices from the reconstructions are shown for low density (e-g), high density (h-j) and HAWK pre-processed (k-m) data. Scale bar 2 µm.

moved to another – in this case, to the −950 nm reconstruction (Figure 3.13j) This results in this image containing almost entirely incorrect localisations. These errors are difficult to spot as they appear similar (but overwhelm) the true structure as indicated by the LD (Figure 3.13g) and HAWK (Figure 3.13m) reconstructions, and can only be determined unequivocally by observing the difference in z-distributions.

3.4 Discussion

Initial tests of HAWK on three of the most common 3D modalities have proven highly promising. As in the initial HAWK publication [78], lateral SMLM artefacts – in which multiple emitters are localised with a bias towards their mutual centres, or a base assumption of sparsity in a cluster of emitters leads to mis-localisation – were almost entirely removed by HAWK pre-processing from astigmatic, DHP, and biplane datasets. This outcome, though it is of course essential for the adoption of the technique in 3D contexts, is not hugely surprising, given that the effects of temporal filtration on pixel intensity-time traces were not expected differ greatly due to minor alterations to the PSF shape. However, prior to experimentation there remained a concerning possibility that HAWK filtration might have a deleterous effect on *axial* localisation precision on these kinds of dataset. As we know (section **??**), HAWK can be thought of as an approach that 'swaps' localisation bias (which can render a dataset worthless) for a small amount localisation uncertainty (which renders an image with lower resolution). The scale of HAWK's localisation precision

loss is related to the scale of artefacts it has removed; that is, the precision loss incurred by HAWK scales with the number and severity of artefacts that would otherwise be present in the final reconstruction. The loss in lateral precision occurs for two main reasons. Firstly, temporal filtration simply incurs a loss in the signal of the image, which leads to a reduced SNR. The resultant reduced curvature of the fitting function means that the molecules cannot be localised to the same precision following HAWK pre-processing. Secondly, laterally adjacent emitters that switch states simultaneously (such that one turns off and one turns on) often cause a slight bias away from their mutual centre (section 3.3.3), which increases the spread of the localisations.

These constitute our current best theories that explain HAWK's lateral precision loss – but each will have completely different effects on different 3D modalities. For instance, the relative curvature of the fitting function between x and y – which we know to be altered by HAWK processing – slightly affects its position on the calibration curves used to localise molecule in z when using the astigmatic fitting method, that may results in a noticeable drop in localisation precision. When using the DHP, it was difficult to ascertain whether HAWK decreased axial localisation precision at higher densities. However, this was only because coupling the DHP with high activation densities was found to lead to such severe undersampling due to the resulting ambiguity in DHP lobe-pairing, that HAWK was entirely necessary for the DHP to be viable over shorter timescales (section 3.3.4). Lastly, biplane SMLM with a single objective was found to be entirely improved by HAWK at higher activation densities for both simulated and experimental data, greatly reducing mislocalisations in z due to artefacts with no obvious performance losses. This success was observed despite the fact that this kind of data suffers from having its signal shared between two cameras.

These results are highly promising, and are important for three main reasons. Firstly, there are currently very few attractive alternatives for performing 3D SMLM at the higher fluorophore activation densities required for live-cell imaging. Additionally, those that do still suffer from the same failure modes as conventional LD methods – they produce artefacts that can render a dataset useless, where HAWK only loses resolution. Secondly, artefacts that occur using the 3D modalities that we tested appear to be much more severe than artefacts that occur laterally. That is, they misplace molecules by a much more significant amount, they are harder to filter out, and are less visually obvious. This is with the exception, of course, of DHP datasets using reconstructed using *Easy-DHPSF*, for which the visual indicators of inaccuracy are quite clear, which brings us to the third major benefit of HAWK

in the context of 3D imaging: it is a pre-processing method. Ordinarily, to reconstruct a HD DHP dataset would require a brand new DHP reconstruction algorithm, because the best currently-available software can only 'handle' the lowest emitter density. Thus, HAWK allows researchers to benefit from the best available software and imaging modalities (for instance, the DHP has a very high localisation precision and functional z-range), without having to 're-invent the wheel', so to speak.

Chapter 4

AFFIRM3D

4.1 Introduction

4.1.1 Anisotropy of microscope resolution

3D data is essential for capturing cells in an environment that more closely resembles their natural physical surroundings. This is because imaging exclusively in 2D environments can potentially lead to misleading conclusions about the behaviour of cells in the body, since cells are mechanically sensitive entities whose signalling pathways are responsive to physical stimulus. However, the PSF is never as well-resolved along the axial direction as it is across the lateral axes. As we have seen, confocal and widefield microscopy, being diffraction-limited modalities, can optimally reach a lateral resolution of about 200 nm (section 1.2.2). Abbe's diffraction limit for lateral resolution in a widefield microscope is given by equation 1.3; however, the limit for its axial resolution is given by:

$$d_z = \frac{2\lambda\eta}{\mathrm{NA}^2} \tag{4.1}$$

Where η is the refractive index of the medium between the objective and the sample. With optimal & realistic values for NA and λ, this results in 2.5-3 times poorer axial than lateral resolution. In a confocal system, the pinhole aperture only allows light from a much shorter vertical distance to be imaged increasing the resolution by a factor of about 1.3 [6], however, relative to its lateral resolution, there is still a severe loss of resolution in z relative to xy. Though the resolution discrepancy between axes can be reduced in a confocal microscope by reducing the radius of the pinhole, this also reduces the amount of light available to form an image, meaning there is a limit to the resolution improvement that can be achieved using this amendment [8]. This discrepancy is also true of related modalities, like spinning-disk confocal microscopy, which uses a spinning array of pinhole apertures to reduce the acquisition time for

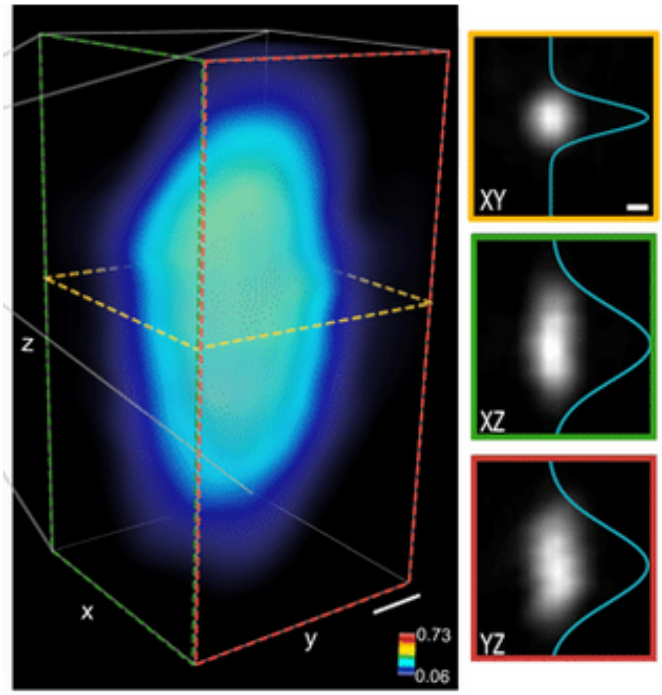


Figure. 4.1. The PSF is much more extended in the axial direction than the lateral directions. Shown is an image of a PSF from *Deep learning enables reference-free isotropic super-resolution for volumetric fluorescence microscopy* [106]. A volumetric image of the PSF is shown on the left, with the colour scale indicating relative imaged signal intensity. This is more extensive along the axial direction – this is also shown by the projected PSF images to the right, wherein the projection down z (top image) has a much thinner intensity profile than the projections down x and y (bottom two images).

the samples. Even in 3D SMLM (section 3.1), whereby the axial location of a molecule mathematically inferred from the PSF dimensions, the axial localisation precision is limited to around half (often less) of the lateral localisation precision in most experimental cases.

One microscopic approach is well-known for its ability to achieve enormous improvements in its axial resolution: interferometric microscopy (IM). Generally this method obtains axial information by splitting the excitatory beam into two before imaging. The beams are recombined after one has excited the sample, and the phase differences in the two beams are used to obtain super-resolved axial information. There are several ways in which this concept can be implemented into standard microscopy systems. In a 4π interferometric microscope [137], the sample is coherently illuminated from both sides in a confocal configuration. Interference of the opposing wavefronts sharpens the focus, dramatically improving the axial resolution. Similarly, another well-known iteration, I^5M [138], places two objectives at opposite sides of the sample to collect the sample information in widefield. The emitted signals are coherently recombined, and the illuminating light (which

has also been split to illuminate the sample in epi-fluorescence from both sides of the sample) creates an interference fringe at the sample to yield further structural information. Interferometry has even been employed in the SMLM imaging context in iPALM [88] in a similar fashion, wherein the light captured two opposing objectives interferes in a three-way beam splitter. Here, the beams are recombined and interfere with each other and are delivered to three CCD cameras at a 120° phase offset. In each of these methods, the axial resolution is superior even to the lateral resolution, because the precision with which the phase differences between multiple beams can be resolved is better than the Rayleigh criterion for the PSF. This results in sub-100 nm axial resolutions for $4\pi M$ and I^5M, and sub-10 nm axial resolution in iPALM [88, 137, 138].

Despite the many impressive improvements to 3D imaging achieved by IM, it suffers from major accessibility shortcomings. Interferometric systems have complex instrumentation and require a high level of technical expertise to build and operate, which is a barrier for many researchers. Imaging the sample from opposing sides can be tricky to implement, requiring specialised sample holders that can themselves be much more difficult to use than in standard FM microscopes that only need a single objective. The equipment can be expensive, and its use cases are quite specific, and thus fewer than for other kinds of microscopy. This means that biologists looking to conduct IM experiments will likely face many obstacles when attempting to do so: do they have access to an IM system in their laboratory, do they know anyone who does? Is their department, company, or university willing to invest in such a system, given it will not likely be used all that often? If they can get access to another researcher's IM system, are they able to perform the experiments themselves, or do they require constant supervision? Taken together, these considerations constitute serious barriers to making IM commonplace in biological imaging research, particularly if the researcher is looking to gather large volumes of data. Because of this, there remains a major demand in 3D FM for a technology that can achieve an augmentation in image z-resolution to achieve resolution isotropy. To overcome the accessibility problems of IM, such a method would have to be cheap and accessible – in this regard, the most welcome solution to resolution anisotropy would be an open-source image processing algorithm.

4.1.2 Convolutional neural networks in biological imaging

CNNs have rapidly become popular and widely used tools in biological imaging, finding applications across the whole field of biological and biomedical research [139, 140] From the beginning of the field, CNNs were widely used in segmentation and

classification, an area where there was often access to a substantial amount of GT data, and algorithms could automate and accelerate tasks that previously had to be carried out by hand [141]. As discussed in section 1.7.4, CNNs come with several distinct advantages for handling image processing tasks: the reduced connectivity between layers reduces the computational overhead required for training, but convolutional feature abstraction means that this generally does not come at great cost to performance when processing images. These features make CNNs the 'go-to' architectural basis for image processing tasks. CNNs are also *translation-equivariant*, meaning that a trained CNN has no particular difficulty in classifying an image following a lateral shift (that would cause the object's identifying features to shift position). This is because convolution involves scanning a kernel across a whole image (usually with padding).

However, it should be noted that convolution is not natively equivariant to all kinds of common transform – typically, they are notoriously sensitive to input rotation. This is but one of many limitations of CNNs that has only been overcome with recent developments in the field. Natively, a CNN that is perfectly capable of performing common computer vision tasks such as image classification can be easily dumbfounded by images containing objects that have been rotated relative to their presentation in the training dataset. Implementing rotation equivariance into CNNs thus becomes a highly valuable addition to the image processing field. This has been achieved by using steerable filters, which is a set of rotated kernels that have been derived from a base filter The image can be passed through this filter to produce a feature map that is invariant to the orientation of the input [142]. These filters' rotational symmetry allows multiple rotations of the filter to be employed at low computational cost due to the reduced number of calculations that need to be made.

There are other shortcomings to using CNNs, but fortunately, due to the overwhelming popularity and utility of the technology, issues are constantly diagnosed and new solutions produced. For instance, when used in classification tasks, CNNs are also highly sensitive to *adversarial noise*, wherein the network becomes highly responsive to particular noise patterns, causing it to misclassify images with high confidence. The *Do Not Disturb Network* (DNDNet) [143] exploits the fact that use of adversarial noise to confuse the network requires knowledge of its parameters. Thus, DNDNet introduces a transform to the input before it is passed through the network that effectively conceals this information, rendering the adversarial noise ineffective.

An example of a popular and successful application has been the detection of

starconvex polygons, which are an excellent model for nuclei [144, 145]. Nucleus detection is an important target in bioimage analysis as it allows cell counting and the quantification of nuclear geometry, and also provides an excellent starting point for cell segmentation, as nuclei are more reliably visually separated, and of more standard shapes, than cells. The addition of further contextual information may be able to boost the accuracy of detectors [146], though this brings up a point that we will return to later: the merits of relatively simple and interpretable algorithms, versus those that use more contextual information but are more opaque.

Aside from segmentation and classification, a second popular area of development in deep learning for bioimage analysis is denoising [147]. Denoising methods have been developed which do not learn specific data of the sample, meaning that algorithms are relatively likely to be unbiased [148, 149]. Success can easily be quantified as an improved SNR, and due to the improvement in perceived quality of the image, and their use as a tool to improve the accuracy of subsequent image processing techniques, these approaches have become highly successful.

By training on specific types of image deep learning networks can be trained to transform between different imaging modalities, for example transmitted light to fluorescence [150], or confocal to STED [105, 151] It should however be noted that these methods are sensitive to the specifics of the labelling and cell structure, and would need to be retrained for each specific case. For already complex image processing tasks, deep learning offers a potential method for speeding them up and achieving better reconstructions, such as in SIM where higher speed [152] and better performance have been demonstrated [153].

There has been a strong push in recent years to make deep learning more accessible to users unfamiliar with PyTorch or TensorFlow, including developing a system for use with ImageJ [154] and google coLab [155], though it should be noted that both may be challenging from the point of view of memory usage (in the case of ImageJ, issues with memory handling may mean it is not possible to analyse larger datasets, while coLab requires a paid subscription for the amount of memory required for many bioimage applications. As these methods become more popular, the challenges involved in using them have come to the forefront of the field, in particular the reliability of these approaches [156] and the lack of GT data in many application areas, affecting both training and testing [157, 158].

4.1.3 Overcoming resolution anisotropy with generative NNs

Broadly, the idea espoused by complex image restoration algorithms is the swelling of the 'design space' – if we think of the various fluorescence microscopy available as trade-offs between temporal resolution, spatial resolution, SNR, and hardware complexity, then post-processing techniques can allow such trade-offs to be more forgiving. In our case, we seek to restore spatial resolution without any increases in hardware complexity. This kind of task – in which information is added to an image using an assumption predicated on complex visual processing – is emblematic of the kinds of tasks that are performed by *neural networks*. Prior to the 'deep learning revolution' such tasks may have been performed by writing functions to perform image transforms from physical principles, utilising such information as the microscope's PSF, the function representing noise distribution, or the SNR, to restore content to an image. The problem with such approaches is that the mathematical assumptions they make about the data they are designed to process are assumed to be universal across the whole dataset, when in fact – in the case of microscopy images – there can sometimes be too much statistical complexity and uncertainty to reliably paramaterise an image restoration algorithm from first principles [159].

Of course, the use of NNs to achieve quality restoration for microscope images is not a new idea – the last decade has seen a blossoming of such methods. The Content-Aware image Restoration (CARE) and Residual Channel Attention Network (RCAN) algorithms restore image SNR when this has been lost to the lower illumination power, which may be required to perform experiments on living cells or avoid fluorophore bleaching [103, 104]. Denoising is a powerful application of deep learning because the GT can be obtained from real data to generate a cost function by simply acquiring images twice, one at high SNR and one at low SNR. The effectiveness of CARE in removing noise to improve image quality when compared with nonlocal means (NLM) noise removal, a non-deep learning approach, is displayed in Figure 4.2. However, these generally employ a pixel-to-pixel comparison of the input image and the GT, which will ultimately encourage the NN to reproduce its training examples by hallucinating structure when the trained network is then applied to process unfamiliar datasets. Applications that involve restoring spatial resolution to an image can also ultilise real data if done soundly: ANNA-PALM [71] creates SMLM images from undersampled data, meaning its cost function can simply compare a full reconstruction to one generated from a subset of the same data. A notable shortcoming of ANNA-PALM is its dependence on an undersampled

dataset, requiring a sparse activation density in the acquisition i.e. a low duty cycle fluorophore, which are relatively uncommon [160]. NNs have also been used to transform a diffraction-limited image into a super-resolution, wherein the cost function compares the input experimental image to the same region of interest captured on a higher-NA objective, or by comparing confocal images to their STED counterparts [105]. Deep-STORM is a SMLM algorithm that have been shown to outperform localisation algorithms that are written from first principles, and is especially notable for its ability to handle quite high activation densities and fast processing times [86, 111]. However, this kind of training can be more problematic still: not only is its training loop likely to hallucinate structure due to its direct comparison of the desired and input images, but it additionally requires a GT image that can only be obtained by simulation. Even though a primary advantage of NNs is that they can be trained to restore data while acknowledging the full statistical complexity of the image, the irony of exclusively simulation-trained ones is that they will underestimate the complexity of real data, which, for a generative network, can result in hallucinatory structure. In spite of this, a frequent difficulty for deep learning networks that do not employ simulations is the requisite acquisition of large volumes of varied experimental data for training. For this reason, hybrid learning approaches such as those used by DECODE [87] are increasingly used to obtain 'best of both' results by coupling a synthetic data training phase with an experimental data training phase. The former allows large volumes of data to be used for training and allows access to the ground truth structure for high-precision loss measurements, while the latter enables the network to learn from the increased statistical complexity of real data. In DECODE's experimental data training phase, the network is challenged to recreate the input image using its molecule location predictions. It out-performed all other tested localisation algorithms on benchmarked data [87, 112] by using the temporal context of the frame being reconstructed, incorporating the uncertainty of its own predictions to the training structure. The success of DECODE speaks to the value of obtaining training data from multiple sources.

An interesting point to note here is that in general, despite the many advantages of NNs over handcrafted algorithms, the latter are built from first principles, meaning they are likely to work similarly on datasets of different structures acquired on different machines. Conversely, NNs are often 'data-centric', meaning that an NN that restores content faithfully to an image of one structure might well hallucinate non-existent or misplaced entities in another, depending on the variety of its training data. Such hallucinations often present as high-frequency information and often only reveal themselves under intense scrutiny, if at all. There are two things

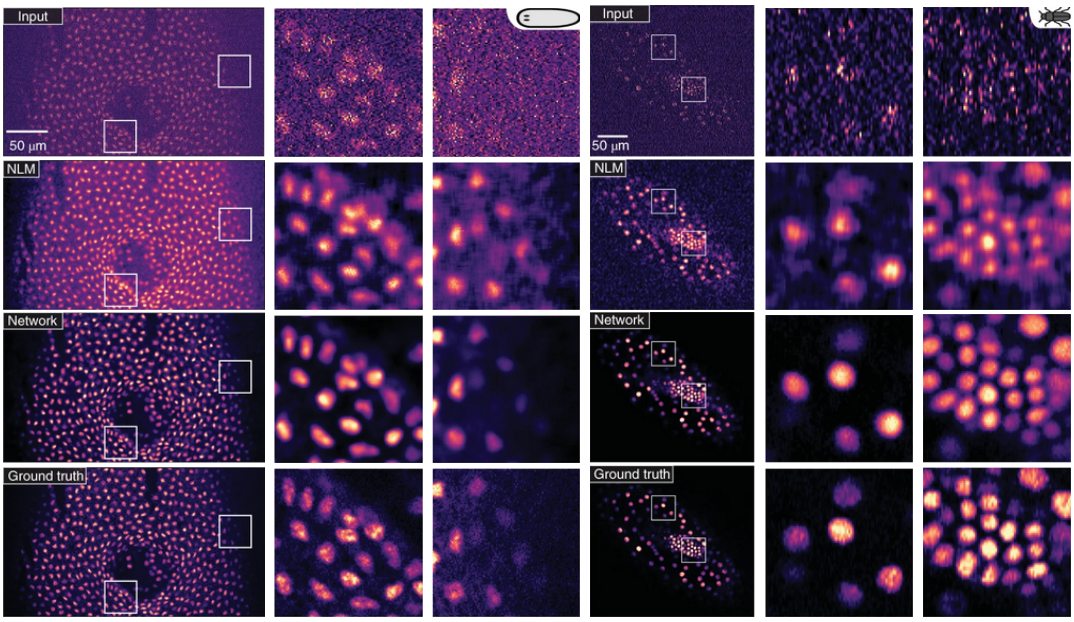


Figure. 4.2. Data from the CARE paper demonstrates effective image improvement using deep learning. Input data and restorations for a nucleus-stained (RedDot1) flatworm (S. mediterranea, left half) and a nucleus-labeled (EFA::nGFP) red flour beetle (Tribolium castaneum, right half). Shown are a single image-plane of a raw input stack (top row), the output of NLM denoising22 (second row), the network prediction (third row), and the high-SNR GT (bottom row).

that contribute to this problem – firstly, inadequate training data, which is very frequently a problem due to the difficulty of acquiring large and varied datasets for training. Secondly, the structure of the training loop can give rise to hallucinations: a prime example in the content restoration context is the use of GANs, whose explicit training motive is to convince a discriminative network of the truthfulness of their output, using persuasive forgeries of images that contain extra information, whether that information is an accurate representation of reality or not [159]. Discriminative networks, while powerful, are limited in the scope of their scrutiny, and will thus often fail to perceive when this extra information does not successfully capture the GT of the structure. Ultimately, this means that the generator is then incentivised to increase the level of hallucination without penalty.

A post-processing algorithm that restores z-resolution to an image requires that we add more information to the image than is natively present, with no reference to an additional source of information. To achieve this requires that the algorithm make an *inference* about the structure being imaged that is 'missing' from the image along the z-axis; it thus requires a *base assumption* about this structure from which to make such an inference. One such assumption, that will likely hold true in many biological scenarios, is that the structures in the region of interest are *themselves*

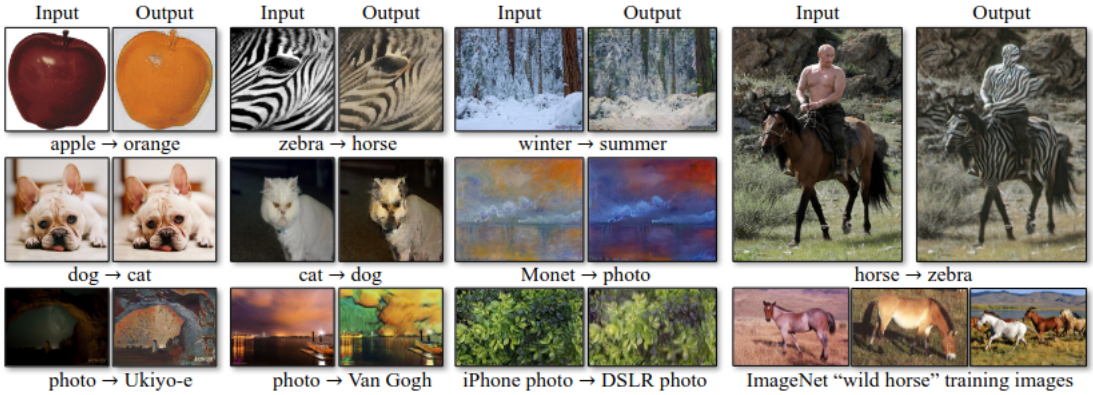


Figure. 4.3. Typical failure modes of generative networks. Displayed here are some of the failed outputs from *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks* [161] Due to a lack of diversity in the training data, the network is not able to render the image translations as expected. The mappings learned by the networks were insufficiently flexible to render the desired outputs.

isotropic. We might consider, for instance, a cell embedded in a 3D matrix. Here, the cell itself makes no distinction between x, y, and z – no direction is 'special', yet the axial direction presents differently when the cell is imaged. Similarly, many organelles will exhibit impartial treatment of the axial direction, regardless of the configuration of their cell. The nuclear pore complex, for instance, sits on the cell's nuclear membrane, which is approximately spherical during interphase, regardless of whether the cell is embedded in a 3D matrix, or struck to a coverslip. There is, of course, one factor in nature that does create a preferential direction in living systems – gravity. However, though gravity certainly has an impact on whole living systems and their tissues at the macro scale [162], and on certain force-sensitive ion channels [163], in the above scenarios wherein a cell is being imaged *in vitro*, the effects of gravity on individual cells or organelles relative to the pressure exerted by the surrounding media, matrices, and tissues can be thought of as negligable. Assuming the lack of a preferential direction in the sample suggests an interesting design for a content-restoration NN cost function. That is, a NN can be trained to optimise an isotropic frequency distribution in a three dimensional image, under the assumption that the actual structure was similarly distributed along each axis x, y, and z. If done prudently, an NN could restore resolution to the z-axis of 3D images while making minimal biased assumptions about the sample structure.

It should be noted here that restoration of axial resolution to volumetric data using a neural network has been undertaken by a couple of notable papers [103, 106]. The architecture of CARE has been exploited to this end by restoring data to individual 'z-slices' from the acquisition. As with other applications of CARE,

they simply created matching pairs of images by using lateral slices as the GT, and modifying these to reduce the resolution to obtain the input dataset. The network was then trained to turn the input images into the GT. As with other applications of CARE, this training loop will encourage the NN to hallucinate structure when applied to novel data, as it will be trained without constraint to reproduce its training examples. A more recent paper from Park et al. [106] has proposed a neural network which explicitly proposes to achieve resolution isotropy in 3D images. It is a highly flexible and powerful approach that uses CycleGAN [161], a two-GAN setup that restores and then removes the increased resolution, binding the output of the first GAN to the original input to minimise structure hallucination. Here, a second GAN F is added to the training loop to perform the inverse function of the first, G, with the aim of transforming the output back into the input. The cost function in a cycle-consistent CycleGAN can be directly supplemented with a loss calculated from the difference between x and $G(F(x))$ to ensure that G does not distort the structures in x. A further innovation of the method proposed by Park et al. is its integration of the training loop into the use-case; that is, the network is used 'blind'. It is trained by processing the input image on a loop to perform gradient descent, and using the discriminator to distinguish between the output's lateral projections and the input's axial projections. The network then gets iteratively better at restoring resolution to the input image until isotropy is achieved. This has the advantage of requiring no reference image such as the GT. Additionally, it will not be biased towards producing the structures it has been trained on, as it is not trained before use. However, there are three problems with this approach. Firstly, the computational demands of such a system on the user are not tenable for many scientists: the developers of this system required 24 GB of memory on an *Nvidia GeForce RTX 3090* GPU. These are expensive and can be prohibitively difficult to obtain – even with this exceptional GPU, running the algorithm on an experimental dataset took almost a full day [106] since training and use are integrated into the same process. This is a significant shortcoming given that a major advantage of a post-processing step to augment axial resolution would be an accessibility increase when compared with hardware-based methods like IM. Secondly, the discriminators in the CycleGAN employed by this technology are not permitted to look at the image in full; rather, they assess only slices or axial projections of the image they are processing. As we have seen, one reason GANs are particularly prone to hallucinating structure because their training motive is to fool the discriminator network. In this case, this means that it only needs to produce images whose axial projections' apparent resolutions are indistinguishable

to the discriminator, but this does not necessarily mean that the generator's output has no hallucinatory features. Lastly, it seems possible that the method's unique feature of training itself on a single image increases the possibility that F and G learn to be increasingly perfect inverse mappings of one-another. This means that G could introduce small structural artefacts into the output image and, so long as the discriminator networks did not recognise it as erroneous, F could simply remove them – satisfying cycle-consistency, while still producing hallucinated structure.

In summary, generative NNs are powerful tools for image processing, but not without their flaws. GANs [110] are a particularly powerful iteration of this process; equally, they are given to particularly problematic failure modes. The example in section 1.7.3 allows us to appreciate the simplicity of a basic GAN training loop. GAN training is typically *unsupervised*, meaning that the dataset does not need to be labelled prior to training. Labelling allows data to be associated with categories to facilitate NN training, but generally requires datasets to be labelled individually by a human. This can be a tedious and enormously labour-intensive process, and is not required in a GAN training loop. However, in most cases, this means that GANs can become very difficult to train because they possess unique failure modes relating to their generalised approach. Chief among these are *mode collapse*, in which the generator learns to fool the discriminator by reproducing a small set of convincing counterfeit images, and *hallucination* (section 4.1.3), where it does so by generating forgeries that do not represent the ground truth. Both failure modes will result in a generated structure that is not faithful to the original – in other words, to train a GAN that outputs *some* isotropic data is quite simple, but to create isotropic data that faithfully captures the structures in the input is difficult. Simply put, GAN outputs are designed to *look* correct, not to *be* correct.

A way around these problems is to use a *hybrid GAN* (hGAN). This simply denotes a GAN whose cost function comprises elements other than the result of the minimax game with the discriminator (the generator's 'adversarial loss'). The additional components to the total loss constrains the generator output avoid certain failure modes. It is a particularly useful technique when the aim of the generator is to transform one type of data into another. For example, by supplementing the minimax loss with a direct loss like L1 or L2 loss, we bind the generator output to the structure of its input, forcing it to modify the input image (instead of, for instance, repeatedly creating the same set of counterfeit images regardless of input). Additionally, the adversarial loss is problematic because it obscures the learning process from the researcher. That is, its only objective is to beat the adversary, and we don't know what criteria the adversary is employing to distinguish between real and

generated data – the networks in this context have extremely poor interpretability. In fact, discriminative CNNs are notorious for being easily fooled into misclassifying images if structured noise is added to them, even if the noisy pictures may look almost identical to a human, though there have been successful attempts to amend this issue [164], though filters have been developed to improve their robustness in this context [143]. However, because the GAN architecture is so effective, it is often useful to retain the adversarial loss as a component of the overall loss, while using more direct measures of the generator's performance to force gradient descent into better local minima.

4.1.4 Use of Fourier-space information

As we have seen, the training of NNs must be approached carefully, because NNs that have been trained to optimise particular structures tend to hallucinate those structures. With this in mind, we sought to enhance the resolution of 3D images using *general* properties of the image, avoiding as much as possible a training procedure that leans heavily on structure-specific properties of the training data. To this end, we designed a network training loop that prioritised the equalising of axial frequency distributions, instead of only using the real-space distribution of signal in 3D space. It was our hope that this would lead to a generative network that could restore isotropic resolution to anisotropic 3D microscope images by re-sampling the pixels and signal in z.

We predicted that use of Fourier-space information in our cost function would have the benefit of removing some of the hallucinatory behaviours exhibited by machine learning algorithms. This is a major advantage in the context of restoring resolution to microscope images, as it reduces the network's propensity for generating false data when confronted with novel structures. An NN that can only process structures that resemble those it has been trained on is not especially useful and certainly not future-proof. Of course, this bias-resistance comes at a price: the frequency domain simply does not contain nearly as much structure-specific information as the real-space domain. That is, frequency-space information alone cannot be used to train a network to produce perfect microscope images. We want our generator to 'equalise' the frequency distributions along x, y, and z. Without real-space information in its cost function, it has no 'incentive' not to produce, for example, images that have lower xy-resolution, instead of images with higher z-resolution.

4.1.5 Choice of training data

One of the most important considerations for any NN is the training data, and whether to train with *simulated* or *experimental* data, or a mix of both. This must be considered before the NN architecture has even been developed, as it will significantly impact the structure of the training loop. There are three primary benefits to using simulations to train an NN:

1. For restoring or adding information, it can be extremely useful to have access to the GT, because it provides the perfect exemplar image from which to derive the loss for each training example. Of course, we only have access to the GT when training using simulated data. As we will see in section 4.2, this benefit is essential to consider due to its significant impact on the structure of the training loop.

2. NNs sometimes require input data to have a discrete dimensionality, meaning each input image must the same shape. This requirement is much easier to satisfy using simulated data, as the data can simply be created with consistent dimensions, as opposed to needing pre-processing. This kind of pre-processing – such as cropping and re-sampling – can increase the computational requirements of training, as well as reduce the resemblace of experimental data to its native appearance.

3. NNs can require enormous volumes of data to be trained effectively, especially ones that are performing complex tasks. Our NN is, of course, expected to be used on large 3D images, meaning it has many weights and biases to optimise. It goes without saying that simulating hundreds or thousands of images for training is far easier than data scraping, that is, finding, downloading, and pre-processing them. For this reason, even NNs that are designed to be trained on experimental data may also have their training supplemented with simulated data.

However, simulating training data also comes with a major disadvantage; namely, that the ultimate goal for the network is to process experimental data, despite the fact that it has never seen any before. Simulated images do not contain the level of detail, complexity, or variability of real structures, as they are generated from a limited number of parameters. They are ultimately not as representative of the data that the NN will be used on. We thus sought ultimately to develop a training loop that had no GT requirements, so that its training could be supplemented with real experimental data. To this end, we aimed to develop a 'transfer learning' structure

to AFFIRM3D. In this set-up, the GAN is pre-trained on simulated data to provide a basic set of parameters that are configured to the task of restoring axial resolution, before subsequent training can be conducted on real data. This allows the benefits of training with both data types to be incorporated into a single algorithm.

Thus, the goals of AFFIRM3D are to:

- restore z-resolution to anisotropically resolved & anisotropically sampled 3D images
- minimise reliance on the GT in the training loop to reduce bias & structure hallucination by using a Fourier spectrum of the axial projections of the input image to generate the largest component of the loss

4.2 Materials & Methods

The code used in this study is available at `https://github.com/squishram/zres_GAN`.

4.2.1 Engineering an appropriate cost function

The aim of our network is to re-sample the z-axis of a 3D image while restoring axial resolution, all whilst incurring the minimum structural bias that can be attained without compromising structure. This is both a complex task and a difficult constraint to circumnavigate, meaning that the cost function needs to be considered carefully. The cost function is the heart and soul of a training protocol, because it asks the question 'what will the network be penalised for, and how much?'. As we saw in sections 4.1.3 and 4.1.4, simply relying on the adversarial loss would result in several shortcomings. For this reason, we supplemented the adversarial loss with two other cost functions: either the mean absolute ('L1') loss or the mean squared ('L2') loss, and the axial Fourier-space ('FZ') loss.

The L1 and L2 losses are simply a direct comparison between the NN's output and its *desired* output. The L1 loss is simply the mean absolute sum of pixel differences, and the L2 loss is the mean squared sum of, and they are calculated as follows:

$$L1 = \frac{1}{N} \sum_{i,j,k} |O_{ijk} - \hat{O}_{ijk}| \tag{4.2}$$

$$L2 = \frac{1}{N} \sum_{i,j,k} (O_{ijk} - \hat{O}_{ijk})^2 \tag{4.3}$$

Where N is the total number of pixels in the image, i, j, and k are the pixel indices in x, y, and z respectively. In our configuration, the actual output is O and the desired output is $\hat{O}$. The L1 loss is instrumental in establishing a desired structure for O. This is in part because of the potential of *mode collapse* in GANs, wherein the generator is able to fool the discriminator by producing a small and limited set of convincing fakes – minimising the loss while failing to perform the actual task of restoring resolution to the input image. L1 loss will reduce the propensity of the network to mode collapse, and improve the quality of O, by binding the structure of the output directly to the input. However, we had to be prudent when assigning a relative weighting the L1 loss: if it is too heavily weighted, this will likely result in a biased network that merely reproduces the examples it has been trained on. To counterbalance the L1 and adversarial losses and their propensity for biasing the network, we supplemented them with the FZ loss.

4.2.2 The axial Fourier-space loss

The FZ loss is calculated by comparing the lateral distribution of frequencies in the input image to the axial distribution of frequences in the generator output. The equation describing its calculation is as follows:

$$\frac{1}{M}\sum |\mathcal{F}(I_x) - \mathcal{F}(O_z)| + |\mathcal{F}(I_y) - \mathcal{F}(O_z)| \tag{4.4}$$

$$\begin{aligned} I_x &= w(\sum_y \sum_z I(x,y,z)) \\ I_y &= w(\sum_x \sum_z I(x,y,z)) \\ O_z &= w(\sum_x \sum_y O(x,y,z)) \end{aligned} \tag{4.5}$$

$\mathcal{F}(f(x))$ is the fast Fourier transform (FFT) of $f(x)$ that has been processed through a high-pass gaussian filter. This ensures that the loss is only calculated from the high-frequency signal, representing the 'finer' or sharper structures, while ignoring any low-frequency noise or background signal. M is the number of samples in the FFT'd signal, and I_d and O_d are the signal distributions along dimension d of the input and output images, respectively. These distributions are calculated by projecting the 3D image down the other two dimensions as shown in Equation 4.5. It is important to note that the signal projections of the images are passed through a windowing function, w, before being passed to the FFT, which is essential to remove ringing artefacts from the FFT'd signal. Simply put, 3D image data is stored in a

cuboidal 'grid' arrangement, the edges of which are not continuous – but the FFT treats such edges as though they were abrupt discontinuities. Windowing functions mitigate the artefacts generated by this issue by smoothly tapering the signal to zero at the edges, reducing discontinuities in the signal and thus reducing the amount of ringing in the FFT output signal.

The assumption built into the FZ loss calculation is that 'for cells embedded in a 3D matrix, the high-frequency distributions along all three axes will be roughly equal' – alternatively, 'when the force of gravity is negligably small, nature does not have a preferred direction'. Its inclusion in the training process comes with two major benefits:

1. As discussed in section 4.2.1, it contributes to training the generator to increase the resolution of the input image *without* binding it to the specific structures present in the training dataset. This means it is expected reduce the likelihood of structure hallucination or otherwise biased output images when the network is presented with novel data.

2. The FZ loss is unique in that, regardless of the specifics of network architecture, it is *never* calculated by comparing to the desired output – rather, it uses information from the input image to train the network. It has no use for a GT image, even if one can be provided. This feature of the FZ loss is what ultimately led us to develop a full training protocol without the GT image, which opens up the possibility of training the network on experimental data.

4.2.3 Handling anisotropically sampled images

A discriminator network in an image processing GAN will often conventionally employ a CNN architecture to condense the image down to a single number in the range $[0, 1]$ signifying the probability that the input is 'real'. This configuration suffers from brittleness with respect to its input, in that the input image will need to be of a discrete dimensionality in order for the network to condense it down perfectly to a single number. In many applications this is fine, as input images can simply be resampled before being passed through the network. However, in microscopy, we wish to avoid this as much as possible, since resampling to meet the dimensionality requirements of the network will either reduce the available information in the input if the image is too large, or interpolate information that isn't present if it is too small. Additionally, almost all 3D micrographs are sampled differently in the axial direction, but the sampling ratio between the axial and lateral axes of the image will vary tremendously from dataset to dataset. Aside from the lower z-resolution, there

is most often less data in z: it is sampled over a smaller distance, and often, the pixels are longer in the axial dimension, limiting the available spatial information.

To overcome the variability in the input sampling, we employed a PatchGAN architecture for AFFIRM3D's discriminator network. This simply means that the discriminator only assesses the 'realness' or 'fakeness' of local patches in its input images, instead of the whole image. In other words, instead of condensing each image down to a single number in the range $[0, 1]$, it is condensed down to several, each representing a patch of the image. Each patch is evaluated separately, and the mean assessment for each patch serving as the number to pass to the GAN loss function (equation 1.9). This was found to have no detrimental effects on the effectiveness of the GAN by Isola et al. and was even associated with a modest increase in the quality of the generator's images when the patches were roughly one-quarter (70×70 pixels) of the full image size (286×286 pixels) [165]. This means that the discriminator will be able to produce an assessment if the training data has significantly reduced sampling in z, allowing us to train AFFIRM3D on data from a variety of setups.

4.2.4 Training data

Simulated training data was made by generating simple simulations of flowing, curvilinear one-dimensional structures propagating in 3D space. This was achieved using a 'random walk' in 3D space, with the step size between each molecule defining the staining density, and each point serving as a single fluorophore-stained epitope. It allowed easy control over the number of molecules, allowing us to interrogate the effects of structure and staining density on training. If a 'walk' led past the edge of the image dimensions, it was simply re-initialised at a random location. The random walk system was designed to mimic long, one-dimensional, subcellular structures, which are typically composed of single molecules or molecule pairs that build iteratively on each other to fulfill different functions. The position of each 'molecule' was calculated according to the following equation:

$$\begin{aligned} p_i &= p_{i-1} + v_i \\ v_i &= R(\theta) \cdot R(\psi) \cdot v_{i-1} \end{aligned} \tag{4.6}$$

Where p_{i-1} is a molecule, v_i is the step taken from p_{i-1} to reach p_i, and $R(\theta)$ is the rotation matrix representing the rotation of a molecule in three dimensions by angle θ. To give an organic appearance to the structure, the direction of the vectors was constrained: the vector joining p_i to step p_{i+1} was limited in its permitted deviation

from the direction of the vector joining step p_{i-1} to step p_i. Equation 4.6 shows how step in the random walk v_{i+1} was generated by simply rotating the previous step v_i by θ through a plane containing itself (the polar angle), and then by ψ about the axis of v_i (the azimuthal angle). To achieve a smooth propagation of the fibres without sharp bends, θ was limited to the range $[0, \pi/20]$ for a step length of 0.5 pixels. These simulations were designed to mimic real cellular structures such as microtubules and actin, though ultimately we would want to use a wider variety of data, including simulations of structures that propagate in two dimensions, as well as real data.

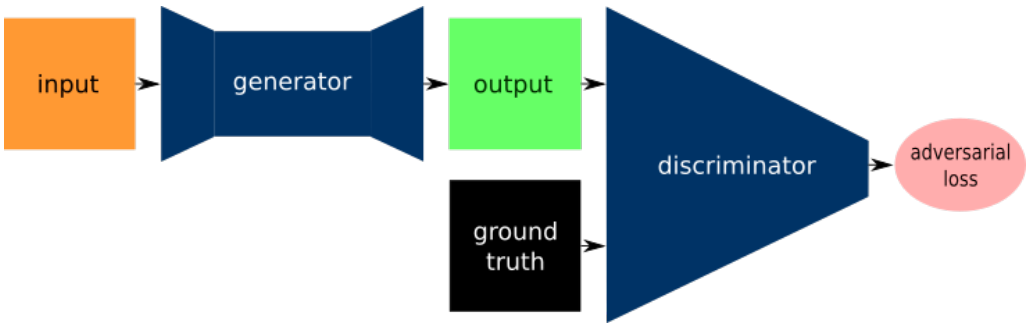
4.3 Results

4.3.1 AFFIRM3D v0.1

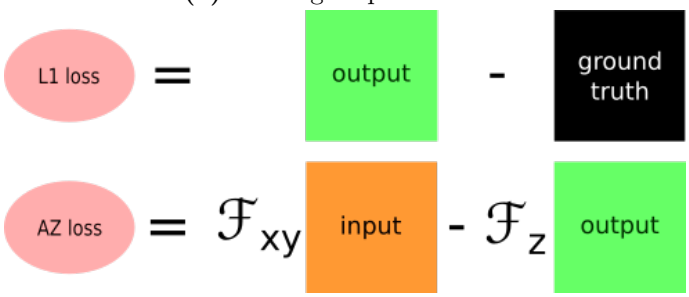
The structure of the training loop underwent several revisions over the course of AFFIRM3D's development. We would begin by developing an NN that could restore z-resolution in the most ideal case; once this was working, we would rework the training loop to make it more widely applicable. In total, we went through two training architectures before our network could be trained without the GT. In the first version of AFFIRM3D, we aimed only to restore z-resolution to isotropically sampled (i.e. cubic) images with poorer resolution in z than in xy. This version also has access to the GT as part of its training architecture, which was used to calculate the L1 and adversarial losses. A flow diagram of the training loop for this version of the network is exhibited in Figure 4.4.

AFFIRM3D v0.1 served the essential purpose of demonstrating a basic proof of principle. While it still included the use of two 'non-adversarial' loss components (the L1 loss and the FZ loss), its training loop was otherwise identical to the original GAN paper [110]. In this context, we were able to freely adjust the scaling factors of the three loss components. Importantly, this allowed us to see what share of the total loss could be safely allocated to the FZ component (which has the advantage of not using the GT or the 'real-space' structure of the image section 4.1.4) without breaking the likeness of the output to the GT.

The effects of AFFIRM3D processing on volumetric simulations of fibre-like structures is displayed in Figure 4.5. To visualise the data, 2D 'slices' were rendered, of either an xy-plane (Figure 4.5a-c) or an xz-plane (Figure 4.5d-f). Both renderings showcase this early version of AFFIRM3D's ability to restore resolution to isotropically sampled data with relatively poor z-resolution. This is evidenced by the fact that the GT image (Figure 4.5b, e) closely resembles the generated

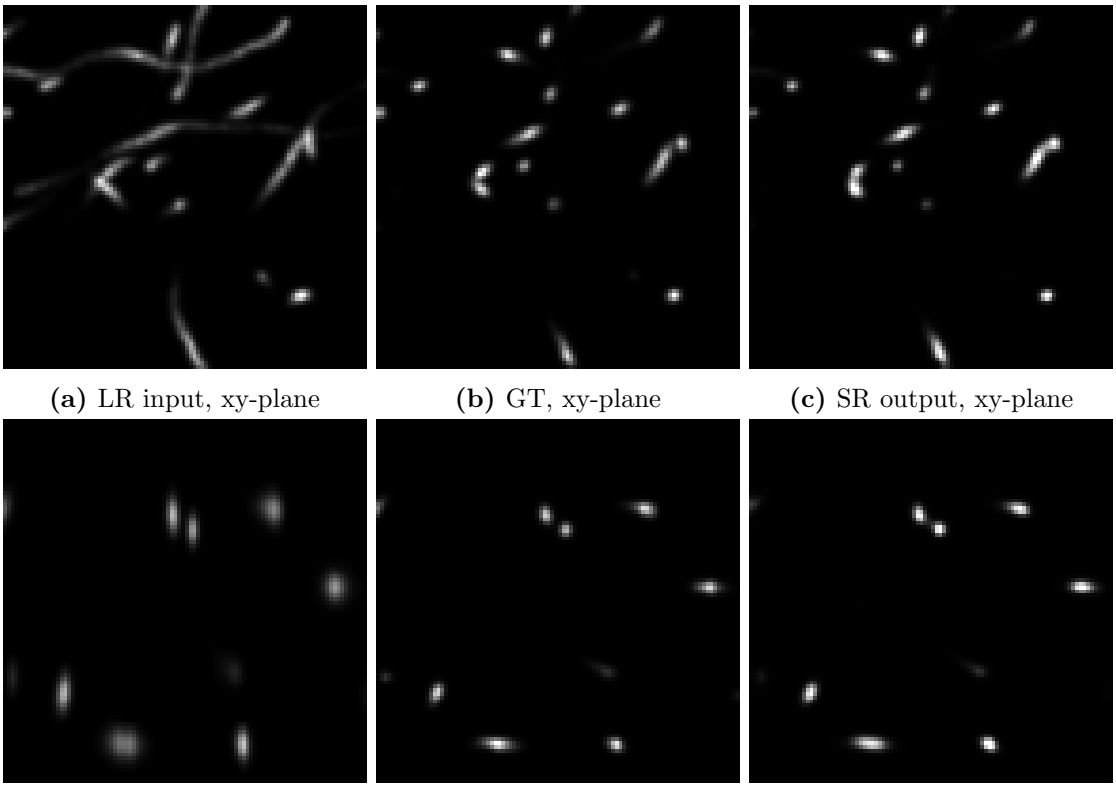


(a) training loop flow chart



(b) loss component calculations

Figure. 4.4. A diagram of the training loop for the first version of AFFIRM3D. (a) displays a flowchart of the training loop, including the obtaining of a value for the adversarial loss. (b) shows where the other two components of the loss for the generator are derived from. Here, $\mathcal{F}_{xy}$ denotes the mean value of the high-pass-filtered FFTs of the distribution of signal in x and y after they have been passed through a windowing function. $\mathcal{F}_z$ is analogous, but only for z (no mean value is taken).



(a) LR input, xy-plane **(b)** GT, xy-plane **(c)** SR output, xy-plane

(d) LR input, xz-plane **(e)** GT, xz-plane **(f)** SR output, xz-plane

Figure. 4.5. Output z-slices and y-slices of the initial version of AFFIRM3D after training on simulated data. (a-c) display the z-slices (xy-plane) and (d-f) display the y-slices (xz-plane). The leftmost column exhibits the slices from the input low z-resolution image with anisotropic resolution. The middle column shows slices from the GT image, which was simulated at the same time as the input low z-resolution image, but was created by convolving with a symmetrical Gaussian. After 10 epochs of training, AFFIRM3D takes the low z-resolution input and produces images whose slices are displayed in the rightmost column as output. The data was rendered with super-resolution pixels of 10 nm, 24 nm lateral FWHM and a 60 nm axial FWHM.

output (Figure 4.5c, f), but not the input image (Figure 4.5a, d). In the xy-slice (Figure 4.5a-c), reduced resolution is evidenced by several fibres crossing the input image that are not visible in the GT or output images. This is because the enlarged, anisotropic PSF 'bleeds through' into adjacent xy-slices in accordance with the lower axial resolution in the input. When looking at the xz-slice (Figure 4.5d-e), this enlargement is even more apparent, because round emitters in the GT and output images appear elliptical, indicating the poorer resolution in z.

4.3.2 Assessment of images generated by deep learning algorithms

The results of deep learning algorithms can of course be assessed using standard resolution assessment metrics such as decorrelation metrics [66]. However, these have substantial disadvantages, as has been covered previously, particularly the lack of a test against artificial sharpening. Ultimately, in the absense of the GT, there may always be limited ability to avoid the generation of sharpened results which do not reflect the sample, particularly in the cases of UNETs trained on more highly resolved structure than that in the data being processed which may be able to reproduce images which relate to the genuine structure of a biological sample, just not the one being imaged in that experiment [103]. However, it is a potential failure mode that would need to be always kept in mind whne using these algorithms.

In segmentation tasks designed to reproduce human performance there is a well defined and accessible GT which the algorithm is aiming to reproduce. In the case of resolution improvement this is not the case, and there is little data available where the same sample is imaged in multiple standard and super-resolution imaging modalities to allow performance of the algorithm to be verified. Acquiring such a dataset would require an international collaboration and a big time investment, but given the popularity of deep learning methods it would undoubtedly be of great value. The absense of any standard way of testing performance against GT in a wide variety of samples is a key limitation of the field. It should be noted that while it is possible to generate data, actual performance of microscopes tends to vary substantially and may be prone to patterns in noise because of, for example, the structure of the camera chip [166]. Simulated data is therefore unlikely to fully reflect the performance of any algorithm on experimental data.

An ideal metric would reward improved resolution and SNR; however, it would additionally reward GT structure being resolved in the reconstructed image when it is not in the original image. It would also penalise hallucinated structure and structure present in the GT which is missed in the reconstruction. Any metric employed should aim to quantify some combination of these properties. As is perhaps inevitable for an area which includes both developers with a microscopy background and those with a computer science background, metrics in use include ones standard to both areas. However, it should be noted that many image assessment metrics used in evaluation of computer vision applications of deep learning are aiming to assess the performance of an algorithm in creating an image *that cannot be distinguished as false by a human.* This is extremely different to the aim of enhancement of images for scientific studies, where the aim is to reproduce an image closer to the

underlying GT, and may lead to an incorrect assessment of performance.

4.3.3 Visual assessment by a human

Visual assessment by a human is somewhat unsatisfactory as it is difficult to reduce a perception of quality and likelihood of reflecting the underlying sample to a single number. However, can be valuable, particularly if paired with data designed to probe potential failure modes. Humans may also, for example, spot false structure which is left unassessed by being thresholded out of assessment, or spot that a particular rendered structure is biologically unlikely. Careful tests partnered with qualitative assessment by an experienced human are therefore a valuable assessment tool for most algorithms.

4.3.4 Line scan

Comparative line scans displayed in a figure are a standard assessment tool for the performance of biological microscopy. They have the advantage that they are visually clear and quantitative, but the disadvantage that the area of assessment is selected by a human, generally in an area of the image where the algorithm under consideration performs well. Such assessment is also generally unable to identify when the FWHM is narrower than the actual structure (as the actual structure is not known and narrower is generally taken to mean better performance).

4.3.5 Normalised root mean squared error

The normalised root mean squared error (NMRSE) is perhaps the simplest quantification metric that can be used to evaluate the similarity of two images. The NRMSE is defined as:

$$\text{NRMSE} = \frac{1}{\text{mean}} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{o} (A_{i,j,k} - B_{i,j,k})^2} \tag{4.7}$$

Where $y_{i,j,k}$ is the pixel value of the original image at position (i, j, k), $B_{i,j,k}$ is the pixel value of the compared image at position (i, j, k), and m and n are the number of rows and columns of pixels in the image. To normalise this metric to obtain the normalised root mean squared error it is standard to divide by the range of values present or the mean. We have chosen to divide by the mean, as it potentially limits the impact of two problems common in microscopy: areas of the image becoming saturated at the maximum possible intensity value, or individual

camera pixels having an uncharacteristically high noise level [167]. The problem of individually high camera pixels occurs in most sCMOS chips, and occurs because each CMOS pixel has an individual gain. It can only be avoided by substituting the averaged value of neighbouring pixels, which itself is undesirable as it means data is being removed in a way which is not transparent to the user. It should be noted that NRMSE is sensitive to the dynamic range of the images being compared. For example, if one image has a higher dynamic range than the other, the NRMSE score will be higher even if the visual difference between the two images is small. This can be a problem when comparing images that have different dynamic ranges, as the NRMSE score will not accurately reflect the visual difference between them.

Closely related is the Peak Signal-to-Noise Ratio (PSNR) metric, which compares the difference between two images by calculating the ratio of the peak signal power to the power of the noise. It is defined as:

$$\begin{aligned} \text{PSNR} &= 20 \cdot log_{10}(\frac{\max}{\sqrt{\text{MSE}}}) \\ \text{MSE} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{i,j,k} - B_{i,j,k})^2 \end{aligned} \tag{4.8}$$

Where max is the maximum possible pixel value and MSE is the Mean Squared Error between the two images. The advantage of PSNR is that it is easy to understand and interpret, but the disadvantage is that it is sensitive to the dynamic range of the images being compared and it does not take into account the structural similarity of the images.

4.3.6 Intersection over union

The *Jaccard index*, or intersection over union, is a method which allows evaluation of two binary masks. It is calculated as:

$$J = \frac{|A \cap B|}{|A \cup B|} \tag{4.9}$$

Where $|A \cap B|$ is the intersection between the two images, that is, those pixels in which both masks are white; $|A \cup B|$ is the union, that is, those pixels in which either mask is white. It is more generally used to evaluate the performance of classifiers. Other related metrics are the *receiver operator curve*, which quantifies performance at different threshold levels, and *precision recall*, which also extends the quantification to different scales. However, these latter two approaches both have the major challenge that a substantial proportion of the weight of the measure

can be in a regime where many more pixels/objects are being detected than actually exist.

4.3.7 Structural similarity index

Structural similarity index (SSIM) was developed as a method to predict the perceived quality of images, by quantifying the structural similarity between two different images. It is calculated for a pair of images, with the value being calculated per 'window' (i.e. patch), quantifying the similarity of the mean and standard deviation per patch. The SSIM is defined as:

$$\mathrm{SSIM}(A, B) = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)} \tag{4.10}$$

Where μ_A and μ_B are the average pixel value of the original and compared images, σ_A^2 and σ_B^2 are the variance of the pixel values of the original and compared images, σ_{AB} is the covariance of the pixel values of the original and compared images, c_1 and c_2 are constants used to stabilize the division in the case of weak denominators (typically derived from the dynamic range of each image). SSIM is based on the idea that adjacent pixel values have an interdependence relationship, that is, that nearby pixels affect each other's values, making the metric more structure-sensitive [168], an assumption that is absent from MRSE-related metrics.

The multi-scale structural similarity Index (MS-SSIM) is an extension of the SSIM metric that takes into account the structural similarity of an image at multiple scales [169]. It calculates the SSIM index at different scales, and then combines these results using a weighting factor to produce a final score. MS-SSIM is defined as:

$$\text{MS-SSIM}(A, B) = \left(\frac{2\mu_A\mu_B + c_1}{\mu_A^2 + \mu_B^2 + c_1}\right)^{\alpha} \left(\frac{2\sigma_{AB} + c_2}{\sigma_A^2 + \sigma_B^2 + c_2}\right)^{\beta}$$

$$\mathrm{MSSSIM}(A, B) = \prod_{k=1}^{K} \mathrm{SSIM}(A, B, k)^{\alpha_k} \tag{4.11}$$

Where $\mathrm{SSIM}(x, y, k)$ is the SSIM index calculated at scale k, K is the number of scales, and α_k is the weighting factor for each scale. The disadvantage of MS-SSIM is that it can be computationally expensive, as it requires the calculation of the SSIM index at multiple scales.

4.3.8 BRISQUE

BRISQUE (Blind/Reference-less Image Spatial Quality Evaluator) is a metric designed to quantify how visually convincing an image is to a human [170]. It is a no-reference image quality assessment method, which means it does not require a reference image to compare against. It assumes that the distribution of pixel intensities of 'natural' images is Gaussian, and a deviation from that is therefore a measure of distortion in the image. The mathematical expression for BRISQUE is:

To assess the distribution of intensities in an image, it first carries out normalisation, with a standard method being mean-subtracted contrast normalisation (MSCN). Four pairwise product images are also created, multiplying the base image with one translated one pixel up, right, up and right, and up and left respectively. These images are used to calculate a feature vector, with the first to elements being the output parameters from fitting the MSCN image to a generalised Gaussian distribution. An asymmetric generalised Gaussian distribution is then fitted to the four multiplied images, with each fit having four parameters.

The algorithm then needs to evaluate how close the image is to the expected intensity distribution. The model can be trained to compare against an existing dataset, so the image can be compared against particular types of image, using an approach such as a support vector machine, with pre-trained models being readily available for photograph-type images. This method has the substantial drawback that it in no way evaluates how correct the reconstructed image is, it simply evaluated how real it is likely to look to a human. This is a very different aim to producing an accurate image, and its use is an artefact of the development of the field (with most of the deep learning computer vision development focusing on generation of images which appear realistic to humans for applications in entertainment). That is, BRISQUE suffers as a micrograph evaluation technique from being trained on other types of images. BRISQUE was therefore not used during our evaluation.

In summary, the appropriate scenarios for the assessment of processed microscope images will depend on the specific goals of the image processing, as well as the type of images being considered. It's important to use a combination of quantitative and qualitative evaluation methods, and to choose the appropriate error metric based on the specific characteristics of the microscope images and the image processing techniques being used.

4.3.9 Performance of the AFFIRM3D algorithm without GT loss

In real imaging scenarios, GT isotropic data is unlikely to be available (or the isotropic imaging method would probably be used in the first place). The algorithm was therefore modified to remove the loss between the isotropic GT data and the generated data. In the new algorithm, an adversary makes xy vs xz and yx slices look the same. For the L2 loss, we have both a real space loss where we convolve the reconstructed image with an additional Gaussian in z predicted to take it to the original resolution. For the FZ loss, the data is projected onto each of the three axes, the power spectral density is calculated, and then the result is high-pass filtered (with a Gaussian filter) to remove the low frequencies, as these correspond to longer scale structure which we are not trying to optimise. To limit the impact of low frequencies, the log of these values is taken. The L2 error is then calculated between the x and z-axes, and the y and z-axes. The results from this analysis is displayed in Figure 4.6.

4.3.10 Optimisation

For the direct optimisation, two different approaches were tested, with the adversary being removed. The first, stochastic gradient descent [171], is a standard approach in deep learning. It is able to tackle problems with very challenging convergence, as the stochastic nature of the algorithm allows a good chance of getting close to the global minimum when trying to optimise over a complex phase space with many local minima. However, it should be noted that one of the main advantages of stochastic gradient descent is its ability to sample from batches, which does not need to be done in this case as the whole dataset is loaded into memory. Here we used the popular gradient descent method Adam [172].

Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) is a quasi-Newton method that uses an estimate of the inverse Hessian matrix [173]. It offers better convergence than stochastic gradient descent, but even the version designed to limit memory usage has much higher memory requirements. This means it is less commonly used in deep learning applications.

In both bases, an L2 norm was used, which is the correct choice for Gaussian noise as it is a maximum likelihood estimator, and Poisson noise is reasonably approximated by Gaussian noise.

In tests, AFFIRM3D reconstructed images of $128 \times 128 \times 128$ from lower resolution, anisotropically sampled data of size $43 \times 128 \times 128$ GT isotropic resolution

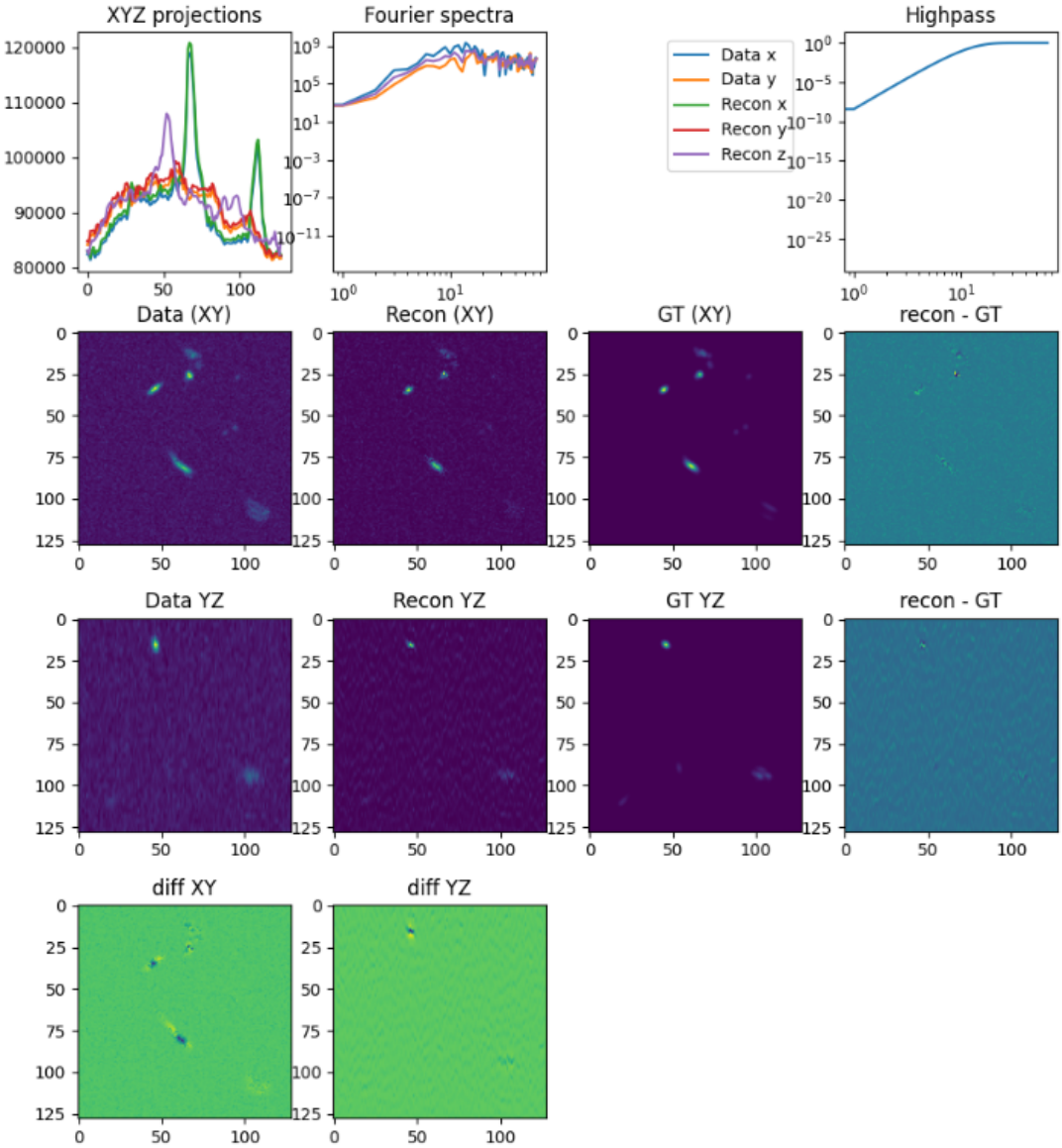


Figure. 4.6. Inputs and outputs of optimisation on assymmetrically sampled data. Right hand most images display differences between the reconstruction and the ground truth and bottom row displays the difference between the start of the optimisation process and the final reconstruction, indicating where in the image changes have occurred due to optimisation

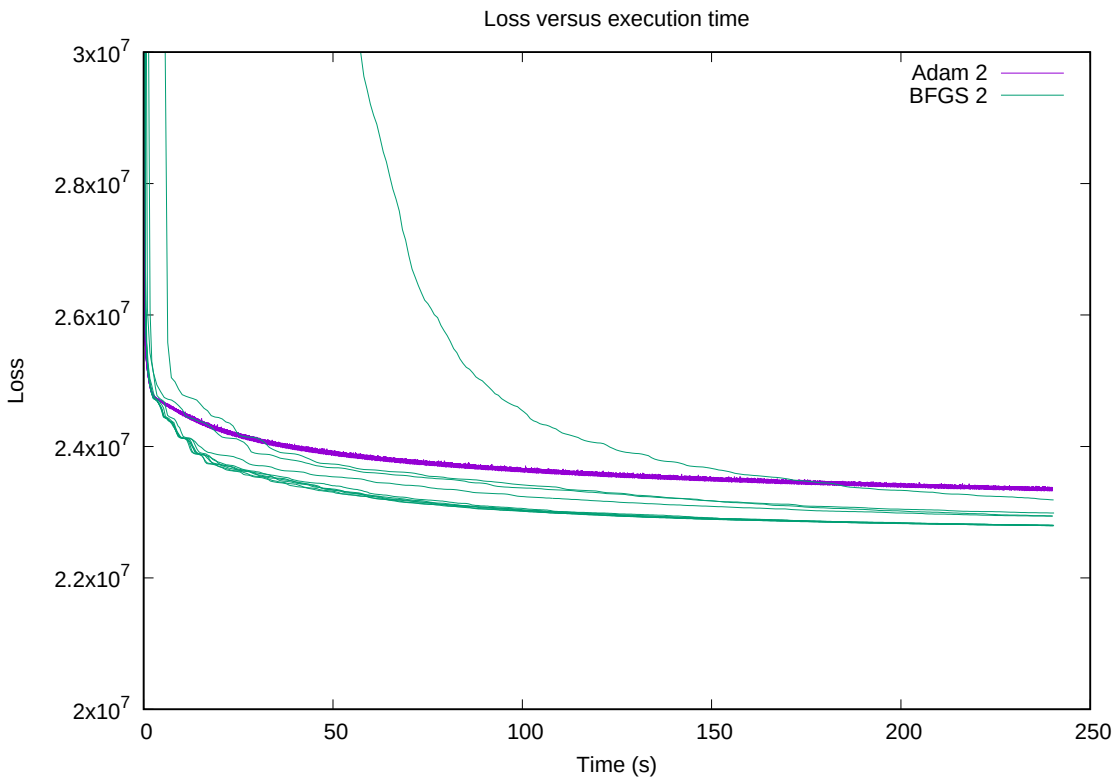


Figure. 4.7. Test of Adam vs LBFGS optimisers for AFFIRM3D. The data was gathered from 10 optimisation training loops consisting of four minutes of trianing each. Though the Adam optimiser shows a greater consistency in its rate of L2 loss reduction, the LBFGS optimiser consistently reaches a lower local minimum.

data of $128 \times 128 \times 128$ was also generated for visual comparison, and to obtain performance metrics. The FWHM of the PSF was $2.4\times$ the pixel size in xy (equivalent to 240 nm for 100 nm pixels). The FWHM of the PSF in z was 2.4 pixels in the isotropic resolution GT, and was 2 pixels in the low resolution data (600 nm if the data had been isotropically sampled at $128 \times 128 \times 128$, then downsampled by a factor of three). Figure 4.7 displays the L2 loss plotted against the training time.

Measure	Mean	Std dev
NRMSE	107.6775	11.4601
PSNR	27.8949	5.8646
Jaccard	0.751	0.096
MS-SSIM	0.3533	0.1230

Table 4.1: Table displaying errors for direct optimisation

Measure	Mean	Std dev
NRMSE	115.9251	13.2396
PSNR	27.2594	5.7456
Jaccard	0.6903	0.0748
MSSIM	0.3225	0.1116

Table 4.2: Table displaying errors for direct optimisation, with the addition of an adversary in the training loop

4.3.11 Performance of initial optimisation

The results from the performance of AFFIRM3D without an adversary are displayed in Table 4.1 The performance of the initial optimisation was evaluated with reference to a number of metrics standard in deep learning. All values were calculated per image plane, with the mean and standard deviation of those values then being calculated. It is interesting to note that the MS-SSIM index performed relatively poorly, whereas the Jaccard index score is fairly high. Examining the data, this seems to arise from mismatches in the noise, highlighting the potential problems with non-thresholded MS-SSIM measurements. However, with thresholded MS-SSIM, the evaluation ignores low intensity areas of the image, potentially failing to penalise the algorithm for creating low-intensity structure.

Surprisingly, the addition of the adversary slightly decreased the quality of the fit parameters, though on human examination the fit appeared marginally improved, highlighting the challenge of parametrising the complex issue of quality effectively. The metrics generated from AFFIRM3D when supplemented with an adversary are displayed in Table 4.4. Further investigation of the parameter space is required to discover if parameters can be identified which outperform direct optimisation.

4.4 Discussion

While AFFIRM3D in its current state constitutes a minimum working version, there remain additional features to implement and more comprehensive testing to undertake. AFFIRM3D offers a reference-free training loop, which makes training it on real experimental data an exciting possibility. Weights and biases can be calibrated (i.e. pre-trained) through training cycles on simulated data, before being applied to various experimental datasets. This regime allows access to the benefits of training on larger datasets without constraing the network, enabling it to deal with more statistically complex data.

The impact of the adversarial network on the current version of the algorithm is limited, and a substantial part of the impact is on the generation of high frequency

noise which does not contribute to the correction of sample structure. This could potentially be improved by weighting the cost function by a blurred version of the input image, upweighting those parts of the image that contain the most information. Alternatively, input images could be changed to a form which compresses low-information areas of the images, reducing their impact on the optimisation process [174]. This would have the advantage that it would reduce memory requirements, potentially allowing the approach to be used on larger image stacks and time series.

Further optimisation of the adversarial system is also necessary, potentially splitting into two adversaries. One would make the views in xy and xz/yz the same, while the other would try to ensure that the did not differ too much from the original.

In terms of evaluation of the output, there are a wide range of approaches currently standard in evaluating the performance of deep image-based image processing. The influx of these methods from an area of computer science and engineering with very different aims has perhaps led to widespread use of metrics that assess for human-perceived similarity rather than effectiveness in informing accurate scientific results. Several types of metric, such as RMSE, MS-SSIM and PSNR have the potential to be informative or misleading depending on precisely how they are used and the information that is needed from the rendered images. A careful consideration of the particular metrics most relevant for each use case will be important in the future.

Finally, it would of course be critical for testing to evaluate the performance of the algorithm on real data, as simulated data does not reproduce the likely variability in pixel response, and the spatial heterogeneity in the point spread function, likely to be present in real data. An ideal test dataset would be one in which both confocal data and iso-STED data were available, and we aim to partner with another lab in the future to gain access to such test images. Currently available images that we have been able to access either have no GT, are super-resolution images with no corresponding experimental low-resolution image, only have improvement in the xy resolution, do not have full 3D data available, with only xz or xy scans of an area, or are light sheet experiments with the GT provided by multi-view reconstruction, a somewhat different process with a lower resolution which is less likely to be well matched to the process we have modelled. The most suitable dataset we could find with GT available was only made available in a pre-processed state, where the upsampling had created line-type artefacts in the image which were easily picked up by the image processing.

Chapter 5

Conclusion & Outlook

5.1 Accessibility & efficacy of SMLM imaging

SMLM is a tremendously powerful technique. Of course, it offers a huge increase in potential resolution when compared with diffraction-limited fluorescence microscopy. However, where it truly stands out from other super-resolution modalities is its exceptional accessibility potential. In terms of hardware, its demands are the lowest of these, offering a much greater breadth of application to the humble widefield fluorescence microscope, and a significantly increased information-richness in the images it produces. Despite this, SMLM experiments are most frequently performed by microscopists and image analysts. It is still often seen as a tool of specialists, and it remains the case that a great deal of published seminal SMLM material is focussed on the continued refinement of the method. There are, broadly, two likely reasons for this that are specifically related to the usability and ease-of-use of the method. Firstly, the sample preparation for SMLM experiments can be much more involved than for other fluorescence microscopy approaches. Because molecules are generally localised and fitted individually, or statistically inferred from multiplex spatial data and fluctuations in signal, the fluorophores need to be very bright; additionally, they need to be capable of blinking or photoswitching easily. For experimenters, this means having a specific knowledge base about fluorophores and buffer preparation, which must be even broader for live-cell imaging. Secondly, algorithm selection for image reconstruction in SMLM is an intricate process, because there are a multiple factors to consider. What is the required time-frame for the experiment? What resolution is needed to get the result needed? What severity of artefacts is expected? What adjustments can be made in order to remove them from the structures of interest? Thirdly, expansion into 3D imaging, an increasingly important step for fluorescence imaging, is a more intimidating step to take in SMLM than it is in, say,

confocal imaging. This is because of the requisite manipulation and mathematical analysis of PSF geometry and the various effects it can produce on localisation precision and accuracy.

The work in this thesis looks to take steps in addressing the second and third issues above. Because molecule localisation and artefact generation are quite mathematically involved, the interplay between the above considerations can, for many, quickly become prohibitively complex. The experiments described here were conducted with two closely related aims in mind:

- To take further steps to overcome the problems of artefacts by providing technologies that can reliably do so. With the technologies that are currently available, it can be easy for non-experts to feel that artefacts are an enigmatic problem. They are difficult to spot and harder still to overcome – and when imaging in 3D, these problems become overwhelmingly complex, even for experienced microscopists.

- To expand the software toolkit available to bioscientists wishing to conduct super-resolution experiments, but for whom the complications of hardware access, and image optimisation and assessment, constitute too high a barrier to entry. A tool is only as valuable as the breadth of applications for which it is actually employed. The more people use SMLM, and feel that they can use it safely and correctly, the more valuable a tool it becomes. Similarly, super-resolved 3D imaging frequently has highly stringent hardware requirements when compared with other kinds of fluorescence microscopy. Because of the increasing importance of 3D imaging for drawing sound biological conclusions, the accessibility of increased axial resolution is paramount for ensuring that as many of these conclusions are drawn as possible.

With these goals in mind, we developed HAWKMAN to assess SMLM images, investigated HAWK's artefact removal on 3D SMLM datasets, and developed AF-FIRM3D to bring isotropic resolution to volumetric imaging.

5.2 HAWKMAN provides advanced and comprehensive image assessment

Artefacts and image bias make up SMLM's most significant challenge. When the GT is known, as in with simulated data, it can seem as though these issues present in fairly predictable and recognisable ways: sharpening, clustering, repeat localisations.

However, when using nothing more than human inspection to search for artefacts in novel experimental data, they can quickly seem mysterious and inscrutable. A major reason for this is the scarcity of SMLM image assessment tools on offer. Armed with a suitable means of testing algorithm performance on experimental data, expert knowledge of the many reconstruction algorithms available would no longer be required by researchers looking to conduct SMLM experiments. There are, by now, many powerful and useful technologies are on offer to assess algorithm performance; be it in terms of the resolution [51], proximity to the widefield image [81], or the ability to capture the lack of structure in randomly distributed data in a statistically accurate manner [80]. However, despite the fact that the problem of localisation bias and other artefacts is as old as the technique itself, still none of these innovations are able to *quantify the super-resolved resolution of local regions at which imaged structures have been faithfully rendered* on experimental data wherein the GT is not available.

HAWKMAN represents the very first technology that has offered such a solution to the problem of artefact detection. Two key insights to achieve this: first, that the GT is not needed to spot bias in SMLM data, only a reconstruction that doesn't have artificial sharpening; second, that only the spatial information, not the intensity distribution, of the reconstruction is needed to detect artefacts. An interesting by-product of the ability to detect artefacts in this way is that disagreement between HAWKMAN's input test and reference images can be indicative of structure that is unresolvable in both inputs. Taken together, these features constitute a huge leap forward in demystifying the problem of artefacts, by making the need for further experiment optimisation (or lack thereof) quick and obvious to detect. There is a conspicuously missing feature in HAWKMAN though, and that is the ability to detect sharpening in volumetric datasets.

Of course, HAWKMAN can be applied to z-slices of 3D datasets, or axial projections. However, these fail to capture the full complexity of 3D datasets. HAWKMAN's central operating mechanic is the reliance on a HAWK'd reference image, from which the location of artefacts can be determined due to the difference in its failure mode at sufficiently high activation densities when compared with non-HAWK'd reconstructions. However, when HAWKMAN was published, we did not yet know if HAWK processing would cause a *3D* SMLM reconstruction algorithm to fail differently than it would otherwise. Hence, the missing requisite results needed to begin HAWKMAN3D development is human-verified data that ensures HAWK's ability to successfully remove artefacts from various 3D SMLM datasets. Testing this capacity of HAWK would be also essential because 3D SMLM datasets contain

a much wider variety of artefacts than 2D ones. With this in mind, along with a recognition of the importance of 3D imaging for future results, the next major undertaking of this thesis was to acquire these results. Testing HAWK in 3D thus additionally paves the way for the development of a 3D SMLM image quality assessment algorithm, another unprecedented development in the field. As we showed in section 3.3.2, however, axial artefacts can be much more complex to analyse. While lateral localisation artefacts tend to remain located in the region of the actual structure, astigmatic artefacts, for instance, can result in localisations being moved to distant positions. It might thus be necessary to include a more nuanced metric than is present in the original HAWKMAN paper, to account for the potential severity of artefactual molecule displacement. At the very least, 3D HAWKMAN would require a method for visualising these artefacts. A good potential solution here would be to include an xz and yz view of the error maps. Ideally, it would also need an approach for handling the increase requisite computer memory required to perform the calculations for a 3D dataset.

5.3 HAWK successfully removes artefacts from 3D datasets

Though the results of these experiments have proven promising, further work must be done to assess the full capabilities of HAWK use on 3D datasets. This would be needed not only to assess the extent of HAWK's improvement to various kinds of dataset, but to ascertain whether HAWK processing introduces any novel issues when using other 3D methods. Potential candidates for further 3D HAWK testing include interferometric microscopy [88], FALCON3D [127], DAISY [135], and hybrid methods such as DHP + biplane [136]. Testing HAWK on these 3D SMLM modalities is important as they potentially represent stronger approaches – in terms of functional z-range, localisation precision, and artefact susceptibility – than the three methods we studied. They were not a priority in our research because they have a lower accessibility in that the requisite specialist hardware is only available to research groups that have the time and resources to obtain (in most cases, build) a specialist system. The astigmatic PSF, DHP, and biplane microscopy are outstanding in that their hardware requirements are minimal, requiring only a cylindrical lens, a phase mask, or a second camera respectively. This made assessing HAWK's efficacy when coupled with these modalities a top priority.

5.4 Improving the resolution of 3D imaging

As discussed in section 4.1, developing generative networks that can add missing information to microscope images without introducing bias is an extremely challenging undertaking. This is because most cost functions for training generative CNNs typically use the specific structures in the training data inform them of how signal ought to be distributed in an image. AFFIRM3D is very much in its early stages of development. However, we have demonstrated in our preliminary experiments that significant axial resolution restoration *is possible*, even without an adversary, and with a significant loss contribution from the Fourier domain, which contains significantly less structure-specific information. Though AFFIRM3D is not the first NN designed with the purpose of resotring isotropic resolution, it is the first that does so with an architecture that is specifically designed to remove bias and hallucination from the final network, once trained. Continued development of AFFIRM3D is an exciting prospect, and we look forward to bridging the accessibility gap to isotropic volumetric micrographs, without introducing fictitious structures into the output image. With this extra axial information more readily available to bioscientists, the imaging of tissues, cells, and cellular sub-structures in environments more faithful to their *in vivo* milieus will become a richer and more accessible source of biological discoveries.

5.5 Concluding remarks

Deep learning in bioscience imaging has seen particularly large growth over the last decade, with the advent of technologies like UNet, GAN, and CycleGAN proving to be exceptionally useful in extracting more information than is natively present biological images. Given the exceptional power of deep CNNs (that is, in terms of the complexity of tasks they can potentially perform, and the processing speed increases that they can potentially offer), as well as their increasing ubiquity, it can be tempting to think that the age of the image processing algorithm from first principles is drawing to a close. However, two-thirds of the technologies presented here, though developed very recently, were developed from first principles, utilising little more than convolution and cross-correlation. This research should also be taken to indicate that, even as the flexibility and interpretability of NNs continues to increase, there are many novel processing techniques written from first principles that are yet to be conceived of and developed. HAWK continues to count among the best options available for SMLM artefact management; its non-learning algorithm contributes to its flexibility, interpretability, and extensibility (in the case of

HAWKMAN).

The totality of the work presented in this thesis presents an optimistic outlook for the future of super-resolution imaging. It looks to offer a way out of the paradigm that super-resolution – in particular, SMLM and super-resolved axial imaging – are valuable discoveries *in themselves*, but remain the purview of physicists. It is my hope that the efficacy, ease-of-use, and most importantly, reliability associated with the techniques outlined and tested here will come to render SMLM reconstruction and image assessment routine practices for bioscientists, instead of being outsourced to specialists. In many ways, however, these results constitute early steps into the realm of standardising this kind of advanced fluorescence imaging. That is, the prospect of further development of these technologies – including HAWKMAN of 3D images, the use of HAWK on other types of 3D data, and the full development, implementation, and final release of AFFIRM3D for retrieving isotropic resolution from 3D images – promises further exciting developments for this field.

Bibliography

1. Abbe, E. On the Estimation of Aperture in the Microscope. *Journal of the Royal Microscopical Society* **1,** 388–423 (1881).
2. Wikipedia. *Abbe sine condition* `https://en.wikipedia.org/wiki/Abbe_sine_condition`.
3. Abramowitz, M. & Davidson, M. W. *Numerical Aperture and Resolution* `https://www.olympus-lifescience.com/en/microscope-resource/primer/anatomy/numaperture/`.
4. Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der Physik* **437,** 55–75 (1948).
5. Tsien, R. Y. Building and breeding molecules to spy on cells and tumors. *FEBS Letters* **579,** 927–32 (4 2005).
6. Sanderson, M. J., Smith, I., Parker, I. & Bootman, M. D. Fluorescence Microscopy. *Cold Spring Harbour Protocols* **35** (2014).
7. Pawley, J. *Handbook of Biological Confocal Microscopy* (2016).
8. Nwaneshiudu, A. *et al.* Introduction to confocal microscopy. *Journal of Investigative Dermatology* **132** (12 2012).
9. Airy, S. G. B. On the Diffraction of an Object-glass with Circular Aperture. *Transactions of the Cambridge Philosophical Society* **5** (1835).
10. Abbe, E. Beitrage zur Theorie des Mikroskops und der mikroskopischen. *Archiv für Mikroskopische Anatomie* **9,** 413–68 (1873).
11. Rayleigh, L. Investigations in optics, with special reference to the spectroscope. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **8,** 261–74 (1879).
12. Klar, T. A. & Hell, S. W. Subdiffraction resolution in far-field fluorescence microscopy. *Opt. Lett.* **24,** 954–956 (1999).

13. Heintzmann, R. & Cremer, C. G. Laterally modulated excitation microscopy: improvement of resolution by using a diffraction grating. *Optical Biopsies and Microscopic Techniques III* **3568** (1999).

14. Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *Journal of Microscopy* **198,** 82–7 (2 2001).

15. Shashkova, S. & Leake, M. C. Single-molecule fluorescence microscopy review: shedding new light on old problems. *Bioscience Reports* **37** (4 2017).

16. Dykstra, M. J. *Biological Electron Microscopy: Theory, Techniques, and Troubleshooting* (Springer, 2012).

17. Godin, A. G., Lounis, B. & Cognet, L. Super-resolution Microscopy Approaches for Live Cell Imaging. *Biophysical Journal* **107** (8 2014).

18. Jones, S. A., Shim, S.-H., He, J. & Zhuang, X. Super-resolution Microscopy Approaches for Live Cell Imaging. *Nature Methods,* 499–505 (2011).

19. Velasco, M. G. M. *et al.* 3D super-resolution deep-tissue imaging in living mice. *Optica* **8,** 442–450 (2021).

20. Heilemann, M. *et al.* Subdiffraction-Resolution Fluorescence Imaging with Conventional Fluorescent Probes. *Angewandte Chemie International Edition* **47,** 6172–6 (2008).

21. Marx, V. Probes: paths to photostability. *Nature Methods,* 187–90 (2015).

22. Grimm, J. B., Brown, T. A., English, B. P., Lionnet, T. & Lavis, L. D. Synthesis of Janelia Fluor HaloTag and SNAP-Tag Ligands and Their Use in Cellular Imaging Experiments. *Super-Resolution Microscopy* **1663,** 179–88 (2015).

23. Van Wee, R., Filius, M. & Joo, C. Completing the canvas: advances and challenges for DNA-PAINT super-resolution imaging. *Super-Resolution Microscopy* **46,** 918–30 (11 2021).

24. Klar, T. A., Jakobs, S., Dyba, M. & Hell, S. W. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Applied Physical Sciences* **97,** 8206–10 (2000).

25. Dyba, M., Jakobs, S. & Hell, S. W. Immunofluorescence stimulated emission depletion microscopy. *Nature Biotechnology,* 1303–4 (2003).

26. Nägerl, U. V., Willig, K. I., Hein, B., Hell, S. W. & Bonhoeffer, T. Immunofluorescence stimulated emission depletion microscopy. *Nature Biotechnology* **105,** 18982–7 (48 2008).

27. Schermelleh, L., Heintzmann, R. & Leonhardt, H. A guide to super-resolution fluorescence microscopy. *Journal of Cell Biology,* 165–75 (190 2010).

28. Hirvonen, L. M., Wicker, K., Mandula, O. & Heintzmann, R. Structured illumination microscopy of a living cell. *European Biophysics Journal,* 807–12 (2009).

29. Fiolka1, R., Shao, L., Rego, E. H., Davidson, M. W. & Gustafsson, M. G. L. Time-lapse two-color 3D imaging of live cells with doubled resolution using structured illumination. *PNAS* **109,** 5311–5 (104 2011).

30. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging with stochastic optical reconstruction microscopy (STORM). *Nature Methods* **3,** 793–5 (2006).

31. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science,* 1642–5 (2006).

32. Hess, S. T., Girirajan, T. P. K. & Mason, M. D. Ultra-high resolution imaging by fluorescence photoactivation localisation microscopy. *Biophysical Journal* **91,** 4258–72 (2006).

33. Cox, S. Super-Resolutionn Imaging in Live Cells. *Developmental Biology* **401,** 715–81 (2015).

34. Jimenez, A., Friedl, K. & Leterrier, C. About samples, giving examples: Optimized Single Molecule Localization Microscopy. *Methods* **174,** 100–114 (2020).

35. Cox, S. & Jones, G. E. Imaging cells at the nanoscale. *The International Journal of Biochemistry & Cell Biology* **45,** 1669–78 (8 2013).

36. Bates, M., Blosser, T. R. & Zhuang, X. Short-Range Spectroscopic Ruler Based on a Single-Molecule Optical Switch. *Physical Review Letters* **94,** 94–5 (2005).

37. Heilemann, M., van de Linde, S., Mukherjee, A. & Sauer, M. Super-Resolution Imaging with Small Organic Fluorophores. *Angewandte Chemie International Edition* **48,** 6903–8 (37 2009).

38. Van de Linde, S. *et al.* Investigating Cellular Structures at the Nanoscale with Organic Fluorophores. *Angewandte Chemie International Edition* **48,** 6903–8 (37 2009).

39. Henriques, R. *et al.* QuickPALM: 3D real-time photoactivation nanoscopy image processing in ImageJ. *Nature Methods* **7,** 339–40 (2010).

40. Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z. & Hagen, G. M. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30,** 2389–90 (2014).

41. Pauli, W. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Physik* **31,** 765–83 (1925).

42. Song, L., Hennink, E. J., Young, I. T. & Tanke, H. J. Photobleaching Kinetics of Fluorescein in Quantitative Fluorescence Microscopy. *Biophysical Journal* **68,** 2588–600 (1995).

43. Wäldchen, S., Lehmann, J., Klein, T., van de Linde, S. & Sauer, M. Light-induced cell damage in live-cell super-resolution microscopy. *Scientific Reports* **5** (2015).

44. Song, L., Varma, C. A., Verhoeven, J. W. & Tanke, H. J. Influence of the triplet excited state on the photobleaching kinetics of fluorescein in microscopy. *Biophysical Journal* **70,** 2959–68 (1996).

45. Song, L., Hennink, E. J., Young, I. T. & Tanke, H. J. Photobleaching Kinetics of Fluorescein in Quantitative Fluorescence Microscopy. *Biophysical Journal* **68,** 2588–600 (1995).

46. Dempsey, G. T. *et al.* Photoswitching Mechanism of Cyanine Dye. *Journal of the American Chemical Society* **131,** 18192–93 (2009).

47. Dempsey, G. T., Vaughan, J. C., Chen, K. H., Bates, M. & Zhuang, X. Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nature Methods* **8,** 1027–36 (2011).

48. Dempsey, G. T., Vaughan, J. C., Chen, K. H., Bates, M. & Zhuang, X. Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nature Methods* **8,** 1027–36 (2011).

49. Shi, X., Lim, J. & Ha, T. Acidification of the Oxygen Scavenging System in Single-Molecule Fluorescence Studies: In Situ Sensing with a Ratiometric Dual-Emission Probe. *Analytical Chemistry* **82,** 6132–8 (2010).

50. Nieves, D. J., Gaus, K. & Baker, M. A. B. DNA-Based Super-Resolution Microscopy: DNA-PAINT. *Genes* **9** (2018).

51. Nieuwenhuizen, R. P. J. *et al.* Measuring image resolution in optical nanoscopy. *Nature Methods* **10,** 577–62 (2013).

52. Rees, E. J. *et al.* Blind assessment of localisation microscope image resolution. *Optical Nanoscopy* **1** (2012).

53. Ober, R. J., Ram, S. & Ward, E. S. Localization accuracy in single-molecule microscopy. *Biophysical Journal* **86,** 1185–1200 (2004).

54. Whelan, D. R. & Bell, T. D. M. Image artifacts in Single Molecule Localization Microscopy: why optimization of sample preparation protocols matters. *Scientific Reports* **5** (2015).

55. Liccardo, F. *et al.* Precision localization of cellular proteins with fluorescent Fab-based probes. *Cold Spring Harbour Laboratory* (2021).

56. Richter, K. N. *et al.* Glyoxal as an alternative fixative to formaldehyde in immunostaining and super-resolution microscopy. *The Embo Journal* **37,** 139–59 (2018).

57. Thompson, R. E., Larson, D. R. & Webb, W. W. Precise Nanometer Localization Analysis for Individual Fluorescent Probes. *Biophysical Journal* **82,** 2775–83 (2002).

58. Long, F., Zeng, S. & Huang, Z.-L. Localization-based super-resolution microscopy with an sCMOS camera Part II: Experimental methodology for comparing sCMOS with EMCCD cameras. *Cold Spring Harbour Laboratory* **20,** 17741–59 (16 2012).

59. Huang, F. *et al.* Video-rate nanoscopy using sCMOS camera–specific single-molecule localization algorithms. *Nature Methods,* 653–8 (2013).

60. Martin-Fernandez, M. L., Tynan, C. J. & Webb, S. E. D. A 'pocket guide' to total internal reflection fluorescence. *Journal of Microscopy* **252,** 16–22 (2013).

61. Fish, K. N. Total Internal Reflection Fluorescence (TIRF) Microscopy. *Current Protocols in Cytometry* (2009).

62. Carter, A. R. *et al.* Stabilization of an optical microscope to 0.1 nm in three dimensions. *Applied Optics* **46,** 421–7 (2007).

63. Lee, S. H. *et al.* Using fixed fiduciary markers for stage drift correction. *Optica Publishing Group* **20,** 12177–83 (2012).

64. Mlodzianoski, M. J. *et al.* Sample drift correction in 3D fluorescence photoactivation localisation microscopy. *Optica Publishing Group* **19,** 15009–19 (2011).

65. Pertsinidis, A., Zhang, Y. & Chu, S. Subnanometre single-molecule localisation, registration and distance measurements. *Nature* **466,** 647–51 (2010).

66. Descloux, A., Grußmayer, K. S. & Radenovic, A. Parameter-free image resolution estimation based on decorrelation analysis. *Nature Methods* **16,** 918–24 (2019).

67. Fox-Roberts, P. *et al.* Local dimensionality determines imaging speed in localisation microscopy. *Nature Communications* **8** (2017).

68. Wolter, S., Endesfelder, U., van de Linde, S., Heilemann, M. & Sauer, M. Resolution limit of image analysis algorithms. *Nature Communications* **10** (2019).

69. Burgert, A., Letschert, S. & Sören Doose, M. S. Artifacts in single-molecule localization microscopy. *Histochemistry and Cell Biology,* 123–31 (144 2015).

70. Helmerich, D. A., Beliu, G., Matikonda, S. S., Schnermann, M. J. & Sauer, M. Photoblueing of organic dyes can cause artifacts in super-resolution microscopy. *Nature Methods,* 253–7 (2021).

71. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localisation microscopy. *Nature Biotechnology* **36,** 460–8 (2019).

72. Grimm, J. B. *et al.* A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nature Methods,* 244–50 (2015).

73. Huang, F., Schwartz, S. L., Byars, J. M. & Lidke, K. A. Simultaneous multiple-emitter fitting for single molecule super-resolution imaging. *Biomedical Optics Express* **2,** 1377–93 (2011).

74. Holdeen, S. J., Uphoff, S. & Kapanidis, A. N. DAOSTORM: an algorithm for high-density super-resolution microscopy. *Nature Methods* **8,** 279–80 (2011).

75. Dertinger, T., Colyer, R., Iyer, G., Weiss, S. & Enderlein, J. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Biomedical Optics Express* **106,** 1377–93 (2009).

76. Cox, S. *et al.* Bayesian localisation microscopy reveals nanoscale podosome dynamics. *Nature Methods* **9,** 195–200 (2012).

77. Culley, S., Tosheva, K. L., Pereira, P. M. & Henriques, R. SRRF: Universal live-cell super-resolution microscopy. *The International Journal of Biochemistry and Cell Biology* **101,** 74–9 (2018).

78. Marsh, R. J. *et al.* Richard J Marsh and Karin Pfisterer and Pauline Bennett and Liisa M Hirvonen and Mathias Gautel and Gareth E Jones and Susan Cox. *Nature Methods* **15,** 689–92 (2018).

79. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9,** 671–5 (2012).

80. Cohen, E. A. K., Abraham, A. V., Ramakrishnan, S. & Ober, R. J. Resolution limit of image analysis algorithms. *Nature Communications* **10** (2019).

81. Culley, S. *et al.* NanoJ-SQUIRREL: quantitative mapping and minimisation of super-resolution optical imaging artefacts. *Nature Methods* **15,** 263–6 (2018).

82. Marsh, R. J. *et al.* Sub-diffraction error mapping for localisation microscopy images. *Nature Communications* **12** (2021).

83. Huang, B., Wang, W., Bates, M. & Zhuang, X. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science* **319,** 810–3 (2008).

84. Pavani, S. R. P. *et al.* Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences* **106,** 2995–2999 (2009).

85. Prabhat, P., Ram, S., Ward, E. S. & Ober, R. J. Simultaneous imaging of different focal planes in fluorescence microscopy for the study of cellular dynamics in three dimensions. *Institute of Electrical and Electronics Engineers* **3,** 237–242 (2004).

86. Nehme, E. *et al.* DeepSTORM3D: dense 3D localisation microscopy and PSF design by deep learning. *Nature Methods* **17** (2020).

87. Speiser, A. *et al.* Deep learning enables fast and dense single-molecule localization with high accuracy. *Nature Methods,* 1082–90 (2021).

88. Shtengel, G. *et al.* Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 3125–3130 (2009).

89. Shechtman, Y., Sahl, S. J., Backer, A. S. & Moerner1, W. E. Optimal Point Spread Function Design for 3D Imaging. *Physical Review Letters* **113** (2014).

90. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5,** 115–133 (1943).

91. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).

92. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature,* 533–6 (1986).

93. Johnson, D. *Back Propagation in Neural Network: Machine Learning Algorithm* `https://www.guru99.com/backpropogation-neural-network.html`.

94. Wengert, R. E. A simple automatic derivative evaluation program. *Communications of the ACM* **7,** 463–4 (1964).

95. Robbins, H. E. A Stochastic Approximation Method. *Annals of Mathematical Statistics* **22,** 400–407 (1951).

96. Kiefer, J. & Wolfowitz, J. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics* **23,** 462–466 (1952).

97. Widrow, B. *Adaptive 'Adaline' Neuron Using Chemical Memistors* (1960), 462–466.

98. Bottou, L., Curtis, F. E. & Noceda, J. Optimization Methods for Large-Scale Machine Learning. *SIAM Review* **60,** 463–4 (2018).

99. Fukushima, K. & Miyake, S. *Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition* in *Competition and Cooperation in Neural Nets* (eds Amari, S.-i. & Arbib, M. A.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 1982), 267–285. ISBN: 978-3-642-46466-9.

100. LeCun, Y., Bengio, Y. & Hinton, G. Deep Learning. *Nature,* 436–44 (2015).

101. Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems* **33,** 6999–7019 (12 2021).

102. Newby, J. M., Schaefer, A. M., Lee, P. T. & Lai, S. K. Convolutional neural networks automate detection for tracking of submicron-scale particles in 2D and 3D. *PNAS,* 9026–31 (155 2018).

103. Weigert, M. *et al.* Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature Methods* **15,** 1090–7 (2018).

104. Chen, J. *et al.* Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes. *Nature Methods* **18,** 678–687 (2021).

105. Wang, H. *et al.* Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nature Methods* **16,** 103–10 (1 2019).

106. Park, H. *et al.* Deep learning enables reference-free isotropic super-resolution for volumetric fluorescence microscopy. *Nature Communications* **13** (2022).

107. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science* **9351,** 234–241 (2015).

108. Liskowski, P. & Krawiec, K. Segmenting Retinal Blood Vessels With Deep Neural Networks. *IEEE Transactions on Medical Imaging* **35,** 2369–80 (2016).

109. Creswell, A. *et al.* Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine* **35,** 53–65 (1 2018).

110. Goodfellow, I. *et al.* Generative Adversarial Nets. *Advances in Neural Information Processing Systems* **27** (2014).

111. Nehme, E., Weiss, L. E., Michaeli, T. & Shechtman, Y. Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica* **5,** 458–64 (2018).

112. Sage, D. *et al.* Super-resolution fight club: assessment of 2D and 3D single-molecule localisation microscopy software. *Nature Methods* **16,** 387–395 (2019).

113. Szegedy, C. *et al.* Intriguing properties of neural networks (2013).

114. Mohseni, S., Zarei, N. & Ragan, E. D. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* **11,** 1–45 (3–4 2021).

115. Sage, D. *et al.* Quantitative evaluation of software packages for single-molecule localisation microscopy. *Nature Methods* **12,** 717–24 (2015).

116. Bradley, D. & Roth, G. Adaptive Thresholding using the Integral Image. *Journal of Graphics Tools* **12,** 13–21 (2007).

117. Birch, P. *et al.* Approximate bandpass and frequency response models of the difference of Gaussian filter. *Optics Communications* **283,** 4942–8 (2010).

118. Lindeberg, T. Scale-space for discrete signals. *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence* **12,** 234–54 (1990).

119. Burgert, A., Letschert, S., Doose, S. & Sauer, M. Artifacts in single-molecule localisation microscopy. *Histochemistry and Cell Biology* **144,** 123–31 (2015).

120. Pfisterer, K., Lumicisi, B. & Parsons, M. Imaging of Human Cancer Cells in 3D Collagen Matrices. *Bio-protocol* (2021).

121. Van Helvert, S., Storm, C. & Friedl, P. Mechanoreciprocity in cell migration. *Nature Cell Biology* **20,** 8–20 (2018).

122. Yamada, K. M. & Sixt, M. Mechanisms of 3D cell migration. *Nature Review Molecular Cell Biology* **20,** 738–52 (2019).

123. Pfisterer, K. *et al.* FMNL2 regulates dynamics of fascin in filopodia. *Journal of Cell Biology* **2019,** 1253–59 (5 2020).

124. Babcock, H., Sigal, Y. M. & Zhuang, X. A high-density 3D localisation algorithm for stochastic optical reconstruction microscopy. *Optical Nanoscopy* **1** (2012).

125. Barsic, A., Grover, G. & Piestun, R. Three-dimensional super-resolution and localisation of dense clusters of single molecules. *Scientific Reports* **4** (2014).

126. Gu, L. *et al.* High-Density 3D Single Molecular Analysis Based on Compressed Sensing. *Biophysical Journal* **106,** 2443–2449 (2014).

127. Birch, P. *et al.* 3D high-density localisation microscopy using hybrid astigmatic/ biplane imaging and sparse image reconstruction. *Biomedical Optics Express* **5,** 3935–3948 (2014).

128. Bourg, N. *et al.* Direct optical nanoscopy with axially localised detection. *Nature Photonics* **9,** 587–93 (2015).

129. Szalai, A. M. *et al.* Three-dimensional total-internal reflection fluorescence nanoscopy with nanometric axial resolution by photometric localisation of single molecules. *Nature Communications* **12** (2021).

130. Li, Y. *et al.* Real-time 3D single-molecule localization using experimental point spread functions. *Nature Methods* **15,** 367–9 (2018).

131. Hagen, G., Ovesny, M., Svindrych, Z., Borkovec, J. & Křížek, P. *ThunderSTORM: Methodology and Algorithms (update)*

132. Bayas, C., von Diezmann, A., Gustavsson, A.-K. & Moerner, W. E. Easy-DHPSF 2.0: open-source software for three-dimensional localization and two-color registration of single molecules with nanoscale accuracy. *Protocol Exchange, from Nature Portfolio* (2019).

133. Gustavsson, A.-K., Petrov, P. N., Lee, M. Y., Shechtman, Y. & Moerner, W. E. 3D single-molecule super-resolution microscopy with a tilted light sheet. *Nature Communications* **9** (2018).

134. Lee, H.-l. D., Sahl, S. J., Lew, M. D. & Moerner, W. E. The double-helix microscope super-resolves extended biological structures by localizing single blinking molecules in three dimensions with nanoscale precision. *Applied Physics Letters* **100** (15 2012).

135. Cabriel, C. *et al.* Combining 3D single molecule localisation strategies for reproducible bioimaging. *Nature Communications* **10** (2019).

136. Badieirostami, M., Lew, M. D., Thompson, M. A. & Moerner, W. E. Three-dimensional localisation precision of the double-helix point spread function versus astigmatism and biplane. *Applied Physics Letters* **97** (2010).

137. Hell, S. & Stelzer, E. H. K. Properties of a 4Pi confocal fluorescence microscope. *Journal of the Optical Society of America* **9,** 2159–66 (12 1992).

138. Gustafsson, M. G., Agar, D. A. & Sedat, J. W. I5M: 3D widefield light microscopy with better than 100 nm axial resolution. *Journal of Microscopy* **195,** 10–16 (1999).

139. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface* (2018).

140. Uhlmann, V., Donati, L. & Sage, D. Deep Learning. *IEEE Signal Processing Magazine* **39,** 73–86 (2 2022).

141. Lucas, A. M. *et al.* Open-source deep-learning software for bioimage segmentation. *Molecular Biology of the Cell* **32,** 823–1005 (9 2021).

142. Weiler, M., Hamprecht, F. A. & Storath, M. *Learning Steerable Filters for Rotation Equivariant CNNs* in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 849–858.

143. Goel, A., Agarwal, A., Vatsa, M., Singh, R. & Ratha, N. K. *DNDNet: Reconfiguring CNN for Adversarial Robustness* in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020), 103–110.

144. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. *Cell Detection with Star-Convex Polygons* in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (eds Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C. & Fichtinger, G.) (Springer International Publishing, 2018), 265–273.

145. Mandal, S. & Uhlmann, V. *Splinedist: Automated Cell Segmentation With Spline Curves* in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (2021), 1082–1086.

146. Hollandi, R. *et al.* nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer. *Cell Systems* **10,** 453–458 (2020).

147. Laine, R. F., Jacquemet, G. & Krull, A. Imaging in focus: An introduction to denoising bioimages in the era of deep learning. *The International Journal of Biochemistry & Cell Biology* **140,** 1060–77 (2021).

148. Laine, S., Karras, T., Lehtinen, J. & Aila, T. *High-Quality Self-Supervised Deep Image Denoising* in *Advances in Neural Information Processing Systems* (eds Wallach, H. *et al.*) **32** (Curran Associates, Inc., 2019).

149. Krull, A., Buchholz, T.-O. & Jug, F. *Noise2Void - Learning Denoising From Single Noisy Images* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

150. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods,* 917–20 (2018).

151. Huang, B. *et al.* Enhancing image resolution of confocal fluorescence microscopy with deep learning. *PhotoniX* **4** (2023).

152. Ling, C. *et al.* Fast structured illumination microscopy via deep learning. *Photon. Res.* **8,** 1350–1359 (2020).

153. Xypakis, E. *et al.* Deep learning for blind structured illumination microscopy. *Scientific Reports* **12** (2022).

154. Gómez-de-Mariscal, E. *et al.* DeepImageJ: A user-friendly environment to run deep learning models in ImageJ. *Nature Methods* **18** (2021).

155. Von Chamier, L. *et al.* Democratising deep learning for microscopy with ZeroCostDL4Mic. *Nature Communications* **12** (2021).

156. Segebarth, D. *et al.* On the objectivity, reliability, and validity of deep learning enabled bioimage analyses. *eLife* **9** (eds Fraser, S. E., Bronner, M. E., Ruderman, D. & Danuser, G.) (2020).

157. Ouyang, W. *et al.* BioImage Model Zoo: A Community-Driven Resource for Accessible Deep Learning in BioImage Analysis. *Cold Spring Harbor Laboratory* (2022).

158. Liu, Z. *et al.* A survey on applications of deep learning in microscopy image analysis. *Computers in Biology and Medicine* **134,** 104523 (2021).

159. Belthangady, C. & Royer, L. A. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nature Methods* **16,** 1215–25 (2019).

160. Costello, I. & Cox, S. Analysing errors in single-molecule localisation microscopy. *The International Journal of Biochemistry and Cell Biology* **134** (2021).

161. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *ICCV* (2017).

162. Becker, J. L. & Souza, G. R. Using space-based investigations to inform cancer research on Earth. *Nature Reviews Cancer* **13,** 315–27 (2013).

163. Tay, A., Schweizer, F. E. & Carlo, D. D. Micro- and nano-technologies to probe the mechano-biology of the brain. *Lab on a Chip* **16,** 1962–77 (11 2006).

164. Momeny, M., Latif, A. M., Sarram, M. A., Sheikhpour, R. & Zhang, Y. D. A noise robust convolutional neural network for image classification. *Results in Engineering* **10** (2021).

165. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. *Computer Science* (2017).

166. Dussault, D. & Hoess, P. Noise performance comparison of ICCD with CCD and EMCCD cameras. *Infrared Systems and Photoelectronic Technology* **5563** (2004).

167. Charbon, E. Single-photon imaging in complementary metal oxide semiconductor processes. *Philosophical Transactions of the Royal Society A* **372** (2012 2014).

168. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13,** 600–12 (2004).

169. Wanga, Z., Simoncelli, E. & Bovik, A. Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers* **2,** 1398–402 (2004).

170. Mittal, A., Moorthy, A. K. & Bovik, A. C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* **21,** 4695–708 (12 2012).

171. Ketkar, N. *Deep Learning with Python A Hands-on Introduction* chap. Stochastic Gradient Descent (APRESS, 2017).

172. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* 2014.

173. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming,* 4695–708 (1989).

174. Cheeseman, B. L., Günther, U., Gonciarz, K., Susik, M. & Sbalzarini, I. F. Adaptive particle representation of fluorescence microscopy images. *Nature Communications* **9,** 51–60 (1 2018).