Technische Universität Dresden

Content-Aware Image Restoration Techniques without Ground Truth and Novel Ideas to Image Reconstruction

---

Dissertation

zur Erlangung des akademischen Grades

Doktor rerum naturalium (Dr. rer. nat.)

vorgelegt an der

Technischen Universität Dresden

Fakultät Informatik

eingereicht von

Tim-Oliver Buchholz
geboren am 21.06.1991 in Locarno, Schweiz

Gutachter:
Prof. Dr. rer. nat. Stefan Gumhold, Technische Universität Dresden
Dr. Anna Kreshuk, EMBL Heidelberg
Dr. Florian Jug, MPI-CBG Dresden

Dresden, den 10. Juni 2021

Summary

In this thesis I will use state-of-the-art (SOTA) image denoising methods to denoise electron microscopy (EM) data. Then, I will present Noise2Void a deep learning based self-supervised image denoising approach which is trained on single noisy observations. Eventually, I approach the missing wedge problem in tomography and introduce a novel image encoding, based on the Fourier transform which I am using to predict missing Fourier coefficients directly in Fourier space with Fourier Image Transformer (FIT). In the next paragraphs I will summarize the individual contributions briefly.

Electron microscopy is the go to method for high-resolution images in biological research. Modern scanning electron microscopy (SEM) setups are used to obtain neural connectivity maps, allowing us to identify individual synapses. However, slow scanning speeds are required to obtain SEM images of sufficient quality. In (Weigert et al. 2018) the authors show, for fluorescence microscopy, how pairs of low- and high-quality images can be obtained from biological samples and use them to train content-aware image restoration (CARE) networks. Once such a network is trained, it can be applied to noisy data to restore high quality images. With SEM-CARE I present how this approach can be directly applied to SEM data, allowing us to scan the samples faster, resulting in 40- to 50-fold imaging speedups for SEM imaging.

In structural biology cryo transmission electron microscopy (cryo TEM) is used to resolve protein structures and describe molecular interactions. However, missing contrast agents as well as beam induced sample damage (Knapek and Dubochet 1980) prevent acquisition of high quality projection images. Hence, reconstructed tomograms suffer from low signal-to-noise ratio (SNR) and low contrast, which makes post-processing of such data difficult and often has to be done manually. To facilitate down stream analysis and manual data browsing of cryo tomograms I present cryoCARE a Noise2Noise (Lehtinen et al. 2018) based denoising method which is able to restore high contrast, low noise tomograms from sparse-view low-dose tilt-series. An implementation of cryoCARE is publicly available as Scipion (de la Rosa-Trevín et al. 2016) plugin.

Next, I will discuss the problem of self-supervised image denoising. With cryoCARE I exploited the fact that modern cryo TEM cameras acquire multiple low-dose images, hence the Noise2Noise (Lehtinen et al. 2018) training

paradigm can be applied. However, acquiring multiple noisy observations is not always possible *e.g.* in live imaging, with old cryo TEM cameras or simply by lack of access to the used imaging system. In such cases we have to fall back to self-supervised denoising methods and with Noise2Void I present the first self-supervised neural network based image denoising approach. Noise2Void is also available as an open-source Python package and as a one-click solution in Fiji (Schindelin et al. 2012).

In the last part of this thesis I present Fourier Image Transformer (FIT) a novel approach to image reconstruction with Transformer networks. I develop a novel 1D image encoding based on the Fourier transform where each prefix encodes the whole image at reduced resolution, which I call Fourier Domain Encoding (FDE). I use FIT with FDEs and present proof of concept for super-resolution and tomographic reconstruction with missing wedge correction. The missing wedge artefacts in tomographic imaging originate in sparse-view imaging. Sparse-view imaging is used to keep the total exposure of the imaged sample to a minimum, by only acquiring a limited number of projection images. However, tomographic reconstructions from sparse-view acquisitions are affected by missing wedge artefacts, characterized by missing wedges in the Fourier space and visible as streaking artefacts in real image space. I show that FITs can be applied to tomographic reconstruction and that they fill in missing Fourier coefficients. Hence, FIT for tomographic reconstruction solves the missing wedge problem at its source.

Acknowledgements

This thesis is the result of my work in the Jug-Lab at the Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG) and would not have been possible without the great support of Florian Jug and many others. Florian Jug is a wonderful supervisor who cares about his students and even when the pandemic prevented us from meeting in person, it didn't take long until we had our own virtual lab in gather.town. I truly appreciate the guidance Florian provided over the last years and all the wonderful scientific discussions as well as the many fun riddles we solved as a group.

I would like to thank Alexander Krull with whom I had the pleasure to work together. Alexander is a great listener and teacher, his patience and ability to explain complicated topics concise is inspiring. I thank Mangal Prakash for being such a wonderful lab partner, always available to discuss new ideas and bring them to paper. It was great fun to exchange the newest tips and tricks in project organization with Manan Lalit, as well as the many discussions we had ranging from culture to novel research topics. I had a lot of fun working on science outreach projects with Deborah Schmidt, who created beatiful displays of our research. I thank current and past members of the Jug-Lab Anna Goncharova, Nuno Pimpão Martins, Tobias Pietzsch, Matthias Arzt, Tom Burke, Joran Deschamps, and Gabriella Turek, as well as all the members of MPI-CBG and CSBD.

I would like to thank my thesis advisory committee, Ivo Sbalzarini, Carlo Vittorio Cannistraci and Gaia Pigino. Special thanks goes out to the whole Pigino-Lab that spent countless hours explaining and acquiring data for my projects. Specifically I want to thank Mareike Jordan for teaching me how to use IMOD and PEET. Furthermore, I am grateful for Anna Kreshuk and Stefan Gumhold for agreeing to review my thesis.

I thank Manan Lalit, Mangal Prakash, Nuno Pimpão Martins, and Robert Haase for their feedback on my writing.

I thank Tamara Stepanyan for joining me in Dresden and the countless wonderful hours we spent together over the last years. Finally, I am grateful for my family, especially my parents Brigitte and Klaus Buchholz for always being there and supporting me in following my dreams.

Contents

List of Figures

List of Tables

Acronyms

ATUM Automated Tape-Collecting Ultramicrotome.

CARE content-aware image restoration.

CNN Convolutional Neural Network.

cryo TEM cryo transmission electron microscopy.

CT computed tomography.

CTF contrast transfer function.

EM electron microscopy.

FBP filtered backprojection.

FDE Fourier Domain Encoding.

FIB-SEM focused ion beam scanning electron microscopy.

FIT Fourier Image Transformer.

FIT: SRes Fourier Image Transformer for super-resolution.

FIT: TRec Fourier Image Transformer for tomograpic reconstruction.

MAE mean absolute error.

MSE mean squared error.

NA numerical aperture.

NAD non-linear anisotropic diffusion.

NLP natural language processing.

ODA outer dynein arm.

PALM Photoactivated Localization Microscopy.

PSF point spread function.

PSNR peak signal-to-noise ratio.

SBF-SEM serial block face scanning electron microscopy.

SEM scanning electron microscopy.

SIM structured illumination microscopy.

SNR signal-to-noise ratio.

SOTA state-of-the-art.

SPA single particle analysis.

STED stimulated emission depletion microscopy.

STORM Stochastic Optical Reconstruction Microscopy.

TGV Total Generalized Variation.

TV Total Variation.

VAE variational auto-encoder.

Chapter 1

Introduction

For many of us, vision is an important sense which allows us to experience the world around us. With our eyes we can observe the surrounding world and understand it better. While we gather information about our surroundings through vision, it is commonly known that this perception of reality can be confounded by illusions, distance and size of objects, in other words vision can be deceptive. This superficial aspect to vision can be dangerous and it is always important to not just accept what we see, but to question and reason about our observations. Engaging our curiosity drives us to take a closer look and many tools have evolved over time and exist for that purpose. Already in the classical antiquity in the Middle East and the Mediterranean people used simple lenses, *i.e.* water filled glass globes, to aid their vision or ignite fires (Sines and Sakellarakis 1987). Ibn al-Haytham was arguably the first scientist to describe the convex lens used for magnification in 1021. His work was translated into Latin in the 13th century, which lead to the development of the first reading glasses, soon after. In 1590 the Dutch opticians Zacharias and Hans Janssen invented the compound microscope by aligning two lenses within a sliding tube (T. C. Kriss and V. M. Kriss 1998). However, it took until the 17th century when Antonie van Leeuwenhoek was able to create the first high quality lenses, which allowed him to accurately describe single cell sized objects which is often seen as the first application of microscopy in biology. Many technical advances happened in the next 150 years and in 1848 Carl Zeiss opened the first microscope workshop in Jena, Germany (van Zuylen 1981). This essentially started the mass production of microscopes and made them widely available for scientific discoveries. With standardized microscopes becoming more prevalent many new imaging methods and techniques got published. The most common microscopy technique is brightfield microscopy, where the sample is illuminated from behind with bright light and the image forms due to light occlusion. An important step in brightfield microscopy was the Köhler illumination proto-

col presented by August Köhler in 1883. By using additional lenses the filament image of the illumination lamp is moved out of the image plane, which results in an even sample illumination (Köhler 1893). Today, on of the most accessible birghtfield microscope is the foldscope (Cybulski et al. 2014), which uses the sun to illuminate the sample. The foldscope enables people around the globe to get a glimpse at the microscopic world and investigate their direct environment. Microscopy is used in many disciplines and especially in biology it is a core research tool. Many biological experiments were only even possible after development of novel imaging techniques and biology is still a strong driving force for imaging developments. The next important imaging technique was discovered in the 20th century by Fritz Zernike and is called phase-contrast microscopy (Zernike 1942), for which he was awarded the Nobel Prize in physics. With the discovery of DAPI as fluorescence staining (Kapuscinski 1995) fluorescence microscopy was born, which allows microscopists to tag specific proteins with fluorescence markers and image them at high resolution. However, all optical systems have a physical resolution limit d, which is linked to the diffraction limit of the illumination light. Ernst Abbé, a physicist working with Zeiss (T. C. Kriss and V. M. Kriss 1998), described this relationship already in 1873 through the relation

$$d = \frac{\lambda}{2n \sin \theta}, \tag{1.1}$$

where λ is the wavelength of the excitation light, n is the refractive index of the lens and θ is the maximal half-angle of light that can enter the lens (Abbe 1873). $n \sin \theta$ is also known as the numerical aperture (NA) and modern optical systems can reach NAs up to 1.6. So for a NA = 1.4 and green light with a wavelength of 550nm a theoretical resolution of $d \approx 200$nm can be reached. With this resolution we can observe cells and localize molecules within cells, however it is impossible to resolve molecules and understand their structure.

Nevertheless, there exist technically advanced methods to go beyond the physical resolution limit of visible light and in 2014, and Eric Betzig, Stefan W. Hell and William E. Moerner were awarded the Nobel Prize in chemistry for having bypassed the diffraction limit of visible light (Möckl et al. 2014). Specifically Stefan W. Hell was rewarded for his work on stimulated emission depletion microscopy (STED) (Hell and Wichmann 1994), Eric Betzig and William E. Moerner set the ground work for single molecule microscopy like Photoactivated Localization Microscopy (PALM) (Betzig et al. 2006) and Stochastic Optical Reconstruction Microscopy (STORM) (Rust et al. 2006). Another method to increase resolution of light microscopy is structured illumination microscopy (SIM) (Guerra 1995). While these methods achieve super-resolution in the field of light microscopy, even higher resolutions are possible with electron microscopes.

The wavelength of electrons depends on their speed, where higher speeds result in shorter wavelengths, which allows us to image at much higher resolutions. Modern electron microscopes routinely operate at sub Ångström (1Å = 0.1nm) resolution. In the following sections I will introduce scanning electron microscopy (SEM) and cryo transmission electron microscopy (cryo TEM), two widely used methods in biological research.

1.1 Scanning Electron Microscopy

Scanning electron microscopy (SEM) is a popular electron microscopy method used to image large samples at high resolution. In SEM an electron beam is focused in a single spot and a dehydrated and stained sample is scanned row by row (Collett 1970). Sample preparation protocols for tissue imaging usually include washing, fixation, dehydration and staining steps. Even though the sample preparation tends to be complex, SEM is often the method of choice to image large tissues at high resolution (Golding et al. 2016). Therefore multiple volumetric SEM techniques have been developed over the last decades. In general SEM methods can be divided into destructive and non-destructive methods. Serial block face scanning electron microscopy (SBF-SEM) (Denk and Horstmann 2004) and focused ion beam scanning electron microscopy (FIB-SEM) (Heymann et al. 2006) are deconstructive, because the imaged surface is either cut off with a diamond knife or milled away with a focused ion beam. In array tomography and serial-section SEM (Horstmann et al. 2012) on the other hand the sample is sectioned before it is put into the microscope, hence these are non-destructive methods. All of these methods have their own advantages and disadvantages. For example non-destructive methods are usually used with complementary imaging approaches like light microscopy to localize fluorescent labellings prior to electron microscopy imaging, however the axial resolution is limited compared to deconstructiv approaches. From the two deconstructive methods FIB-SEM achieves higher axial resolution than SBF-SEM. However, the milling process is FIB-SEM is more time consuming, hence SBF-SEM continues to be the method of choice for large tissue samples (Shami et al. 2019). But automated imaging of large tissues is not only possible with deconstructive methods. The Automated Tape-Collecting Ultramicrotome (ATUM) (Baena et al. 2019) is a fully automated serial-sectioning method which enables automated imaging of large tissues in a non-destructive manner and was first used to image a mouse brain (Kasthuri et al. 2015).

A highly automated FIB-SEM setup was used at the HHMI Janelia Research Campus to acquire the largest synaptic level connectome of a large portion of the fly brain (Scheffer et al. 2020). The reconstructed connectome contains about

25′000 neurons and it took about 3 months to image. Usually, scanning speeds used in SEM connectomics projects are in the range of 0.5-4MHz and imaging a 8000×8000 pixel image at 1MHz takes about one minute. Hence, methods to improve imaging speeds are required to make SEM more time and therefore cost efficient.

One way to improve imaging speeds is to capture more electrons with improved detectors. Another highly technical and expensive approach is the use of a multi-beam SEM with 91 parallel electron beams (Crosby et al. 2016). A cost efficient alternative would be to simply increase the scanning speeds of established SEM setups. Unfortunately, this means trading high signal-to-noise ratio (SNR) for scanning time, which means that the acquired images become significantly more noisy and downstream processing becomes more difficult or even impossible.

In 2018 Weigert *et al.* used content-aware image restoration (CARE) methods to restore low SNR fluorescence microscopy data (Weigert et al. 2018). They achieved astonishing results by training deep neural networks with acquired image pairs of low- and high-quality. This so called supervised training approach teaches a network to uses image context to predict a clean denoised image from a noisy observation. In this thesis I will show how these supervised CARE techniques from light microscopy can be directly applied to SEM image data. By training such CARE networks we can increase scanning speeds of SEM setups 40- to 50-fold and restore high quality images digitally from fast scanned noisy observations.

1.2 Cryo Transmission Electron Microscopy

While SEM is often the method of choice for connectomics, cryo transmission electron microscopy (cryo TEM) is the preferred approach in structural biology. Cryo TEM belongs to the family of cryo electron microscopy (EM) methods, for which Jacques Dubochet, Joachim Frank, and Richard Henderson were awarded the Nobel Prize in Chemistry in 2017. In cryo TEM, or in cryo EM in general, the sample is rapidly frozen to cryogenic temperatures by plunge freezing (Dobro et al. 2010). During this rapid freezing process the water can not assemble into crystalline ice but embeds the sample in vitreous ice (Jacques Dubochet et al. 1988). From a biological perspective using water as sample substrate is desirable. This enables imaging of lipid complexes which usually collapse during dehydration. Furthermore the samples do not need to be chemically fixed *i.e.* the biological structures are literally frozen in time and not altered (Bhella 2019).

In cryo TEM the electrons pass through the sample. If they interact with

the atoms in the sample they become elastically scattered and are affected by a phase shift. Others, which we call undeflected, pass just through the sample without interaction and are not affected by a phase shift. The cryo TEM image is then formed by phase-contrast between the elastically scattered and undeflected electron wave. The observed image is modified by the contrast transfer function (CTF), which is the Fourier transformed point spread function (PSF), of the optical system. The PSF describes how a point light source is deformed by the optical system. The CTF for in-focus images in cryo TEM attenuates low resolution frequencies, which makes identification of molecules impossible due to low contrast between foreground and background. Hence, cryo TEM practitioners use defocus to increase phase-contrast in lower frequencies. However, this trades low frequency contrast at the cost of high frequency information (Bhella 2019). Furthermore, cryo TEM images suffer from low SNR because of beam induced sample damage. Unfortunately, electrons interact negatively with organic material, which leads to denaturation of the samples while they are imaged, hence only a total of $100e^-/\text{Å}^2$ to $120e^-/\text{Å}^2$ can be used for cryo TEM imaging (Knapek and Dubochet 1980). In summary cryo TEM experts have to choose appropriate defocus and electron dose during the imaging process and are still left with rather low contrast and low SNR images.

Single particle analysis (SPA) and cryo tomography are two approaches that use cryo TEM to answer different questions in structural biology. SPA enables us to resolve individual molecules at near atomic resolution and cryo tomography allows us to look into biological samples at high resolution. Both approaches are extremely important in modern structural biology and we will discuss in the next two sections how both of these approaches deal with low contrast and low SNR.

1.2.1 Single Particle Analysis

In single particle analysis (SPA) a single type of molecule is replicated and purified in solution. The solution is then applied to a cryo TEM sample grid, which is plunge frozen and inserted into the microscope. From which a single transmission image is acquired. Due to the purification process it is given that each object – each density – in the image belongs to the same particle class. However, each particle has its own random orientation, which enables 3D tomographic reconstruction of the particle.

During tomographic reconstruction of the molecular complex, also called sub-tomogram averaging, the individual densities are identified/picked (Bepler, Morin, et al. 2019; Voss et al. 2009; Wagner et al. 2019), aligned and averaged together. The central (also Fourier) slice theorem (Bracewell 1956) relates the 2D Fourier

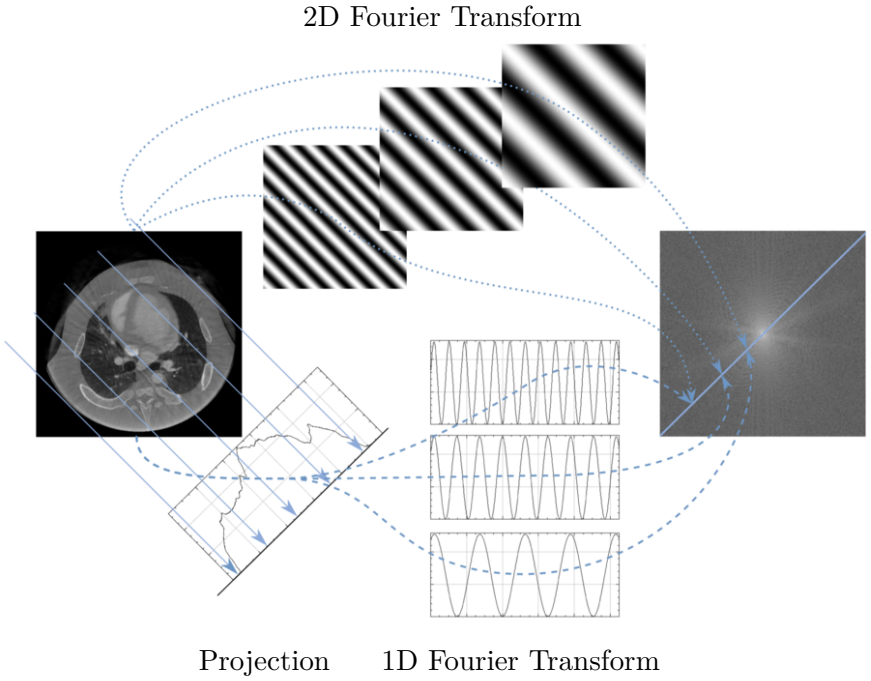

Figure 1.1: The Fourier (or central) slice theorem relates the 1D Fourier transform of a projection to the central slice, which is perpendicular to the projection direction, in 2D Fourier space of the imaged sample. This image is taken from the text book (Maier et al. 2018).

transformation of a 2D projection image to a slice in the 3D Fourier space of the sample. More specifically, the 2D Fourier transform corresponds to the slice in 3D Fourier space, which is perpendicular to the projection direction. This relation holds also for 1D projections and 2D reconstructions and is illustrated in Figure 1.1. As a consequence of this, it is given that two random 2D projection images of the same particle class have a single line in common (the common line), where the two Fourier transformed slices cross in the reciprocal space of the 3D sample. By finding these common lines of the individual Fourier transformed 2D projections it is possible to compute a particle alignment and generate a 3D reconstruction from many (10′000 to 100′000) random 2D projections (Jonić et al. 2008).

RELION (Scheres 2012) and EMAN2 (Tang et al. 2007) are two widely used software packages for particle alignment and sub-tomogram averaging in SPA. Another recent development in the field is spearheaded by Dimitry Tegunov with his two contributions WARP and M. WARP is a real-time preprocessing pipeline for cryo TEM data (Tegunov and Cramer 2019) and M is a novel particle refinement framework which uses a deformation model to correct for optical aberrations (Tegunov, Xue, et al. 2021). Another extremely interesting work is cryoDRGN by Zhong *et al.* where they use neural networks to reconstruct flexible *i.e.* heterogeneous particles and are able to interpolate between different confor-

mational states (Zhong et al. 2021). The advances in digital post-processing of SPA data combined with ever better imaging hardware allow reconstruction of molecular complexes at near atomic resolution. And in 2020 the first atomic resolution results were reported, where individual atoms are visible in the particle reconstruction (Nakane et al. 2020). However, we can not observe these molecules in their native environments, let alone investigate interactions between different molecules.

1.2.2 Cryo Tomography

Cryo electron tomography, on the other hand, allows us to image complete biological systems. Although, samples up to 250µm can be plunge frozen, only thin samples (up to 400nm thickness) can be transmission imaged (Golding et al. 2016). Above 400nm the electron beam will not be able to penetrate (Gan and Jensen 2012). From such a sample a tilt-series is acquired by rotating the sample from *e.g.* -60 to 60 degrees in 2 degree steps. Rotations below and above ±60 degrees are usually infeasible, because of relative increasing sample thickness and sample holder geometry. Note, that the dose per tilt-angle is much lower than the total dose of a SPA image, therefore individual tilt-series images have a much lower SNR compared to SPA. Hence, any tomographic reconstruction from such noisy images suffers from low SNR as well.

Cryo tomography experts use binning and classical filtering techniques like non-linear anisotropic diffusion (NAD) (Frangakis and Hegerl 2001) to enhance image quality. These steps are necessary to enable visual inspection of tomographic data in tools like IMOD (Kremer et al. 1996). Furthermore, since we deal now with extremely crowded environments, automated picking pipelines as they are used in SPA are not necessarily applicable anymore. And more often than not, structural biologists working with cryo tomography have to hand pick individual particles of interest to perform sub-tomogram averaging. However, the aforementioned filtering techniques are rather weak compared to modern content-aware deep learning solutions and require delicate hyper-parameter tuning to work best for given frequency ranges.

In this thesis I will apply SOTA deep learning solutions to cryo tomographic data. Providing a novel approach to tomographic image restoration, which optimizes denoising over a wide range of structural sizes, hence enabling cryo tomography practitioners to see more in their data. In the next section we will discuss tomographic reconstruction in general and have a closer look at the artefacts which can occur.

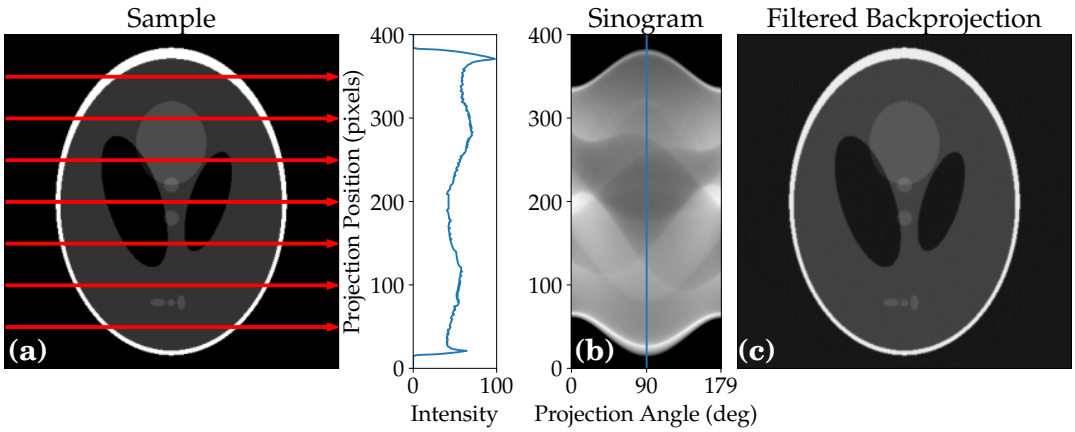

Figure 1.2: In tomography a detector is rotated around a sample and transmission images are acquired at defined acquisition angles. The red arrows, in Subfigure **(a)** correspond to a transmission image acquired at a 90° angle. For a 2D sample we will acquire 1D intensity measurements, see intensity plot. All acquired 1D projections are usually arranged in a sinogram (see Subfigure **(b)**), where each column corresponds to an acquisition angle. From a sinogram we can then reconstruct the 2D image via Filtered Backprojection (FBP), also called inverse radon transform (Kak et al. 2002; Ramesh et al. 1989) shown in Subfigure **(c)**.

1.3 Tomographic Reconstruction

Tomographic reconstruction is used to restore a 3D image from a series of 2D projections or a 2D image from a series of 1D projection images. Formally, tomographic reconstruction is the inverse transformation of the Radon transform. The Radon transform (Kak et al. 2002; Radon 1917) is obtained by either rotating a detector around a sample, or by rotating the sample, and acquiring a series of density measurements at defined projection angles α_i. In cryo tomography, as we have seen above, the sample is rotated and a so called tilt-series consisting of multiple 2D projections is acquired. A wider known application is computed tomography (CT) from bio-medical imaging, where a 1D detector array is rotated around the sample, acquiring a series of density measurements. In CT these density measurements are often visualized as sinogram, where each column corresponds to a single 1D projection see Figure 1.2. Like in cryo tomography, where total electron dose is a limiting factor, total radiation dose for the patient is a safety concern in medical imaging.

Reducing total radiation dose can be achieved in two ways: (i) reducing dose per projection or (ii) reducing number of projections. The first approach leads to increased noise levels in the measurements, which results in noisy tomographic reconstructions see Figure 1.3 (b). As for the second approach, reducing the number of projection angles results in tomographic reconstruction artefacts see Figure 1.3 (c) and (d). These reconstruction artefacts are linked to an undersam-

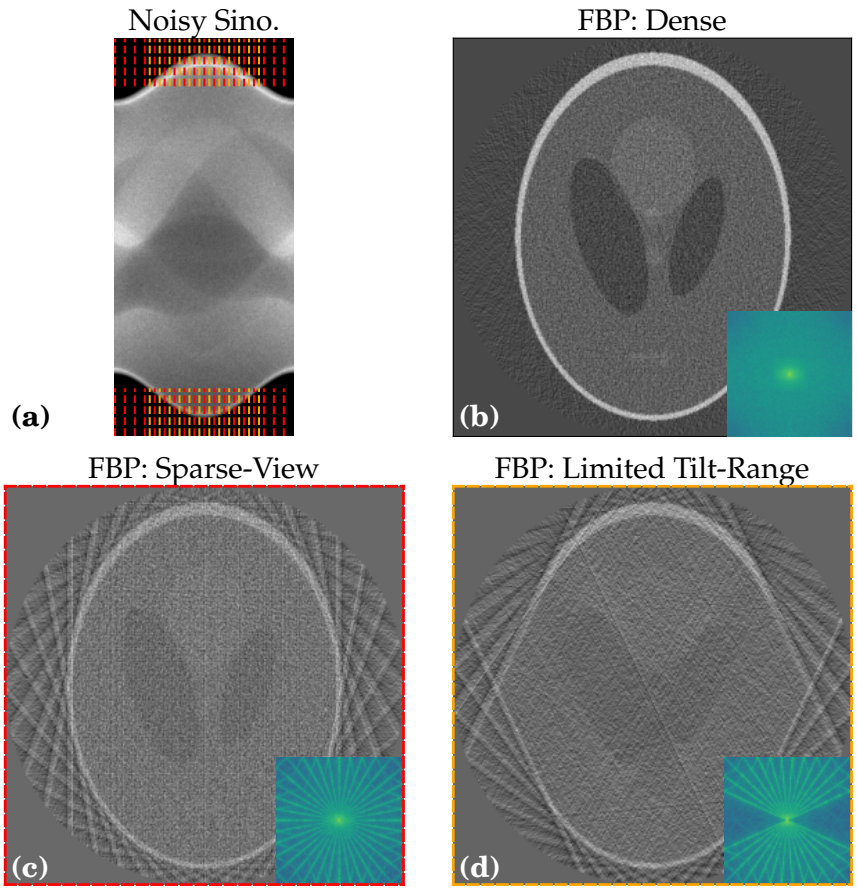

Figure 1.3: Subfigure **(a)** shows the sinogram from Figure 1.2(b) with additional Poisson noise. The observed noise leads to a noisy FBP reconstruction as shown in Subfigure **(b)**. Subfigure **(c)** shows the FBP reconstruction from a sparse-view acquisition, where only the projections indicated by the red dashed lines in (a) were used. Subfigure **(d)** shows a tomographic reconstruction from a sinogram which was acquired with a limited tilt-range (yellow dashed lines in (a)), as it is the case in cryo tomography. The insets in (b-d) show the Fourier spectra of the reconstructions with visible missing wedges in (c) and (d). These missing wedges are the reason for the streaking artefacts in image space.

pled Fourier space as described by the Fourier slice theorem (see Figure 1.1. The Fourier slice theorem states that the Fourier coefficients of each 1D projection at a given angle α_i coincide with the 2D Fourier coefficients that lie on the line that crosses the DC component at angle α_i (Bracewell 1956). Hence, a sparse sampling of projection angles leaves many 2D Fourier coefficients unobserved, leading to so called missing wedge artefacts see Figure 1.3. Note, the same is true for 3D tomographic acquisitions as they are used in cryo tomography. There exists a plethora of work that deals with different combinations of noise and number of projections, some of them we will discuss shortly. Many of these techniques were developed for biomedical imaging and some of them can directly be applied to cryo tomography, while others are infeasible due to increased runtimes for 3D reconstructions

or because of image size. Most modern methods try to cope with less dose and fewer projections, while keeping reconstruction quality high, which would allow practitioners to reduce radiation further. In the following paragraphs we will go over some of the different tomographic reconstruction approaches.

The most common tomographic reconstruction method is filtered backprojection (FBP) or inverse radon transform (Kak et al. 2002; Ramesh et al. 1989). Each projection image can be interpreted as the sum over all intensities along the projection direction. To obtain a reconstruction the measured intensity is then replicated along the projection direction. This is repeated for each projection angle and the final restored image is the sum over all of these "backprojected" projection images. Simply doing this leads to a reconstruction which is blurred by the point spread function $\frac{1}{r}$, where r corresponds to the distance of the current pixel from the projection rotation center. This blur can be reduced by Fourier filtering the backprojected image *i.e.* deconvolving the reconstruction. To speed up the reconstruction process, the Fourier filtering can be applied first to the lower dimensional projections before backprojecting them. This improves reconstruction time without compromising reconstruction quality, hence the name filtered backprojection. Many different filtering approaches (Ramp, Shepp-Logan, Cosine, Hamming, Hann) exist to deal with varying amount of noise. However, filtered backprojection is not able to reduce missing wedge artefacts.

The missing wedge artefacts are directly related to missing information in the Fourier space (see Figure 2.6 (c) and (d)), so any method trying to remove these artefacts has to be able to generate this information in some way. Classical reconstruction algorithms achieve this by optimizing

$$\underset{\hat{y}}{\operatorname{argmin}}\, \mathcal{D}(x, \Phi(\hat{y})) + \lambda \mathcal{R}(\hat{y}), \tag{1.2}$$

where $\hat{y}$ is the reconstructed image, Φ is the radon transform *i.e.* the forward projection operation, x is the observed sinogram (or tilt-series), $\mathcal{D}$ is the data term, $\mathcal{R}$ is a regularizer and λ is a hyper-parameter controlling the relative weight of regularization versus data affinity. A common data term is the L_2-Norm and classical regularizers are Total Variation (TV) (Rudin et al. 1992) or Total Generalized Variation (TGV) (Bredies et al. 2010). Today, these engineered regularizers are also replaced by shallow neural networks (Adler and Öktem 2018; Hauptmann et al. 2018). Optimizing such objective functions is done with iterable solvers (Chambolle and Pock 2011; Gilbert 1972; Gordon et al. 1970) and high quality reconstructions can be generated from a few noisy projections. However, due to their iterative nature these solvers are comparably slow and tuning the regularization weight λ is non-trivial.

More recent approaches use deep neural networks as post-processing of FBP. To train such a network in a supervised fashion pairs of low quality reconstructions, obtained by standard FBP, and high quality ground truth images are required to train the deep neural network (H. Chen et al. 2017; Jin et al. 2017). Once the network is trained reconstruction and artefact removal is fast and exceptional results can be achieved. However all these reconstruction methods deal with missing wedge artefacts in real image space *i.e.* they use regularizers or post-processing to smooth out or remove these artefacts in the reconstruction. But we know, from the Fourier slice theorem, that these artefacts originate in the Fourier space where certain frequencies are just not present. This lead me to investigate if these artefacts can be removed by directly predict missing Fourier coefficients in Fourier space. In the Chapter 4 of this thesis we will investigate this line of thought and with Fourier Image Transformers (FITs) a novel tomographic image reconstruction approach working directly in Fourier space restoring missing Fourier coefficients directly is presented, with it eliminating reconstruction artefacts.

1.4 Overview and Contributions

In this thesis I will present at multiple approaches to image denoising without ground truth data and present a novel idea to approach tomographic reconstruction. In Chapter 2, I will demonstrate how CARE methods for denoising can be applied to electron microscopy data. In particular, I present a way to speed up image acquisition in SEM by using fast and slow scanned image pairs to train supervised CARE models. Furthermore, I present cryoCARE, a Noise2Noise (Lehtinen et al. 2018) trained CARE denoising approach for cryo tomograms, where acquisition of paired low- and high-quality training data is impossible. In Chapter 3, we will take a step back from specific imaging modalities like cryo TEM and look at image denoising in general. I will introduce Noise2Void, the first self-supervised image denoising method based on deep learning. Noise2Void allows us to train content-aware image denoising networks with only single noisy observations. However, Noise2Void can not be applied to reconstructed tomograms and denoising the projections leads to enhanced reconstruction artefacts. In Chapter 4, I will address tomographic reconstruction artefacts namely the missing wedge artefact. We will turn toward Transformer based architectures. Transformers are currently setting new gold standards in virtually all natural language processing (NLP) tasks (Devlin et al. 2018; Radford et al. 2018). In the future, I expect that an increasing number of computer vision tasks will be solved with Transformers and I present such methods. Unlike convolutional neural networks Transformers are applied to 1D input sequences and it has been shown that Transformers are able to complete pixel sequences corresponding to a flattened image (M. Chen et al. 2020; Katharopoulos et al. 2020; Parmar et al. 2018). Here I propose a novel 1D image

encoding which is based on the Fourier transformation of the image. In particular, the proposed Fourier Domain Encoding (FDE) fulfills the property that each prefix encodes the full image, however at reduced resolution. This allows to easily train an *auto-regressive* Transformer for image super-resolution. Additionally, I will present an *encoder-decoder* Transformer setup for tomographic reconstruction. Unlike convolutional based post-processing restoration approaches, I aim at filling in the missing wedges directly in Fourier space by restoring the missing Fourier coefficients.

The contributions of this thesis can be briefly summarized as follows:

Chapter 2: Content Aware Image Restoration for Electron Microscopy

- Proposing a content-aware image denoising approach for SEM to enable faster image acquisition.
- cryoCARE a content-aware image denoising approach for cryo TEM tomograms based on Noise2Noise (Lehtinen et al. 2018).
- Open-source implementation of cryoCARE[1] and integration into Scipion[2] (de la Rosa-Trevín et al. 2016), a cryo electron microscopy image processing framework.

Parts of this chapter are published in (Buchholz, Jordan, et al. 2019; Buchholz, Krull, et al. 2019).

Chapter 3: Noise2Void - Self-Supervised Denoising

- Introduction of Noise2Void, a novel approach for training denoising Convolutional Neural Networks (CNNs) that requires only a body of single, noisy images.
- Comparison of Noise2Void trained denoising results to results obtained with existing CNN training schemes (Lehtinen et al. 2018; Weigert et al. 2018) and a non-trained method (Dabov et al. 2007).
- A sound theoretical motivation for Noise2Void as well as a detailed description of an efficient publicly available implementation[3].

Parts of this chapter are published in (Krull, Buchholz, et al. 2019).

[1] https://github.com/juglab/cryoCARE_pip
[2] https://github.com/scipion-em/scipion-em-cryocare
[3] https://github.com/juglab/n2v

Chapter 4: Fourier Image Transformer

- Proposing Fourier Domain Encoding (FDE) a novel sequential image encoding.
- Demonstrating Fourier Image Transformer (FIT) for super-resolution by training an *auto-regressive* Transformer on the FDE.
- Demonstrating FIT for tomographic reconstruction, which resolves missing wedge artefacts directly in Fourier space.
- Open-source implementation of FIT in PyTorch[4].

Parts of this chapter are under review.

In the text of this thesis, the pronoun "we" generally refers to the author and the reader.

[4] https://github.com/juglab/FourierImageTransformer

Chapter 2

Denoising in Electron Microscopy

Contents

Over the last decades, tremendous technological advancements have been made in light microscopy (LM) and electron microscopy (EM). Employing fluorescent light microscopes in workflows, imaging beyond the resolution limit, acquiring image volumes at high temporal resolution, and capturing many hours of video material is now routinely done, which enables imaging of processes in living

cells and tissues that were previously unobservable. Electron microscopes can go far beyond the resolution limit of light microscopy, and modern EM approaches enable us to see cellular building-blocks in their native cell and tissue context. Despite all technological progress, electron microscopy images tend to have a low signal-to-noise ratio (SNR).

As we have seen in Section 1.1 SEM imaging is slow, if we would like to acquire at high SNR. Scanning speeds can be drastically increased, but then high SNR is traded for faster acquisition times. Unfortunately low SNR images make solving downstram tasks hard, for example, when such images are post-processed to obtain a neural connectivity map (connectome) of the brain (Kasthuri et al. 2015). Hence, the scanning speed is a major limiting factor when acquiring large image volumes. On the other hand, we have looked at cryo TEM, which can image thousands of pixels within seconds, much faster than any SEM setup. However, the lack of staining in combination with beam induced sample damage (Knapek and Dubochet 1980) prevents acquisition of high quality, high SNR images in cryo TEM. Both EM methods would greatly benefit from modern state-of-the-art (SOTA) denoising approaches to ease post-processing in cryo TEM and enable faster scanning speeds without compromising downstream processing in SEM. In this chapter I will first discuss multiple image denoising methods from simple fixed filter approaches like mean-filtering to current SOTA methods based on deep neural networks. In Section 2.1 we look at some classical denoising approaches and follow up with modern supervised denoisers in Section 2.2. In Section 2.3 I describe how SOTA deep learning approaches from fluorescence microscopy can be applied to SEM data. This allows to significantly speed up SEM acquisitions without loss in image quality. Then, in Section 2.4, the seminal Noise2Noise work by Lehtinen *et al.* is introduced. In Section 2.5 I will present multiple approaches to denoise 2D cryo TEM data and reconstructed 3D cryo TEM tomograms. Section 2.6 introduces the used and developed open-source software packages. Finally, I close with a discussion in Section 2.7.

Contributions:

- Using content-aware image denoising to enable faster SEM acquistions.
- cryoCARE a content-aware image denoising approach for cryo TEM tomograms based on Noise2Noise.
- Open-source implementation of cryoCARE[1] and integration into Scipion[2], a cryo EM image processing framework.

Parts of this chapter are published in (Buchholz, Jordan, et al. 2019; Buchholz,

[1] https://github.com/juglab/cryoCARE_pip
[2] https://github.com/scipion-em/scipion-em-cryocare

Krull, et al. 2019).

2.1 Image Denoising

The dominant noises in bio-medical imaging are Gaussian and Poisson noise. This is due to the inherent mechanisms behind imaging – while Gaussian noise is induced by analog to digital convertion in the detector, Poisson noise is inherently given by the stochastic observation process of photons or electrons. An important property of Gaussian and Poisson noise is, that they are both zero-centered around the signal *i.e.* the expected value of the noise for a given signal is zero. We can formalize this relationship into

$$\mathbb{E}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{s} + \boldsymbol{n}) = \boldsymbol{s} + \mathbb{E}(\boldsymbol{n}) = \boldsymbol{s} + \mathbf{0}, \tag{2.1}$$

where $\boldsymbol{x}$ represents an image consisting of of the ground truth signal $\boldsymbol{s}$ and a random noise contribution $\boldsymbol{n}$. Hence, the image quality depends on a better estimate of the random noise contribution, which can be achieved by increasing exposure *i.e.* maximising number of captured photons or electrons. Note, that it doesn't matter if the capture is continuous or over multiple individual images over which we can sum later on. But, it is important that the imaged scene is perfectly still and does not move, else motion blur is introduced into the image. Hence, the first step in image denoising is optimizing the imaging protocol to gather as much signal as possible. However, there exist plenty of applications where exposure has to be limited. For example in time-lapse imaging in biology due to phototoxicity, in cryo tomography due to beam induced sample damage (Knapek and Dubochet 1980) or in computed tomography (CT) due to radiation concerns for the patient.

If noise is a limiting factor for down stream image processing, denoising offers different solutions. Simple denoising approaches convolve a noisy image with a fixed kernel. The convolution operation computes for each output pixel the weighted sum over the input pixels, where the weights for each pixel are given by the kernel. By increasing the size of the kernel, the receptive field *i.e.* the amount of information aggregated by the filter is increased. However, by increasing the kernel size we trade resolution for signal.

This trade-off can be nicely observed if the convolution is performed via Fourier space. A relationship described by the convolution theorem, which states that a real image space convolution corresponds to an element-wise multiplication in Fourier space between the Fourier transformed kernel and image. In Figure 2.1 we can see that the Gaussian filter attenuates the high frequencies in the Fourier domain, which leads to a noise reduction in the image domain. This makes sense,

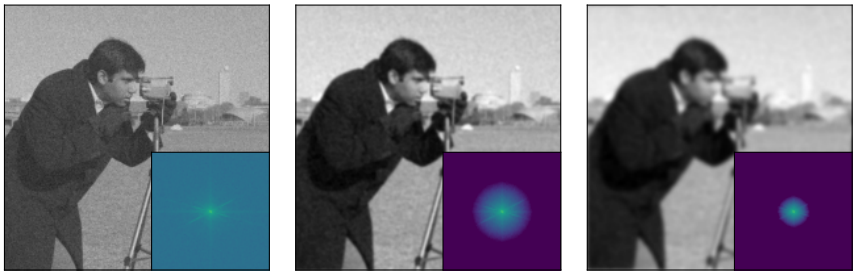

Figure 2.1: The left most image shows a noisy version of the cameraman image and the inset is the Fourier transformation of it. The center image is a Gaussian filtered version of this image with $\sigma = 1$ and the right image is Gaussian filtered with $\sigma = 2$. The insets show the Fourier transformation of the image. In the Fourier transform the low frequency information is displayed in the center and each concentric ring contains higher frequency information. The high frequencies are most affected by pixel-wise independent noise, hence applying a Gaussian filter which attenuates high frequencies (middle and right subfigure) is able to suppress pixel-wise independent noise. However, all frequencies are attenuated equally, also high frequencies responsible for sharp edges, hence the denoised images become blurrier with larger σ.

since noise is random for each pixel and therefore captured by high frequencies in the Fourier spectrum. However, by increasing the kernel size also lower frequencies get attenuated and eventually zeroed out. This overall loss in high-frequency information manifest as blur in the denoised image.

Using a single fixed kernel on images with content of different scales and shapes is suboptimal. A single convolution-based filter can only be optimized for a single given structure and will introduce blur artefacts in all other regions of the image. Anisotropic diffusion is an advanced image denoising algorithm which tackles this problem by introducing an edge detection component, which tunes down the filter response near edges to keep them sharp and contrasted (Frangakis and Hegerl 2001). Other methods like non-local mean filtering (Buades et al. 2005) or BM3D (Dabov et al. 2007) are based on the idea that natural images usually contain a large amount of repeating patterns. They perform internal grouping of similar looking image regions and combine them to produce denoised outputs. Intuitively, these methods denoise similar looking patches by averaging them and build up a denoised image by stitching together all of these denoised patches. However, such iterative and internal statistics based methods have long run times, which makes them cumbersome to apply to large image data.

Generally speaking, all of the above mentioned denoising techniques rely on at least one hyper-parameter *e.g.* the kernel size in the case of a Gaussian filter. These hyper-parameters are non-trivial to optimize and often a lot of domain

expertise is required to get decent denoising results. However, if we would have access to some ground truth signal we could optimize hyper-parameters with respect to some quality measurement like mean squared error (MSE) or peak signal-to-noise ratio (PSNR).

2.2 Supervised Image Restoration

Supervised training of deep neural networks is the current SOTA in image restoration and image denoising. Instead of hand crafting individual filters or priors, deep neural networks are trained with pairs of low- and high-quality images $(\boldsymbol{x_i}, \boldsymbol{y_i})$ combined with a suitable loss function $\mathcal{L}$. We can interpret a deep neural network as a highly parameterized function, which maps an input $\boldsymbol{x_i}$ to an output $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x_i}) = \hat{\boldsymbol{y}}_{\boldsymbol{i}}, \tag{2.2}$$

where $\boldsymbol{\theta}$ are the trainable parameters of the neural network. By employing a loss function

$$\boldsymbol{e} = \mathcal{L}(\hat{\boldsymbol{y}}_{\boldsymbol{i}}, \boldsymbol{y_i}) \tag{2.3}$$

we obtain the error $\boldsymbol{e}$ between the current prediction $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ and the ground truth target $\boldsymbol{y_i}$. This error, also called loss, is then backpropagated (Rumelhart et al. 1986) through the neural network and for each weight the individual error contribution is computed. With gradient descent, we can then iteratively update the neural network weights until convergence. Following this procedure enables us to train the neural network function $f_{\boldsymbol{\theta}}$ to map any input $\boldsymbol{x_i}$ to any output $\boldsymbol{y_i}$. But our hope is to obtain a network which generalizes well to unseen data, for which we *do not* have access to ground truth.

2.2.1 Training and Validation Loss

The best way to get a neural network to generalize better over a given data distribution is to increase the amount of training data pairs. However, a large enough neural network trained until convergence will still be able to learn a perfect one-to-one mapping for all training samples and perform poorly on new unseen test images. Therefore it is necessary to monitor the neural network performance on unseen validation data during training. It is important that validation and training data are disjoint datasets, however they should capture the same data distribution. By plotting the training and validation loss we can learn a couple things about the trained neural network. In Figure 2.2 (a) we can observe that both loss curves are going down over time. These curves contain two bits of information, first validation- and training-loss are close to each other, this means that the training- and validation-datasets are reasonably similar *i.e.* they both

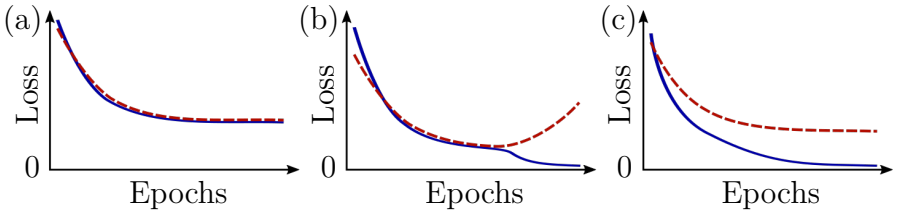

Figure 2.2: Shown are three common loss curve plots with the blue line corresponding to the training-loss and the red dashed line corresponding to the validation-loss. Subfigure **(a)** shows the loss curves of a under-parameterized model, hence the loss curves can not reach 0. In Subfigure **(b)** loss curves of an over-parameterized model are shown. The model is able to capture the data distribution and even overfits in the last third of the training *i.e.* it does not generalize well to the unseen validation data. In Subfigure **(c)** the validation loss is much higher than the train loss, this indicates that the data distribution of the validation dataset differs from the training data distribution.

capture the same data distribution. Second, the losses have flattened out *i.e.* reached a plateau, but are still far away from zero. Flattening loss curves tell us that the training has converged, however the gap towards zero informs us that the neural network has not enough capacity to capture all essential features of the training data. Essentially, the model has used its limited capacity to learn some of the most common features of the data. By increasing the size of the neural network we can increse its capacity and it will be able to learn more specialized features, see Figure 2.2 (b). Where both learning curves are going down initially, and reach a lower plateau compared to subfigure (a). However, towards the end the training-loss suddenly drops again and reaches quasi zero, while the validation loss simultaneously starts growing. This behaviour indicates overfitting of the model to the training data. Overfitting means, that the network has learned a perfect representation of all training samples, however it does not generalize well to the unseen validation data. In such a case we can either reduce the model size, add more training data or we employ the early stopping strategy (Finnoff et al. 1993) and just use the model corresponding to the lowest validation loss. In Figure 2.2 (c) we see that the validation-loss is much higher than the training-loss. This indicates that the validation- and training-datasets capture different distributions and while the model is able to learn all details of the training data, it is unable to generalize to the different validation-dataset. This is a good indicator to investigate the datasets, which potentially leads to an increase in required training- and validation-data. Understanding the losses is important to train deep neural networks, however as LeCun stated in (Y. A. LeCun et al. 2012), efficient training with backpropagation in reality is often a lot harder than it seems. Not every neural network will converge with any optimizer and loss function, which makes training of deep neural networks also an engineering challenge. In the next section, I will highlight the neural network architectures which have been

proposed for the task of image denoising.

2.2.2 Neural Network Architectures

The first convolutional neural network (CNN) used for image denoising was presented by Jain *et al.* in 2009. They interpret the denoising task as a regression task and the CNN is trained to minimize a loss calculated between its prediction and clean ground truth data (Jain and Seung 2009). This basic setup builds the backbone of many SOTA CNN denoisers available today. It is noteworthy to mention that GPUs were becoming commercially available around the same time[3]. Hence, training of deep neural networks started to become feasible.

Zhang *et al.* use a very deep CNN architecture for image denoising in (Zhang et al. 2017). Their architecture, unlike previous architectures, does not predict the clean image, it predicts the noise and the clean image is then computed in a subsequent subtraction step. Essentially, the network computes a residual image, an idea presented by He *et al.* (He et al. 2016).

Around the same time Mao *et al.* presented a very deep encoder-decoder architecture (Mao et al. 2016) for image denoising. They also use residual or skip connections between corresponding encoding and decoding modules *i.e.* information in the forward and backward pass can skip further compression. Note, that these architectures by Zhang *et al.* and Mao *et al.* completely dispense with pooling layers.

In 2015 Ronneberger *et al.* presented the U-Net architecture for the segmentation of neuronal structures in electron microscopy images (Ronneberger et al. 2015), however it got adapted for denoising tasks later on (Weigert et al. 2018). In general the U-Net consists of convolution blocks, down- and up-sampling layers and most importantly skip-connections, a depiction is shown in Figure 2.3. The convolution blocks contain two convolution layers followed by the ReLU (Nair and Hinton 2010) activation function. The first convolution block takes a single channel input image and convolves it with `n_first` different learnable kernels, which results in an intermediate image representation with `n_first` feature maps. These `n_first` feature maps are then passed through the second convolution layer, which applies another `n_first` trainable convolution kernels to produce a new intermediate image representation of `n_first` feature channels. This output is then down-sampled by applying a 2×2 pooling operation (*e.g.* average- or max-pooling). Then the next convolution-block is applied, but this time the number of feature maps is doubled by the first convolution layer. This procedure can be re-

[3] https://web.archive.org/web/20130624034844/https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/

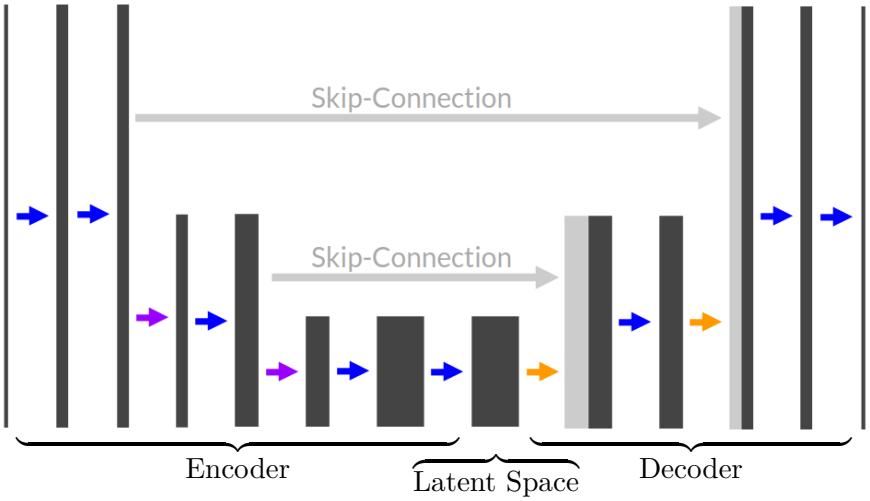

Figure 2.3: A depiction of a U-Net with `n_depth` = 2. The blue arrows are convolution layers, the purple arrows are pooling layers, the orange arrows represent up-sampling layers and the grey arrows are the skip-connections. The first convolution layers on the left side, the encoder, double the number of feature channels. In the decoder, the right side of the U-Net, the first convolution layers half the number of feature channels.

peated arbitrarily often and each pooling operation adds an additional depth level to the U-Net until a predefined depth (`n_depth`) of the U-Net is reached. Then an up-sampling (*e.g.* up-convolution, pixel-shuffle (Shi et al. 2016) or linear interpolation) layer is employed, which doubles the size of each feature channel in X- and Y-dimension while halving the number of feature channels. Now, before the next convolution-block is applied, the feature channels from the same level of the encoding side are concatenated or summed to the up-sampled feature channels, these feature maps have skipped the down-sampling, encoding and compression, hence the name skip-connection. The following convolution-bock halves the number of feature channels with the first convolution *i.e.* the model can choose between the up-sampled feature maps and the feature maps from the skip-connection. This up-convolution followed by convolution-block procedure is repeated until depth zero is reached again. Finally, the remaining feature maps are reduced with a single 1×1 convolution layer to the number of required output channels. In the original U-Net paper the authors used `n_first` = 64, `n_depth` = 5, max-pooling, up-convolutions and added the feature channels from the skip-connections to the up-sampled feature maps. Furthermore, the last convolution reduced the 64 feature maps to 2 outputs – foreground and background (Ronneberger et al. 2015).

The skip-connections are essential for the U-Net, because they allow the network to let high-resolution details flow from the encoder to the decoder, without being compressed. The encoder of such a network, compresses the input image

into a high dimensional latent space representation, however each latent space feature map is much smaller than the original input image. The decoder inverts this operation. Given a latent space representation the decoder builds up a full image. In some sense encoder and decoder are highly specialized filter-banks. The encoder filter-bank decomposes an image into a compressible representation and the decoder filter-bank uses these compressed signals to choose the appropriate combination of filters to restore a full image. Encoder-decoder networks tend to learn low-resolution features first and gradually restore higher resolution features. However, with limited capacity these encoder-decoder networks are not able to fully restore the original image due to missing high-resolution features. This is where the skip-connections of the U-Net prove advantageous, because they by-pass the compression and the network can learn to integrate or ignore high-resolution features. Because of its simplicity the U-Net architecture got quickly adapted in bio-medical image processing and is one of the most used neural network backbones for bio-image analysis.

A U-Net can be trained for virtual any image-to-image task like segmentation, deconvolution, up-sampling or denoising as long as we have access to pairs of input and target data. In 2018 Weigert *et al.* presented their work on content-aware image restoration (CARE) for fluorescence microscopy images (Weigert et al. 2018). They describe multiple strategies, from simulation to carefully imaged pairs of low- and high-quality images, to obtain training data to train a U-Net. One of these tasks is image denoising, which they achieve by imaging the same sample once with low exposure and a second time with high exposure, resulting in suitable training pairs to train CARE networks. In the following section, I will use this CARE approach from light microscopy and apply it to SEM images with the goal of increasing image acquisition speeds without loss of image quality.

2.3 SEM-CARE

In scanning electron microscopy (SEM) the limiting factor is often acquisition time, leading to long and expensive projects. With SEM-CARE I use CARE to speed up image acquisitions, by restoring fast scanned low SNR images to high quality images which can be used for down stream processing like connectomics. This work is a collaboration with Réza Shahidi and Gáspár Jékely from the Living Systems Institute at the University of Exeter (Exeter, UK).

2.3.1 SEM-CARE Experiments

My collaborators imaged ultrathin sections (30nm) of an EPON-embedded larva of the marine annelid worm *Platynereis dumerilii* using a Zeiss Gemini 500

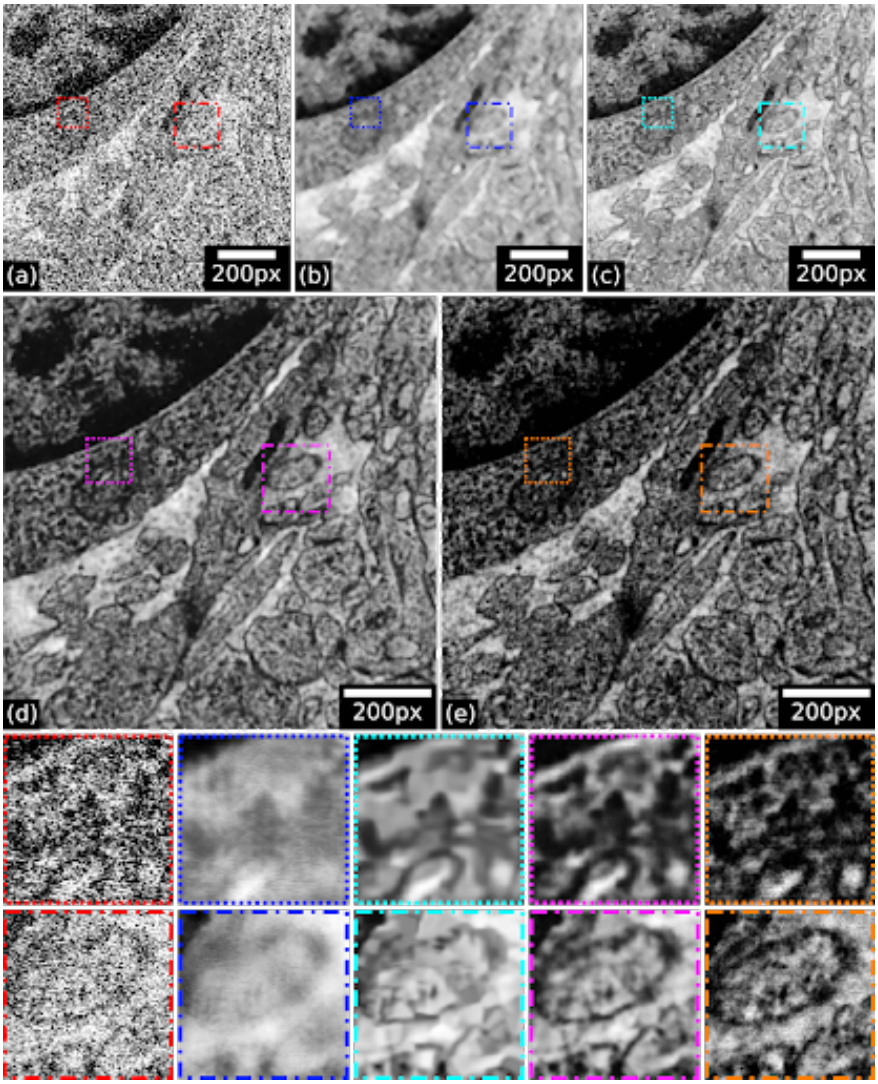

Figure 2.4: Results of SEM-CARE. The upper row of images shows **(a)** the noisy input image (scanned at 5 MHz), and two baseline denoising methods, namely **(b)** Non-Local Means and **(c)** BM3D. The second row of images shows **(d)** SEM-CARE results, and **(e)** the ground-truth, i.e. an average of 4 scans at 0.2 MHz. The remaining two rows show the insets of (a-e) in respective order, additionally indicated by color and line-style.

SEM. Platynereis is an ideal specimen for whole-body connectomics, because of the transparent embryos, synchronous fertilization of many eggs and deterministic/stereotypic development so variance between multiple individuals is low in terms of position of nuclei (Fischer et al. 2010). The collaborators collected sections as ribbons on conductive ITO glass (Pluk et al. 2009). For post-staining, a solution of uranyl acetate and lead citrate was used. To train a CARE image denoising network, pairs of low- and high-quality ground truth images are required. The ground truth images were obtained by scanning the same region 4 times at 0.2MHz and averaging the four images together. The corresponding low-quality

image was scanned at 5MHz. The high speed scanned images suffered from severe noise and could not be used in their raw form for downstream image analysis *e.g.* connectome tracing.

I have used this data to train a CARE denoising model to restore fast scanned SEM images to sufficient quality, such that downstream image analysis is possible again. To this end I had to ensure that the training data is pixel-perfect aligned, else the CARE network can not learn an exact mapping from noisy observations to clean observations, which will manifest as blur in the denoised images. Luckily, image registration is a well understood problem and many powerful methods are readily available (Klein et al. 2009; Schindelin et al. 2012; Thevenaz et al. 1998). In this work I used the free Fiji plugin StackReg (Thevenaz et al. 1998) to align the low- and high-quality images pixel perfect.

Then to train CARE, I extracted 32′768 randomly positioned image patches of size 128×128 from a total of 8 images (jointly counting 471 megapixels). No additional patch augmentation was used. From the extracted patches I used 10% as validation data and trained a default CARE denoising network with `n_depth` $= 2$, 5×5 kernels, and a linear activation function in the last layer. A batchsize of 16 and an initial learning rate of 0.0004 was used. Further the mean absolute error (MAE) was used as loss function. The best performing network on the validation set is evaluated on the test data.

2.3.2 SEM-CARE Results

After training, one fast-scanned, low-quality image which was excluded from the training set was restored. Additionally, I used non-local means (Buades et al. 2005) and BM3D (Dabov et al. 2007) as baselines, two self-supervised denoising algorithms. All results and the corresponding ground truth (slowly scanned and averaged image of the same sample) are shown in 2.4. In addition, I summarize computed PSNR and SSIM (Zhou Wang et al. 2004) (higher is better) values for all baselines and our CARE results in Table 2.1.

	PSNR	SSIM
Input (5MHz)	6.62	0.09
NLM	9.25	0.16
BM3D	9.41	0.37
SEM-CARE	16.56	0.47

Table 2.1: Quantitative measurements comparing the restoration results of the 5MHz SEM acquisition restored with Non-Local Means (NLM), BM3D, and CARE to the 0.2MHz 4 times average ground-truth SEM acquisition.

2.4 Noise2Noise

So far we have looked at CARE networks which required pairs of low- and high-quality images. A neural network is trained by using the low-quality image as input and we ask the network to predict a high-quality output. Naturally, a randomly initialized network will produce some random output far from the high-quality image. But we can compute the error between the current prediction and the high-quality image. This loss is then backpropagated through the network and the neural network weights $\boldsymbol{\theta}$ are gradually updated, optimizing

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{y_i} = f_{\boldsymbol{\theta}}(\boldsymbol{x_i}), \boldsymbol{s_i}) \tag{2.4}$$

where $\mathcal{L}$ is a suitable loss function computing the error between our neural network output $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ and the ground truth $\boldsymbol{s_i}$. Once this process converges we can feed different low-quality images, obtained with the same optical setup and of similar structures, through the network and will obtain high-quality images.

A seminal discovery was presented by Lehtinen *et al.*, which removes the ground truth requirement for supervised denoising tasks. They show that the ground truth observations $\boldsymbol{s_i}$ can be replaced by noisy observations $\boldsymbol{y'_i} = \boldsymbol{s_i} + \boldsymbol{n'}$ as long as $\mathbb{E}(\boldsymbol{y'_i}|\boldsymbol{x_i}) = \boldsymbol{s_i}$, which requires the noise to have an expected value of $\mathbf{0}$ given the signal. Crucially this condition holds true for Gaussian and Poisson noises, the two most common noises in imaging, which means that we don't have to carefully acquire high quality images corresponding to ground truth but a second noisy observation is sufficient. This means that supervised training for denoising can be done with pairs of noisy images $(\boldsymbol{x_i}, \boldsymbol{x'_i})$, hence the name Noise2Noise. The only requirement is, that $\boldsymbol{x_i} = \boldsymbol{s_i} + \boldsymbol{n}$ and $\boldsymbol{x'_i} = \boldsymbol{s_i} + \boldsymbol{n'}$ have independent noise contributions $\boldsymbol{n}$ and $\boldsymbol{n'}$. In practice this means that acquiring two independent images, either in quick succession or with two cameras would provided all required training data to train a content-aware denoising network with the mean squared error (MSE) loss. And even though we are attempting to learn a mapping from a noisy input to a noisy target, the training will still converge to the correct solution. Intuitively we are asking the neural network to do the impossible, to predict a random noise value given an independent random input value and since the expected value of the noise is $\mathbf{0}$, the best guess *i.e.* the prediction is $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{s_i} + \mathbf{0}$ the underlying ground truth signal (Lehtinen et al. 2018).

In some sense the Noise2Noise approach is already indirectly used in (Weigert et al. 2018). By acquiring high quality target images for training a second "noisy" observation of the sample is acquired, just that the noise in this case is not detectable anymore. Naturally, it is beneficial to have as much information as possible available in the training data *i.e.* reducing the noise in

the network input and target will always yield better results.

2.5 cryoCARE

With SEM-CARE I demonstrated the use of CARE denoising networks by acquiring pairs of low and high signal-to-noise ratio (SNR) images, which is possible due to the SEM staining. In cryo-TEM this is not possible anymore. Even the most optimized protocols produce very noisy, low contrast images. Hence, for cryo-TEM data a combination of CARE and the Noise2Noise idea from (Lehtinen et al. 2018) enable us to train cryoCARE networks. It has to be mentioned that Dimitry Tegunov has also implemented Noise2Noise based denoising in Warp (Tegunov and Cramer 2019) as well as Bepler *et al.* in the Topaz-Denoiser (Bepler, Kelley, et al. 2020). Also Palovcak *et al.* implemented a Noise2Noise based cryo TEM denoiser and additionally investigated subtomogram averaging of denoised particles (Palovcak et al. 2020).

In the following subsections I will present various ways to use the Noise2Noise idea to denoise single cryo-TEM projections or fully reconstructed cryo tomograms. The feasibility of cryoCARE is demonstrated on two datasets. The first dataset, called Tomo110, was acquired by Mareike Jordan from the lab of Gaia Pigino on a 300 kV Thermo Fisher cryo TEM Titan Halo with a Gatan K2 direct electron detector. This detector can acquire images in dose-fractionation mode (movie mode), which enables us to test all proposed cryoCARE variations. The Tomo110 data is available via our example notebooks on GitHub[4]. The second dataset, EMPIAR-10110[5], is publicly available via the EMPIAR database (Iudin et al. 2016) and consists of a complete tomographic series of tilted projections.

All presented experiments were performed with the open-source CSBDeep framework by Uwe Schmidt and Martin Weigert (Weigert et al. 2018). More specifically a U-Net (Ronneberger et al. 2015) with `n_depth` = 2, a convolution kernel size of three, and a linear activation function at the last layer is used. Moreover a per-pixel mean squared error (MSE) loss is employed. In all experiments 10% of extracted training data is used as validation data only.

2.5.1 Restoration of cryo TEM Projections

Here I describe three ways to train cryoCARE networks on adequately prepared pairs of cryo TEM projections.

[4] https://github.com/juglab/cryoCARE_T2T
[5] http://dx.doi.org/10.6019/EMPIAR-10110

Training using acquired image pairs: The most straight forward way to combining CARE (Weigert et al. 2018) and Noise2Noise (Lehtinen et al. 2018) is to acquire pairs of images for which the noise is independent. To this end, Mareike Jordan acquired such image pairs on a 300 kV Thermo Fisher cryo-TEM Titan Halo that is equipped with a K2 direct electron detector from Gatan. More precisely, she acquired images of *Chlamydomonas reinhardtii* cilia in dose-fractionation mode (movie mode) (Li et al. 2013). The acquired frames were then splitted in two halves, and averaged without additional alignment, resulting in the equivalent of two independently acquired images at half the available electron dose each[6]. From such pairs of images I extracted 1000 randomly selected patch-pairs of size 128×128 which are used to train a CARE network with the Noise2Noise regime. After training, I used the trained network to restore all image pairs and retrieve the final result by pixel-wise averaging the two individual restorations (see Fig. 2.5 (d)).

Training using tomographic tilt-angle pairs: For readily acquired, not dose-fractionated data, the previously described scheme cannot be applied. Archived data for which only single acquisitions exist can therefore not be used for training cryoCARE networks. For existing tilt-series, acquired for tomographic reconstruction, I asked myself if pairs of neighboring tilt-angles could be used for training. I used IMOD (Kremer et al. 1996) to align and register all acquired tilt-angles. As before, training was performed on 1000 randomly selected patch pairs of size 128×128 taken from adjacent tilt-angle projections. Final restorations are retrieved by applying the trained network to both tilt-angles individually followed by pixel-wise averaging (see Figure 2.5 (c)).

Training using dose-fractionated movie frames: Since the data was acquired on a Gatan K2 direct detector, I was able to go an additional step further. Instead of using two acquired images, as described initially, I can leverage the fact to have many more frames acquired. As it is usually done during dose-fractionation, I additionally corrected for motion-blur of the sample by registering the individual frames using MotionCor2 (Zheng et al. 2016). Then I sum all even and odd frames to retrieve two images with independent noise. This interleaved frame-splitting is advantageous because induced beam damage will be equally shared in both independent images. Again I trained on 1000 randomly selected patch pairs of size 128×128, and created the final restored projection by applying the network to both images followed by pixel-wise averaging (see Fig. 2.5 (e)).

In Figure 2.5 (c) and Figure 2.8 (c) cryoCARE restoration results based on

[6] Note that each image in such a pair has an even lower SNR due to the halved electron dose.

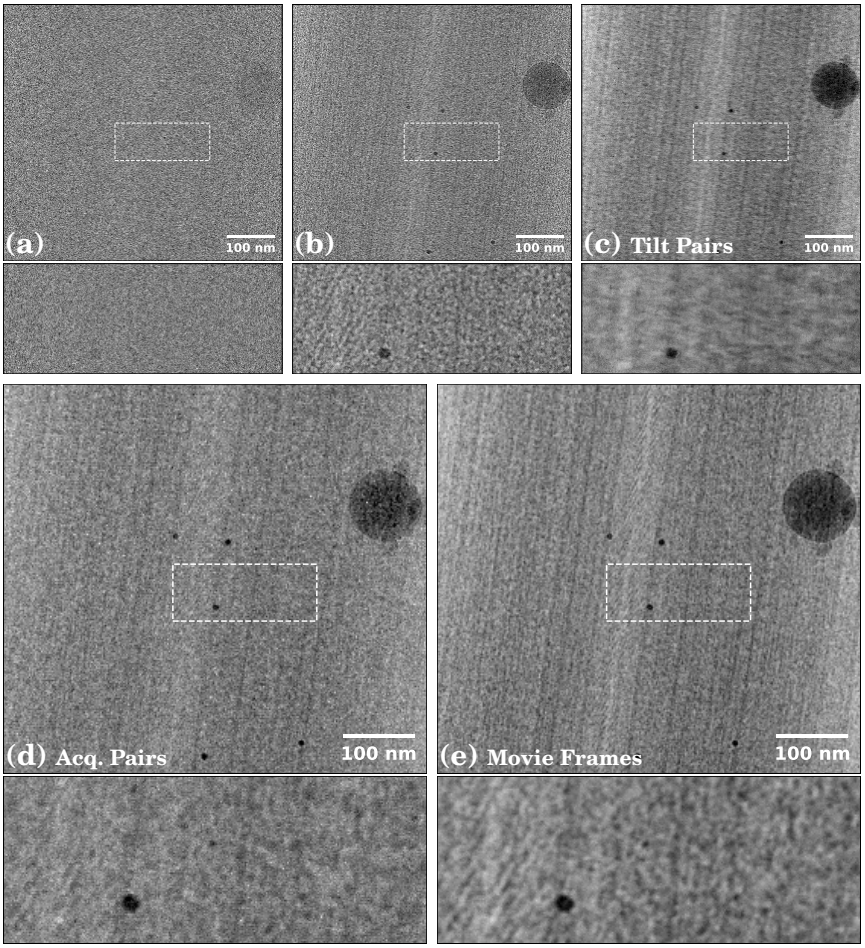

Figure 2.5: cryoCARE results on a 2D cryo TEM projection. Subfigures and insets show: raw input data **(a)**, median filtered restoration baseline **(b)**, cryoCARE results when trained on tomographic tilt-angle pairs **(c)**, on acquired image pairs **(d)**, and on dose-fractionated movie frames **(e)**.

tilt-angle pairs are presented. This approach leads to restored images that appear blurry. This is expected, because the neighboring tilt-angles are not and **can not** be pixel-wise perfectly registered, due to the rotational projection geometry. Hence, the CARE network learns to map slightly displaced image features to one another *i.e.* the closest solution is a compromise between the two structures, which appears blurry.

This problem is circumvented by training cryoCARE on specifically acquired pairs of images or by using dose-fractionated and aligned movie frames. Restoration results of these approaches are shown in Figure 2.5 (d,e).

2.5.2 Restoration of cryo TEM Tomograms

A canonical idea to reconstruct denoised tomograms is to use restored movie frame tilt-angles (like in Figure 2.5 (e)). This does, unfortunately, amplify the

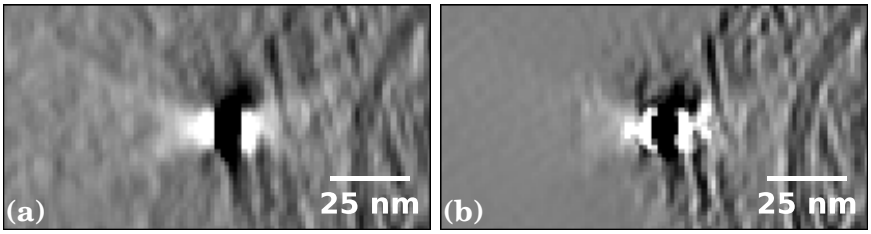

Figure 2.6: Tomogram reconstruction artefacts. Tomograms reconstructed from restored tilt-angles lead to strong missing-wedge artefacts **(a)**. This problem is reduced using the proposed TOMO2TOMO training scheme **(b)**.

missing wedge artifacts at high-gradient locations (see Fig. 2.6). Since neural networks are complex non-linear filters and tilt-angle reconstructions are performed independently, the predicted intensities for a given structure is not necessarily consistent across restored tilt-angles. These inconsistent amplitudes are likely the reason for the amplification of the observed missing wedge artifacts. However, this problem can be addressed with the TOMO2TOMO network training regimes described in the following paragraphs. The TOMO2TOMO approaches work directly on reconstructed tomograms. All tomographic reconstructions were performed with ETOMO, which is part of IMOD (Kremer et al. 1996).

Training using even-odd acquisitions: This protocol is designed to work for conventionally acquired tilt-series, when no direct detector is available. Here all tilted projections are split in two sets based on their acquisition number. From all tilt angles with an even/odd acquisition number, two data-independent tomograms are reconstructed and used to train a 3D CRYOCARE network on $1'200$ randomly selected 3D sub-volumes of size $64 \times 64 \times 64$. The final restored tomogram is obtained by applying the trained network to both tomograms followed by voxel-wise averaging (see Figure 2.7 (d)).

Training using dose-fractionated movie frames: In case the available data was acquired in dose-fractionation mode (movie mode), I propose a slightly different protocol. For each tilt-angle, similar to the cryo TEM projection approach on dose-fractionated data, the frames are aligned, split in even/odd frames and summed. The two sets of independent tilt-angle projections can then be used to reconstruct two independent tomograms. I trained as before and created the final restored tomogram by applying the trained network to both tomograms followed by voxel-wise averaging. The advantage of this approach is that the angular sampling for both tomograms is denser and consistent, hence leading to better results (see Figure 2.7 (d)).

In Figure 2.7 TOMO2TOMO even-odd acquisitions and TOMO2TOMO dose-fractionated reconstructions are compared to the reconstructed raw tomogram

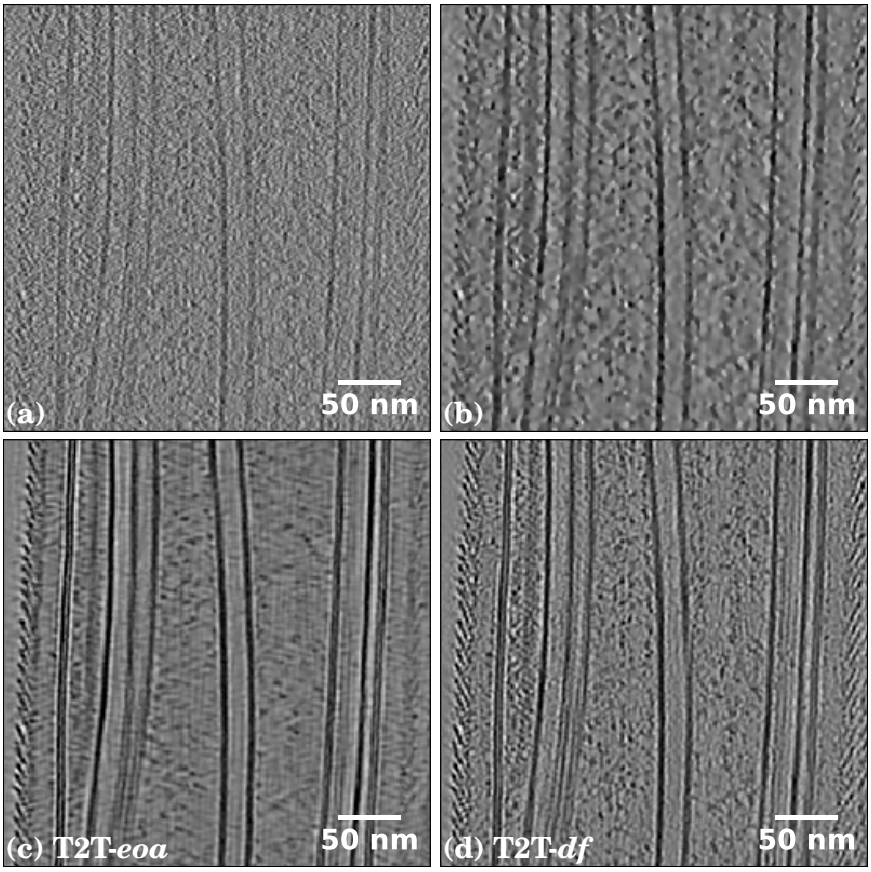

Figure 2.7: cryoCARE results on a 3D cryo TEM tomogram. Subfigures show: a section through the raw tomogram **(a)**, the non-linear anisotropic diffusion filtered baseline **(b)**, cryoCARE results when trained on even- and odd-tilt angle tomograms (T2T-eoa) **(c)** and trained on dose-fracitonated movie frame splits (T2T-df) **(d)**.

and a non-linear anisotropic diffusion (NAD) (Frangakis and Hegerl 2001) filtered baseline on Tomo110. In Figure 2.8 (bottom row) the Tomo2Tomo method using even-odd acquisitions is demonstrated.

2.5.3 Automated Downstream Analysis

In order to test if the restored images are beneficial for downstream analysis of cryo tomograms, I developed the following segmentation and detection workflow. To segment and detect *Chlamydomonas reinhardtii* outer dynein arm (ODA) a U-Net (Ronneberger et al. 2015) was trained on manually created and refined ground truth generated with PEET (Heumann et al. 2011; Nicastro et al. 2006), a sub-tomogram averaging software. The predicted segmentation were then normalized and Otsu thresholded (Otsu 1979). Each connected component was then filtered according to its size in voxels and each remaing component was treated as one detected ODA. Since only a single hand annotated tomogram was available, the

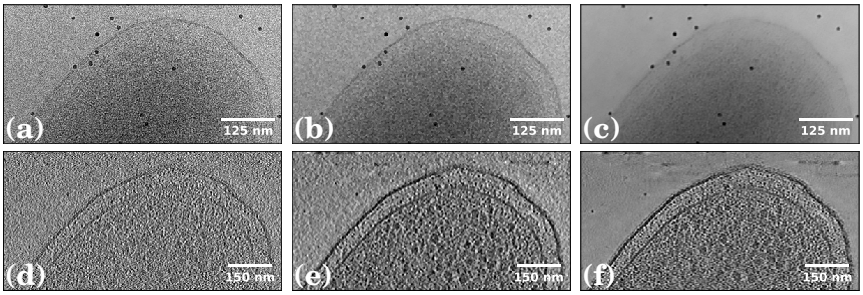

Figure 2.8: cryoCARE restoration on the publicly available EMPIAR-10110 dataset. **(a)** Raw projection (single tilt angle). **(b)** Median filtered baseline. **(c)** The projection restoration results trained with neighboring tilt-angle pairs. **(d)** Raw tomogram. **(e)** NAD filtered baseline. **(d)** The Tomo2Tomo restorations based on even-odd acquisitions.

ground truth annotations were split into 383 training and 712 test annotations. For the neural network training no additional data augmentation was used. The described automated segmentation and detection workflow was applied to the raw and cryoCARE restored data. In Figure 2.9 we can appreciate that cryoCARE is beneficial for automated downstream processing.

2.6 Implementations and Availability

All presented SEM-CARE experiments were conducted with the existing open-source software package CSBDeep[7]. For cryoCARE I developed a custom CSBDeep wrapper and example notebooks with the initial publication. Later on I integrated cryoCARE into Scipion[8] (de la Rosa-Trevín et al. 2016) with the help of Jorge Jiménez de la Morena and Pablo Conesa. As part of the Scipion integration the cryoCARE wrapper became a standalone Python package which is available via pip[9]. The latest development of the cryoCARE package allows users to lazily load training data from multiple tomograms, hence enabling training of more robust and better generalized networks, while keeping the memory footprint as low as possible *i.e.* a normal sized workstation is sufficient to train cryoCARE.

In early 2021 the developer of IMOD (Kremer et al. 1996) has reached out and is interested in integrating cryoCARE directly into IMOD. Unfortunately a direct integration of cryoCARE into IMOD is unlikely to happen, since the deployment of deep learning solutions is still rather complex with respect to different hardware, drivers and required libraries. Nonetheless, the IMOD developers are looking into options which will ease the data preprocessing with IMOD for

[7] https://github.com/CSBDeep/CSBDeep
[8] http://scipion.i2pc.es/
[9] https://pypi.org/project/cryoCARE/

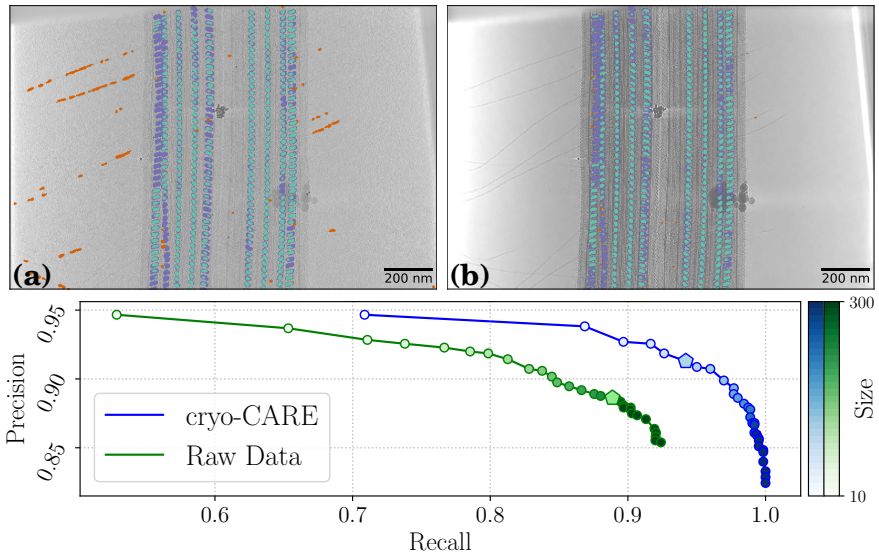

Figure 2.9: Automated downstream analysis on raw data **(a)** and a Tomo2Tomo restored tomogram using the dose-fractionated data approach **(b)**. Ground truth voxels are shown in violet, true-positives in turquoise, and false-positives in orange. Precision-recall plots on increasing segment size threshold (see main text) are shown below. The pentagons correspond to subfigures (a) and (b).

cryoCARE.

2.7 Discussion

In this chapter we have seen how SOTA image restoration techniques can be applied to SEM and cryo TEM data. When using SEM, faster acquisition times are desirable if very large image volumes need to be recorded. My results using SEM-CARE indicate that 40- to 50-fold speed-ups can be achieved without substantial loss in quality. Low-quality acquisitions used in the experiments have been acquired using a 200 times faster scanning speeds than the ground truth images.

In cryo TEM, data is usually heavily filtered with relatively simple filtering techniques like NAD before it is manually investigated. cryoCARE, as we have seen, leads to highly contrasted and well resolved 2D and 3D data. The proposed Tomo2Tomo approach on dose-fractionated movie frames is a simple and powerful tool for content-aware tomographic image restoration.

2.7.1 Tasks Facilitated through cryoCARE

An often overlooked task from non cryo-tomography experts is manual data inspection. However, talking to practitioners immediately reveals the importance of cryo tomographic data visualization. Only by looking at the data it is possi-

ble to build new and confirm hypothesises. Sometimes the structures of interest are extremely rare and hard to see in raw or naively filtered data. This is where cryoCARE can enable easier and less tiring data browsing. The same holds true for manual particle picking tasks, which is often necessary since cryo tomograms are crowded environments where automated picking solutions from SPA are not applicable.

I have also shown that cryoCARE restorations can lead to improved automated segmentation results. An essential feature of cryoCARE is that training data can be generated by the microscope itself and does not require tedious human labeling. While end-to-end pipelines on raw data might need huge amounts of labeled data to also co-learn to restore the noisy data, cryoCARE helps to uncouple these two tasks – a pre-processing step that does not need human labels and a segmentation stage that is likely to require lesser amounts of training data.

Since cryoCARE facilitates particle picking, a canonical question to ask is if cryoCARE processed tomograms can be used for sub-tomogram averaging. It is clear that a single particle instance taken from a denoised tomogram has a better SNR than one extracted from the raw data. However, this advantage quickly diminishes for averages of multiple particles. Palovcak *et al.* report that sub-tomogram averages from CNN denoised particles are only slightly worse than averages of only raw particles (Palovcak et al. 2020). From the Noise2Noise perspective this is to be expected, since the trained model can only learn a single fixed estimate of the denoised particle. The cryoCARE approach does not contain a generative component which could add extra high frequency signal to create higher resolution particles. In other words a cryoCARE network can at best replace each noisy particle with the learned average of all particles in the training data. Hence, each average created from cryoCARE denoised data will at best be equal to a raw data average of the same data.

Finally, I am extremely excited about the integration of cryoCARE into Scipion (de la Rosa-Trevín et al. 2016) and the ongoing developments in IMOD (Kremer et al. 1996) to facilitate data pre-processing. With the integration of cryoCARE into these famous cryo TEM tomography processing packages, I am confident that cryoCARE will be used by an increasing number of researchers in the future.

Chapter 3

Noise2Void: Self-Supervised Denoising

Contents

In the previous chapter I used supervised CARE and trained deep neural networks to denoise SEM and cryo TEM data. In this chapter we will take a step back from specific data modalities and focus on the training task itself. So far neural network training for image denoising requires pairs of training data. Usually we expect these training pairs to consist of low- and high-quality images, but as Lehtinen *et al.* have shown with Noise2Noise this requirement can be relaxed and the high-quality image can be replaced by a second low-quality image as long as the noise contributions are independent. With cryoCARE we have discussed an application of Noise2Noise to cryo TEM. However, often times access to a ground truth image or even a second noisy observation is not possible. For example during live imaging the sample will move between two observations,

or if we only have access to an electron microscope with an old detector that does not support dose-fractionated movie acquisitions, or sometimes we just don't have access to the used optical setup anymore.

In such cases we have to fall back to unsupervised image denoising approaches like classical naive filtering or more advanced internal statistics methods like non-local means (Buades et al. 2005) or BM3D (Dabov et al. 2007). Classical filtering approaches are fast, but as we have seen in Chapter 2 Section 2.1 they are limited in restoration quality leading to blurred results. On the other hand we have methods like BM3D which produce good denoising results but are slow *e.g.* running BM3D on a 992×832 image takes about 4.6 seconds, the same image takes <1 second with a trained CARE network. All these methods exploit the core assumption that the signal $\boldsymbol{s}$ in a given image is not statistically independent. In other words, by observing just the neighborhood of an occluded pixel, we can make a sensible (above chance) prediction of the hidden pixel intensity. A large body of work, *e.g.* (Roth and Black 2005; Tappen et al. 2007), explicitly modeled these interdependencies via Markov Random Fields (MRFs). However, convolutional neural networks (CNNs) such as the ones used in the previous chapter, produce much better results and provide faster inference times.

In this chapter I will introduce Noise2Void: the first self-supervised denoising approach for CNNs. Noise2Void allows us to train deep neural networks if we only have access to single noisy observations, addressing the shortcomings of supervised training approaches like CARE and Noise2Noise. In Section 3.1 we will look at the image formation process from a probabilistic perspective. Next, we will take another look at fully convolutional neural networks and their receptive fields in Section 3.2. Then in Section 3.3 a detailed detailed description of Noise2Void and its efficient implementation is provided. I evaluate Noise2Void in Section 3.4, in particular, I evaluate the performance of Noise2Void on the BSD68 dataset (Roth and Black 2009) and simulated microscopy data[1]. Then I compare the results to the ones obtained by a traditionally trained network (Weigert et al. 2018), a Noise2Noise trained network, and BM3D (Dabov et al. 2007), a powerful but training-free baseline. Additionally, Noise2Void training and prediction is applied to four biomedical datasets: cryo-TEM images from (Buchholz, Jordan, et al. 2019), SEM images from (Buchholz, Krull, et al. 2019) and two datasets from the Cell Tracking Challenge[2] (Ulman et al. 2017). For these examples, I have only access to ground truth data for the SEM experiments (see previous Chapter). For the cryo TEM data I have access to a second noisy observation, hence I can train a Noise2Noise network. However, Noise2Void can be applied to all four

[1] For simulated microscopy data we know the perfect ground truth.
[2] http://celltrackingchallenge.net/

datasets showcasing the tremendous practical utility of Noise2Void. Finally we discuss the findings in Section 3.5 and summarize some of the followup works to Noise2Void.

Contributions:

- Introduction of Noise2Void, a novel approach for training denoising CNNs that requires only a body of single, noisy images.
- Comparison of Noise2Void trained denoising results to results obtained with existing CNN training schemes (Lehtinen et al. 2018; Weigert et al. 2018) and a non-trained method (Dabov et al. 2007).
- A sound theoretical motivation for the Noise2Void approach as well as a detailed description of an efficient implementation.
- Publicly available implementation of Noise2Void[3].

Parts of this chapter are published in (Krull, Buchholz, et al. 2019) and I would like to thank Alexander Krull in particular for the great collaboration on this paper. I have learned a lot from him.

3.1 Probabilistic Image Formation

So far we have looked at noisy images $\boldsymbol{x}$ as a combination of ground truth signal $\boldsymbol{s}$ and some random noise contribution $\boldsymbol{n}$

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}. \tag{3.1}$$

This interpretation is convenient and allows us to think of denoising as a simple subtraction process. However the image formation process is better described as a draw from the joint distribution

$$p(\boldsymbol{s}, \boldsymbol{n}) = p(\boldsymbol{s})p(\boldsymbol{n}|\boldsymbol{s}). \tag{3.2}$$

Let us assume $p(\boldsymbol{s})$ to be an arbitrary distribution satisfying

$$p(\boldsymbol{s}_i|\boldsymbol{s}_j) \neq p(\boldsymbol{s}_i), \tag{3.3}$$

for two pixels i and j with a certain distance of each other. That is, the pixels $\boldsymbol{s}_i$ of the signal are not statistically independent. With respect to the noise $\boldsymbol{n}$, let us assume a conditional distribution of the form

$$p(\boldsymbol{n}|\boldsymbol{s}) = \prod_i p(\boldsymbol{n}_i|\boldsymbol{s}_i). \tag{3.4}$$

[3] https://github.com/juglab/n2v

That is, pixels values $\boldsymbol{n}_i$ of the noise are conditionally independent given the signal. We furthermore assume the noise to be zero-mean

$$\mathbb{E}\left[\boldsymbol{n}_i\right] = 0, \tag{3.5}$$

which leads to

$$\mathbb{E}\left[\boldsymbol{x}_i\right] = \boldsymbol{s}_i. \tag{3.6}$$

In other words, if we were to acquire multiple images with the same signal, but different realizations of noise and average them, the result would approach the true signal. An example of this would be recording multiple photographs of a static scene using a fixed tripod-mounted camera.

3.2 Receptive Field

Until now we have treated CNNs as functions which map low-quality images to ground truth images. Here I want to introduce a slightly different but equivalent view on such networks. Lets just consider a single predicted pixel $\hat{\boldsymbol{y}}_i$ in the output of the CNN and reason about the information aggregated in it. From earlier we know that each convolutional filter has a receptive field and that the computed output is the weighted sum over all pixels within the receptive field. A CNN is essentially just a stack of multiple convolution filters with non-linearities in between. The receptive field of a stack of convolutions can recursively be computed by

$$r_{out}(r_{in}, k, j) = r_{in} + (k-1) \cdot j, \tag{3.7}$$

with r_{in} being the size of the receptive field of the previous convolution or 1 if it is the first convolution, k representing the kernel size and j being the stride of the convolution (Dumoulin and Visin 2016). For pooling operation the receptive field grows by the pooling factor *i.e.* each 2×2 pooling operation doubles the size of the receptive field. Hence, every predicted output pixel of a CNN has a certain *receptive field* $\boldsymbol{x}_{\mathrm{RF}(i)}$ of input pixels, usually a square patch around that pixel. With this knowledge we can now consider a CNN as a function that takes a patch $\boldsymbol{x}_{\mathrm{RF}(i)}$ as input an outputs a prediction $\hat{\boldsymbol{y}}_i$ for the single pixel i located at the patch center. Following this view, the denoising of an entire image can be achieved by extracting overlapping patches and feeding them to the network one by one. Consequently, a CNN can be defined as the function

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{RF}(i)}) = \hat{\boldsymbol{y}}_i, \tag{3.8}$$

where $\boldsymbol{\theta}$ denotes the trainable CNN parameters.

In supervised training we are presented with a set of training pairs $(\boldsymbol{x}^j, \boldsymbol{s}^j)$,

each consisting of a noisy input image $\boldsymbol{x}^j$ and a clean ground truth target $\boldsymbol{s}^j$. By again applying the patch-based view of the CNN, we can see the training data as pairs $(\boldsymbol{x}^j_{\mathrm{RF}(i)}, \boldsymbol{s}^j_i)$. Where $\boldsymbol{x}^j_{\mathrm{RF}(i)}$ is a patch around pixel i, extracted from training input image $\boldsymbol{x}^j$, and $\boldsymbol{s}^j_i$ is the corresponding target pixel value, extracted from the ground truth image $\boldsymbol{s}^j$ at same position. These pairs can be used to tune the parameters $\boldsymbol{\theta}$ to minimize pixel-wise loss

$$\arg\min_{\boldsymbol{\theta}} \sum_j \sum_i L\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}^j_{\mathrm{RF}(i)}) = \hat{\boldsymbol{y}}^j_i, \boldsymbol{s}^j_i\right). \tag{3.9}$$

Here the standard MSE loss

$$L\left(\hat{\boldsymbol{s}}^j_i, \boldsymbol{s}^j_i\right) = (\hat{\boldsymbol{s}}^j_i - \boldsymbol{s}^j_i)^2, \tag{3.10}$$

is considered.

Now let us consider the training procedure according to (Lehtinen et al. 2018). Noise2Noise allows us to cope without clean ground truth training data. Instead we start out with noisy image pairs $(\boldsymbol{x}^j, \boldsymbol{x}'^j)$, where

$$\boldsymbol{x}^j = \boldsymbol{s}^j + \boldsymbol{n}^j \text{ and } \boldsymbol{x}'^j = \boldsymbol{s}^j + \boldsymbol{n}'^j, \tag{3.11}$$

that is the two training images are identical up to their noise components $\boldsymbol{n}^j$ and $\boldsymbol{n}'^j$, which are, in the probabilistic image generation model, just two independent samples from the same distribution (see Eq. 3.4).

We can now again apply the patch-based perspective and view the training data as pairs $(\boldsymbol{x}^j_{\mathrm{RF}(i)}, \boldsymbol{x}'^j_i)$ consisting of a noisy input patch $\boldsymbol{x}^j_{\mathrm{RF}(i)}$, extracted from $\boldsymbol{x}^j$, and a noisy target $\boldsymbol{x}'^j_i$, taken from $\boldsymbol{x}'^j$ at the position i. As in supervised training, the parameters are tuned to minimize a loss, similar to Eq. 3.9, this time however using the noisy target $\boldsymbol{x}'^j_i$ instead of the ground truth signal $\boldsymbol{s}^j_i$. Even though we are attempting to learn a mapping from a noisy input to a noisy target, the training will still converge to the correct solution. The key to this phenomenon lies in the fact that the expected value of the noisy input is equal to the clean signal (Lehtinen et al. 2018) (see Eq. 3.6).

3.3 Noise2Void Training

Now let us go a step further. Nothing prohibits us to derive both parts of the training data, the input and the target, from a single noisy training image $\boldsymbol{x}^j$. However, if we were to simply extract a patch as input and use its center pixel as target, the network would just learn the identity, by directly mapping the value at the center of the input patch to the output (see Figure 3.1 (a)).

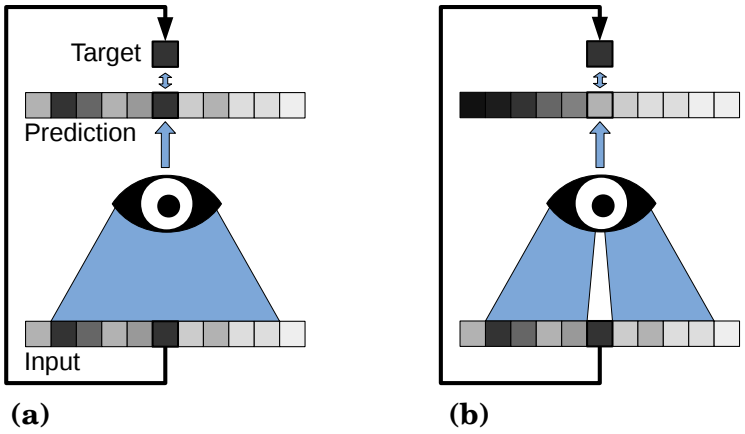

Figure 3.1: A conventional network versus the proposed blind-spot network. **(a)** In the conventional network the prediction for an individual pixel depends an a square patch of input pixels, known as a pixel's *receptive field* (pixels under blue cone). If such a network is trained using the same noisy image as input and as target, the network will degenerate and simply learn the identity. **(b)** In a *blind-spot network*, as proposed, the receptive field of each pixel excludes the pixel itself, preventing it from learning the identity. I show that blind-spot networks can learn to remove pixel wise independent noise when they are trained on the same noisy images as input and target.

To understand how training from single noisy images is possible nonetheless, let us assume that we use a network architecture with a special receptive field. We assume the receptive field $\tilde{\boldsymbol{x}}_{\mathrm{RF}(i)}$ of this network to have a blind-spot in its center. The CNN prediction $\hat{\boldsymbol{y}}_i$ for a pixel is affected by all input pixels in a square neighborhood except for the input pixel $\boldsymbol{x}_i$ at its very location. Let us call this type of network a *blind-spot network* (see Figure 3.1 b).

A blind-spot network can be trained using any of the training schemes described above. Like with a normal network, supervised training or Noise2Noise, using a clean target, or a noisy target respectively can be applied. The blind-spot network has a little bit less information available for its predictions, and a slight drop in accuracy is expected compared to a normal network. Considering however that only one pixel out of the entire receptive field is removed, we can assume it to still perform reasonably well.

The essential advantage of the blind-spot architecture is its inability to learn the identity. Let us consider why this is the case. Since we assume the noise to be pixel-wise independent given the signal (see Eq. 3.4), the neighboring pixels carry no information about the value of $\boldsymbol{n}_i$. It is thus impossible for the network to produce an estimate that is better than its *a priori* expected value (see Eq. 3.5).

The signal however is assumed to contain statistical dependencies (see Eq. 3.3). As a result, the network can still estimate the signal $\boldsymbol{s}_i$ of a pixel by looking at its surroundings.

Consequently, a blind-spot network can be trained with input patch and target value being extracted from the same noisy training image. During training the empirical risk

$$\arg\min_{\boldsymbol{\theta}} \sum_j \sum_i L\left(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}^j_{\mathrm{RF}(i)}), \boldsymbol{x}^j_i\right), \tag{3.12}$$

is minimized. Note that the target $\boldsymbol{x}^j_i$, is just as good as the N2N target $\boldsymbol{x'}^j_i$, which has to be extracted from a second noisy image. This becomes clear when we consider Eqs. 3.11 and 3.4: The two target values $\boldsymbol{x}^j_i$ and $\boldsymbol{x'}^j_i$ have an equal signal $\boldsymbol{s}^j_i$ and their noise components are two independent samples from the same distribution $p(\boldsymbol{n}_i|\boldsymbol{s}^j_i)$.

We have seen that a blind-spot network can in principle be trained using only individual noisy training images. However, implementing such a network that can still operate efficiently is not trivial. As workaround special masking schemes are used, which replace the pixel value in the center of each input patch with a randomly selected value. This effectively erases the pixel's information and prevents the network from learning the identity. The following replacment strategies are part of NOISE2VOID:

- **Uniform Pixel Selection (UPS)** replaces the value of the selected pixel i with a randomly selected pixel value from a square window around i. This includes the pixel itself.
- **Gaussian (G)** changes the value of the selected pixel i by adding random Gaussian noise.
- **Gaussian Fitting (GF)** fits a 1D Gaussian distribution to the pixel values within a small square neighborhood around the pixel i with the center pixel included and draws a sample from this distribution as replacement value.
- **Gaussian Pixel Selection (GPS)**

These masking schemes are necessary to use out of the box U-Nets for NOISE2VOID training.

3.3.1 Implementation Details

A naive implementation of the above training scheme is unfortunately still not very efficient. An entire patch has to be processed to calculate the gradients for a single

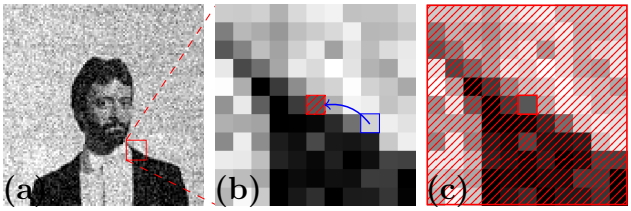

Figure 3.2: Blind-spot masking scheme used during NOISE2VOID training. **(a)** A noisy training image. **(b)** A magnified image patch from (a). During NOISE2VOID training, a randomly selected pixel is chosen (blue rectangle) and its intensity copied over to create a blind-spot (red and striped square). This modified image is then used as input image during training. **(c)** The target patch corresponding to (b). I use the original input with unmodified values also as target. The loss is only calculated for the blind-spot pixels which is masked in (b).

output pixel. To mitigate this issue, the following approximation technique is used. Given a noisy training image $\boldsymbol{x}_i$, patches which are bigger than the receptive field of the network are randomly extracted.. For example a U-Net (Ronneberger et al. 2015) from the CSBDeep (Weigert et al. 2018) framework with `n_depth` $= 2$ and kernel size 3×3 has a receptive field size corresponding to a patch of 22×22 pixels. If the kernel size is increased to 5×5 the receptive field grows to 40×40 pixels. Hence, the default size of randomly extracted patches is set to 64×64 pixels. Within each patch randomly N pixels are selected, using stratified sampling to avoid clustering. Then these pixels are masked and the original noisy input values are used as targets at their position (see Figure 3.2). For the masked pixels the loss is then calculated simultaneously and backpropagated, while ignoring the rest of the predicted image. This is achieved by setting the loss for non-masked pixels to zero in a customized loss function.

The publicly available implementation of NOISE2VOID[4] is based on the CSBDeep (Weigert et al. 2018) framework. Furthermore a Fiji (Schindelin et al. 2012) implementation of NOISE2VOID[5] is providing a one-click solution to deep learning based image denoising.

3.4 Experiments

Now I will evaluate NOISE2VOID on natural images, simulated biological image data, and acquired microscopy images. The NOISE2VOID results are compared to results of supervised CARE and NOISE2NOISE trained CARE, as well as results of BM3D (Dabov et al. 2007), a non deep learning approach. From a methodological perspective the deep learning based approaches can be divided in

[4] https://github.com/juglab/n2v
[5] https://imagej.net/N2V

two steps: (i) the training phase, during which a the neural networks are optimized w.r.t. a suitable loss function and (ii) the inference or test phase, during which the neural network weights are fixed and only applied to novel unseen data. The BM3D approach on the other hand has no training phase and is directly applied to the test images. However, BM3D will build up an internal denoising state from scratch for each test image, which results in increased run times compared to deep learning inference.

All neural networks in the following experiments were trained for 200 epochs with 400 steps per epoch and the data is normalized to 0-mean and 1-standard deviation. Various Noise2Void training and prediction notebooks are publicly available[6].

3.4.1 Natural Images

For the evaluation on natural image data I followed the example of (Zhang et al. 2017) and took 400 gray scale images with 180×180 pixels of which randomly 1% were chosen as validation data and the rest served as training data. For testing the gray scale version of the BSD68 dataset was used. Noisy versions of all images are generated by adding zero mean Gaussian noise with standard deviation $\sigma = 25$. Furthermore, data augmentation is used on the training dataset. More precisely, each image was rotated three times by 90° and also all mirrored versions were added. During training random 64×64 pixel patches from this augmented training dataset were drawn.

The network architecture used for all BSD68 experiments is a U-Net (Ronneberger et al. 2015) with `n_depth` = 2, kernel size 3, batch normalization, and a linear activation function in the last layer. The network has `n_first` = 96 feature maps on the initial level, which get doubled while the network gets deeper. A learning rate of 0.0004 and the default CARE learning rate schedule, halving the learning rate when a plateau on the validation loss is detected are used. The validation loss is computed on a fixed set of randomly chosen pixels in the validation dataset.

I used batch size 128 for traditional training and batch size 16 for Noise2Noise, where I found that a larger batch leads to slightly diminished results. For Noise2Void training I used a batch size of 128 and simultaneously manipulated $N = 64$ pixels per input patch (see Section 3.3.1), as before with an initial learning rate of 0.0004.

In the first row of Figure 3.3, I compare Noise2Void results (with *Uniform*

[6] https://github.com/juglab/n2v/tree/master/examples

Masking Types

Masking	Kernel	Loss	Features	PSNR
UPS (3×3)	3×3	MSE	96	26.98
UPS (5×5)	3×3	MSE	96	27.71
UPS (7×7)	3×3	MSE	96	27.26
UPS (50×50)	3×3	MSE	96	27.42
GF (3×3)	3×3	MSE	96	27.51
GF (5×5)	3×3	MSE	96	27.31
GF (7×7)	3×3	MSE	96	27.47
GF (50×50)	3×3	MSE	96	27.35
G (5)	3×3	MSE	96	27.24
G (10)	3×3	MSE	96	26.52
GPS	3×3	MSE	96	27.31

Other Parameters

Masking	Kernel	Loss	Features	PSNR
UPS (5×5)	5×5	MSE	96	27.60
UPS (5×5)	3×3	MAE	96	27.58
UPS (5×5)	5×5	MAE	96	26.99
UPS (5×5)	3×3	MAE	32	27.33
UPS (5×5)	5×5	MAE	32	27.36

Table 3.1: Results achieved with various masking methods and different parameter settings on the BSD68 dataset.

Pixel Selection UPS with a 5×5 pixel window) to the ones obtained by BM3D, traditional CARE training, and Noise2Noise training. While on visual inspection, all results look similarly good, the PSNR value of the classical training method is clearly best. As mentioned earlier, Noise2Void is not expected to outperform other training methods, as it can utilize less information for its prediction. Still, here we observe that the denoising performance of N2V even drops moderately below the performance of BM3D (which is not the case for other data).

In Table 3.1 performance of all proposed masking schemes on the BSD68 dataset are compared. Furthermore results obtained with different kernel sizes and with the mean absolute error (MAE) loss are presented.

3.4.2 Light Microscopy Data

Simulated Data

The acquisition of close to ground truth quality microscopy data is either impossible or at the very least, difficult and expensive. Since ground truth data is required to compute desired PSNR values, I decided to use a simulated dataset for our second set of experiments. To this end, I used simulated membrane labeled

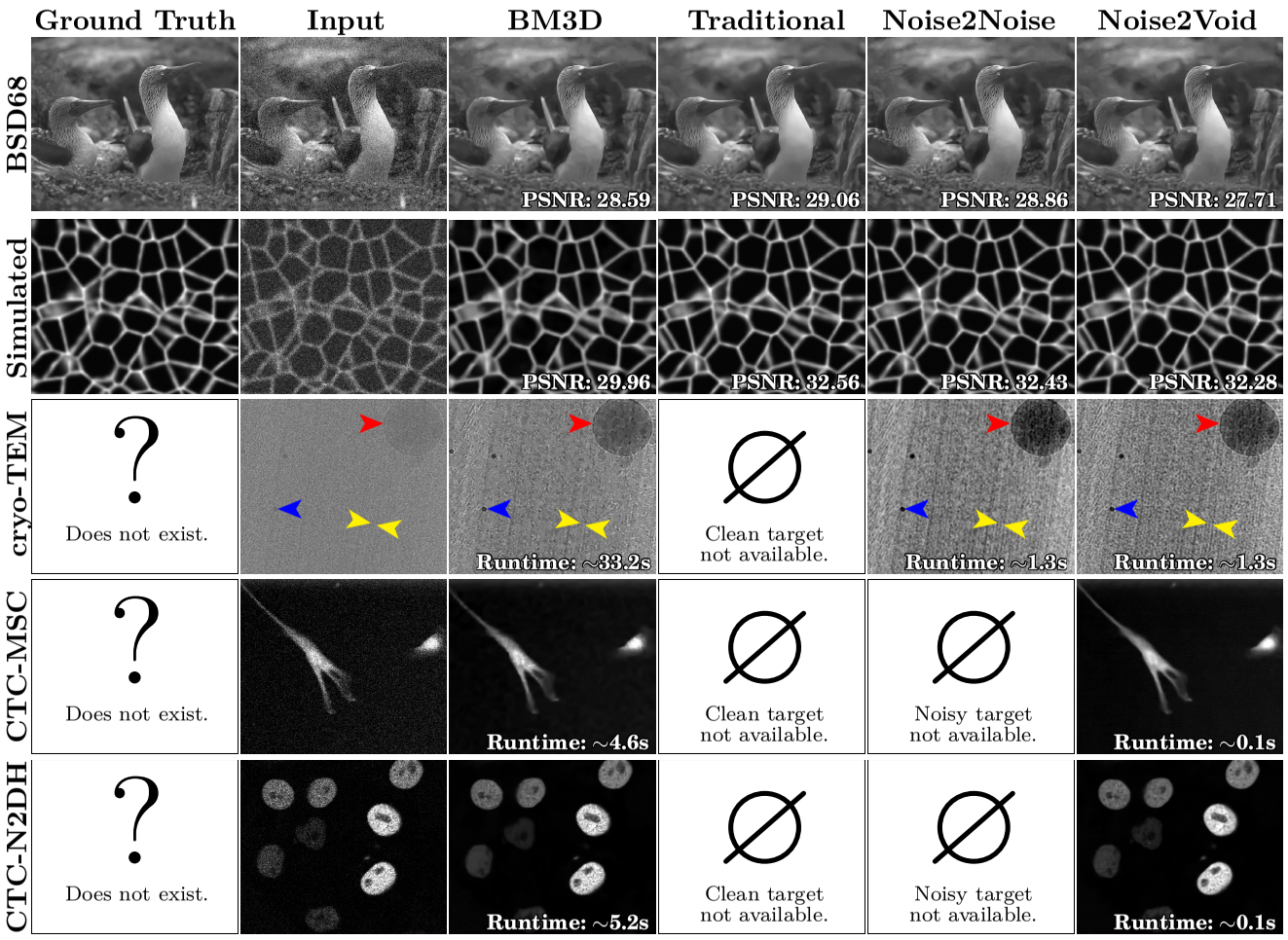

Figure 3.3: Results obtained by traditionally trained and Noise2Noise trained compared to Noise2Void trained denoising networks. For BSD68 data (row one) and simulated data (row two) all three training schemes are applicable. For cryo TEM data ground truth images are unobtainable. Since pairs of noisy images are available, I can still perform Noise2Noise training. Red, yellow, and blue arrowheads indicate an ice artifact, two tubulin protofilaments that are known to be $4nm$ apart, and a $10nm$ gold bead, respectively. For the CTC-MSC and CTC-N2DH data (rows 4 and 5, respectively) only single noisy images exist. Hence, neither traditional nor Noise2Noise training is applicable, while the proposed Noise2Void training scheme can still be applied.

cells *epithelia* provided by Alexander Dibrov and mimicked the typical image degradation of fluorescence microscopy by first applying Poisson noise and then adding zero-mean Gaussian noise. More specifically I normalized the simulated membrane images to the range $[0, 1]$. Then I added a constant value of 0.2 to simulate background illumination, followed by multiplication with a factor λ to account for the exposure. Then shot noise is simulated by drawing for each pixel from a Poisson distribution conditioned on the shifted and scaled membrane image. Finally readout sensor noise is simulated by adding zero-mean Gaussian noise with a standard deviation of $\sigma = 1$. This scheme allows simulation of low- and high-exposure images, using $\lambda = 20$ and $\lambda = 10'0000$ respectively. For the NOISE2NOISE experiments a second noisy observation was generated with $\lambda = 20$. The high-exposure images were used as ground truth images for the CARE training and to calculate the PSNR during testing. Due to the different scaling factors λ for low- and high-exposure images, I had to rescale the outputs of NOISE2NOISE and NOISE2VOID prior to the PSNR computation. I used the same data augmentation scheme as described in Section 3.4.1.

The network architecture used for all experiments on simulated data is a U-Net (Ronneberger et al. 2015) of with `n_depth` = 2, kernel size 5, batch norm, `n_first` = 32, and a linear activation function in the last layer. Traditional and NOISE2NOISE training was performed with batch size 16 and an initial learning rate of 0.0004. The NOISE2VOID training was performed with a batch size of 128. I chose to simultaneously manipulate $N = 64$ pixels per input patch (see Section 3.3.1) and use the *Gaussian Pixel Selection* (GPS) masking method. Again the standard CARE learning rate schedule was used for all three training methods.

In the second row of Figure 3.3 one can appreciate the denoising quality of NOISE2VOID training, which reaches virtually the same quality as traditional and NOISE2NOISE training. All trained networks clearly outperform the results obtained by BM3D.

Real Light Microscopy Data

I tested NOISE2VOID on two fluorescence microscopy datasets from the Cell Tracking Challenge (Ulman et al. 2017). The first dataset, Fluo-C2DL-MSC (CTC-MSC) consists of two movies and I only used the provided image data without any additional ground truth annotations for segmentation. I extracted 256 randomly selected patches of size 80×80 pixels from each movie frame. To ensure that each patch contains some foreground signal I computed the standard deviation over the pixel intensities of each patch and rejected patches with standard deviation below 1250. From each frame a single patch is used as validation data only. On

the training data I employed the same augmentation strategy as before.

The second dataset, Fluo-N2DH-GOWT1 (CTC-N2DH) also consists of two movies. Like before, only the image data without any ground truth annotations for segmentation are used. Like for the CTC-MSC dataset I extracted 256 randomly selected patches of 80×80 pixels, but this time patches with a standard deviation below 5 were rejected.

Since no ground truth images or second noisy observations are available, only self-supervised image denoising methods like Noise2Void and BM3D can be used to denoise these data. The last two rows of Figure 3.3 show the results of BM3D and Noise2Void. We can see that the Noise2Void trained network gives subjectively smooth and appealing results, while requiring only a fraction of the BM3D runtime.

The network architecture used for all experiments on real microscopy (light and electron) data is a U-Net (Ronneberger et al. 2015) of `n_depth` = 2, kernel size 3, batch norm, `n_first` = 32, and a linear activation function in the last layer. For an efficient training of Noise2Void $N = 64$ pixels per input patch (see Section 3.3.1) were simultaneously manipulate and with the *Uniform Pixel Selection* (UPS) masking method with a 5×5 window. I used a batch size of 128 and a initial learning rate of 0.0004.

3.4.3 Electron Microscopy Data

The same network setup as above was used for the EM experiments.

SEM Data

I also applied Noise2Void to SEM data acquired by our colaborators from the Jékely lab (see previous chapter for more details). Since, I have access to slow- and fast-scanned SEM images I can also train a fully supervised denoising network and compute PSNR numbers for the noisy input, the supervised prediction and the Noise2Void prediction see Figure 3.4. Both neural networks were trained on 442 training and 68 validation patches of 96×96 pixels and I used data augmentation *i.e.* all four 90° rotations and mirroring.

Cryo-TEM Data

In cryo-TEM, the acquisition of high-SNR images is not possible due to beam induced damage (Knapek and Dubochet 1980) as discussed in the previous chapter. We have already seen that cryoCARE can be used to denoise cryo TEM data by employing the Noise2Noise training paradigm. Now we will look at a

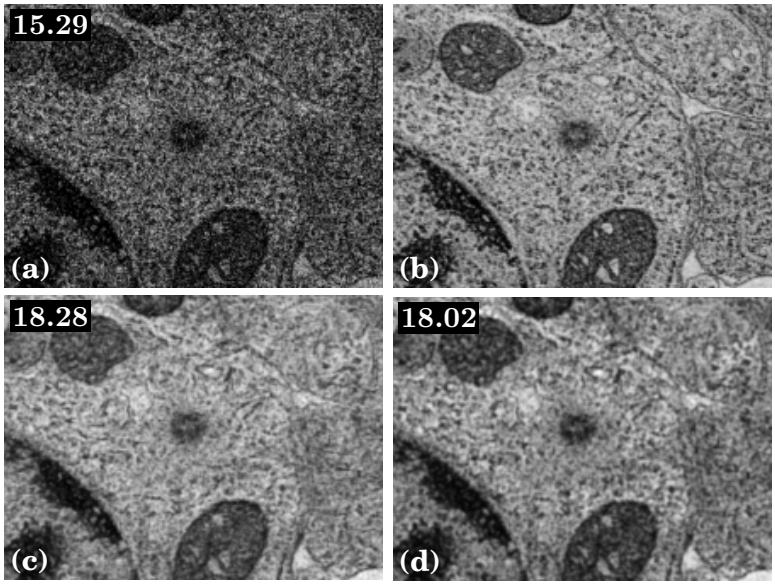

Figure 3.4: Subfigure **(a)** shows a low-SNR SEM image from the Jékely lab (see previous chapter). Subfigure **(b)** shows the corresponding high-SNR ground truth image. Subfigure **(c)** shows the denoising result of a supervised trained CARE network using pairs of low- and high-quality images. Subfigure **(d)** shows the denoising result achieved with Noise2Void using only sinlge noisy observations for training. The numbers in the top left corners of (a), (c) and (d) are the PSNR values with respect to the ground truth image (b).

Noise2Void trained network on the same data. The network was trained on a single cryo TEM projection of 7676×7420 pixels and of which 435 overlapping patches of 512×512 pixels were extracted. Of these patches 10% are used as validation data and no data augmentation is used for training.

In terms of inference runtime Noise2Noise and Noise2Void are equal, which is expected, and about 25 times faster than BM3D. Furthermore, BM3D seems to struggle more with fine details (indicated in Figure 3.3) compared to Noise2Void.

As mentioned in the previous section, ground truth quality microscopy data is typically not available. Hence, I can no longer compute PSNR values.

3.4.4 Errors and Limitations

In this section we will look at extreme error cases of Noise2Void predictions and discuss the limitation of Noise2Void to pixel-wise independent noises. We will start by looking at the denoising results of real world images *i.e.* the BSD68 data, for which Noise2Void performed least convincing. Figure 3.5 shows the ground truth image, and prediction results of a fully supervised CARE trained and a Noise2Void trained network. The upper row contains the image with the

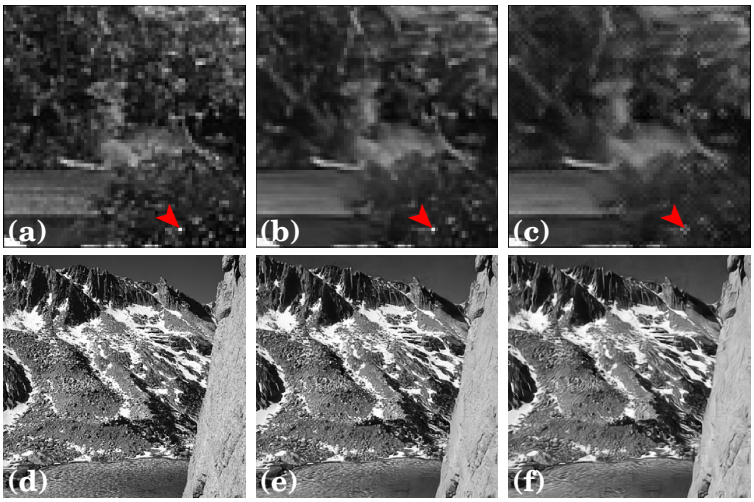

Figure 3.5: Failure cases of NOISE2VOID trained networks. **(a)** A crop from the ground truth test image with the largest individual pixel error (indicated by red arrow). **(b)** Result of a network trained with available ground truth. **(c)** Result of a NOISE2VOID trained network. The network fails to predict this bright and isolated pixel. **(d)** A crop from the ground truth test image with the largest total error. **(e)** Result of a network trained with ground truth targets. **(f)** Result of a NOISE2VOID trained network. Both networks are not able to preserve the grainy structure of the image, but the NOISE2VOID trained network loses more high-frequency detail.

largest squared single pixel error and the lower row shows the image with the largest sum of squared pixel errors.

NOISE2VOID works on the assumptions that the signal is pixel-wise dependent on each other (see Equation 3.3), while the noise is pixel-wise independent of each other (see Equation 3.4). The images in Figure 3.5 show either a single bright pixel surrounded by dark pixels (top row) or highly irregular patterns of rubble in the mountains and a breeze on a lake (lower row). Since these signal constellations are rare in the training data it is more difficult for the neural network to pick these patterns up and restore them. This is also true for supervised approaches, however these have additionally access to all input pixels, while in NOISE2VOID the center pixel is removed from the input. In cases where the center pixel coincides with a single bright pixel NOISE2VOID has no access to this information at all. Therefore, the loss of reconstruction quality is expected.

In Figure 3.6 another, in some sense opposite, limitation of NOISE2VOID can be observed. It is impossible for NOISE2VOID to distinguish between ground truth signal and structured noise. As soon as the noise, *e.g.* a checkerboard pattern or lines, span over multiple pixels the pixel-wise independence assumption for noise is violated (see Equation 3.4). This phenomenon was first observed when NOISE2VOID was applied to real low exposure microscope images, where the de-

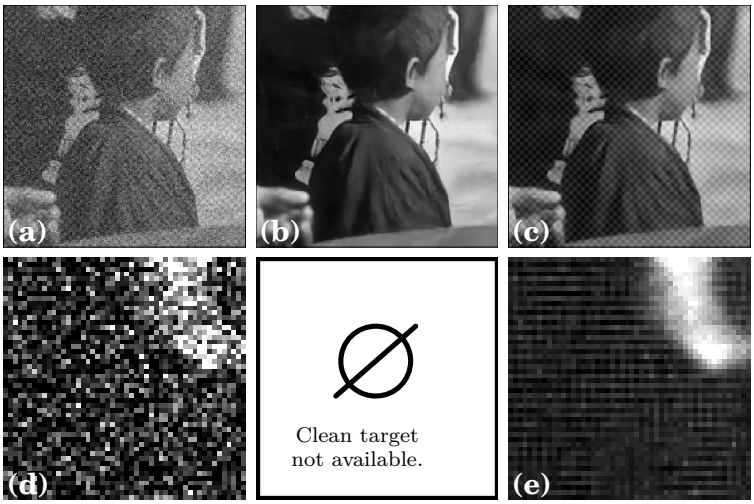

Figure 3.6: Effect of structured noise on NOISE2VOID trained network predictions. Structured noise violates the assumption that noise is pixel-independent (see also Eq. 3.4). **(a)** A photograph corrupted by structured noise. The hidden checkerboard pattern is barely visible. **(b)** The denoised result of a traditionally trained CNN. **(c)** The denoised result of an NOISE2VOID trained CNN. The independent components of the noise are removed, but the structured components remain. **(d)** Structured noise in real microscopy data. **(e)** The denoised result of an NOISE2VOID trained CNN. A hidden pattern in the noise is revealed. Note that due to the lacking training data, it is not possible to use NOISE2NOISE or the traditional training scheme in this case.

noised images contained lines (lower row in Figure 3.6). This limitation can be replicated by superimposing a faint checkerboard pattern before artificially applying noise to a ground truth image (see top row in Figure 3.6). The checkerboard pattern is well hidden in the noisy observation and if a supervised CARE model is trained with clean ground truth targets, *i.e.* ground truth without the checkerboard, the model is able to remove the noise as well as the checkerboard pattern. However, NOISE2VOID will only remove the pixel-wise independent noise contributions and treat the checkerboard pattern as part of the ground truth signal. The striped pattern in the microscope images can be traced back to a systematic error of the imaging system.

3.5 Conclusion and Followup Work

I have introduced NOISE2VOID, a novel training scheme that only requires single noisy acquisitions to train denoising CNNs. We have seen its application to a variety of image modalities *i.e.* photography, fluorescence microscopy, cryo TEM and SEM. And as long as both initial assumptions of predictable signal and pixel-wise independent noise are met, NOISE2VOID trained networks can compete with supervised trained networks. We have also looked at examples where these

assumptions are violated, showing us limitations of Noise2Void.

Concurrent with the initial presentation of Noise2Void at CVPR in 2019, Batson *et al.* presented a similar method for self-supervised training of neural networks and other systems, also based on the idea of removing parts from the input (Batson and Royer 2019). Since then, multiple incremental papers by different groups have been published, which deal with different shortcomings of Noise2Void and we will go over some of them in the following paragraphs.

Probabilistic Noise2Void by Krull *et al.* predicts an intensity distribution for each output pixel instead of a single value. This distribution is characterized by 800 simultaneously predicted values by the CNN. Then the predicted distribution is combined with a suitable noise model, which results in a complete probabilistic description of the noisy observation and the ground truth signal. Once the network is trained, they use minimum mean squared error inference to predict the clean ground truth signal. The required noise model is crucial to this method, however it is only dependent on the optical system and not on the image data (Krull, Vičar, et al. 2020). While probabilistic Noise2Void uses a histogram based noise model, Prakash *et al.* go a step further in (Prakash, Lalit, et al. 2020) and replace the noise model with a Gaussian mixture model, which leads to more robust results. Additionally, they explore bootstrapping the noise model via Noise2Void directly from the image data to denoise.

In Noise2Void the pixel is masked by replacing it with random values, conditioning the model to ignore the center pixel. Laine *et al.* propose a different network architecture which uses shifted convolutions and receptive fields which only grow in a single direction. They create a true blind spot by feeding the same image four times each with one of the 90° rotations (S. Laine et al. 2019). Honzátko *et al.* propose a dilated convolution with a blind-spot in its kernel, alleviating the rotation requirement from before and leading to a smoother coverage of the receptive field (Honzátko et al. 2020).

Broaddus *et al.* tackle the problem of structured noise removal with self-supervised training by increasing the size of the blind-spot (Broaddus et al. 2020). This extension is especially useful to handle striping noise artefacts produced by some microscope cameras. StructN2V is also part of the publicly available Noise2Void implementation.

Noise2Same by Xie *et al.* proposes to apply the model once to the un-masked and once to the masked input image. Then the reconstruction loss is computed between the un-masked output and the noisy image. This alone would lead to

a network which learns the identity. To avoid this, they introduce a second loss which is computed between the masked pixels of the un-masked and masked output and enforces them to be equal. These two losses are combined, resulting in significantly better image denoising results compared to previous self-supervised image denoising methods (Xie et al. 2020).

DivNoising by Prakash *et al.* takes a different route compared to all other approaches mentioned above and builds on the probabilistic noise model formulation in Probabilistic Noise2Void by Krull *et al.* However, the U-Net architecture is replaced by a variational auto-encoder (VAE) (Kingma and Welling 2014), which allows generating multiple diverse denoising results, which then can be aggregated into a single denoised solution (Prakash, Krull, et al. 2021).

These are some of the follow-up works to Noise2Void and it is exciting to see these works develop and push self-supervised image denoising performance closer to fully supervised image denoising approaches where ground truth data is available. I believe that there is still a lot of interesting research to be done in the field of self-supervised image denoising. I expect novel neural network architectures and loss function to emerge in the future. Already today, Prakash *et al.* have replaced the standard U-Net with VAE and it will be interesting to see which architectures will be used in the future. I expect that novel ideas will generate networks which are able to distinguish between wanted and unwanted signal. One day we might be able to separate a given image into noise, structured noise and ground truth signal.

Chapter 4

Fourier Image Transformer

Contents

So far we have discussed content-aware image restoration, in particular image denoising, with supervised training using pairs of low- and high-quality images, with Noise2Noise training using pairs of low-quality images and self-supervised training with Noise2Void, which only requires single noisy observations. The backbone of all methods discussed until now was the U-Net (Ronneberger et al. 2015), a convolutional encoder-decoder architecture with skip-connections. The U-Net encodes a corrupted image into a latent space embedding, from which the U-Net decoder reconstructs a restored image. High resolution details can be

passed through the skip-connections as well as gradients during backpropagation. However, in this last chapter we will step away from the U-Net architecture and turn towards Transformers, a very different type of neural networks originally proposed for natural language processing (NLP) and explore their application to image restoration tasks.

Transformer architectures are currently setting new standards on virtually all natural language processing (NLP) tasks (Devlin et al. 2018; Radford et al. 2018). Recently, Transformers were also successfully applied to image classification tasks (Dosovitskiy et al. 2020; Ramachandran et al. 2019) and pixel-by-pixel image generation (M. Chen et al. 2020). Hence, Transformers might be the next big step in computer vision related tasks.

The key novelty of Transformers is their self-attention mechanism (Vaswani et al. 2017), allowing them to learn and utilize long ranging dependencies in data. Recently, this mechanism is applied to longer and longer input sequences of words or other elements, *e.g.* pixels (M. Chen et al. 2020). In this chapter I will investigate if pixel sequences are the only valid image representation to train *auto-regressive* Transformer models in the spirit of (M. Chen et al. 2020). In particular, I want to use a representation where each prefix of such a descriptive sequence encodes the full image at lower resolution. Therefore, I introduce *Fourier Domain Encodings* FDEs, which do have this desired property and, as I will show, can successfully be used to train auto-regressive Fourier Image Transformer for super-resolution (FIT: SRes).

Additionally, I will investigate how an *encoder-decoder* based Fourier Image Transformer can be trained on a set of Fourier measurements and then used to query arbitrary Fourier coefficients, which I use to improve sparse-view computed tomography (CT) image restoration by filling in missing Fourier coefficients, hence removing the missing wedge reconstruction artefacts directly in Fourier space. I will call this approach Fourier Image Transformer for tomograpic reconstruction (FIT: TRec). I demonstrate this by providing a given set of projection Fourier coefficients to the encoder-decoder setup and use it to predict Fourier coefficients at arbitrary query points. This allows the prediction of a dense, grid-sampled discrete Fourier spectrum of a high quality CT reconstruction.

In Section 4.1 I discuss the relevant transformer literature. In Section 4.2 I introduce the novel Fourier Domain Encoding (FDE) and training strategies for auto-regressive and encoder-decoder transformer models. Finally in Section 3.4 super-resolution and tomographic reconstruction experiments are presented and evaluated.

Contributions:

- Fourier Domain Encoding (FDE) a novel sequential image encoding.
- Fourier Image Transformer for super-resolution (FIT: SRes) by training an *auto-regressive* transformer on the FDE.
- Novel approach to tomographic reconstruction, which aims to resolve missing wedge artefacts directly in Fourier space.
- Open-source implementation of Fourier Image Transformer in PyTorch[1].

Parts of this chapter are under review and available as arXiv pre-print[2].

4.1 Transformers

Transformer architectures are revolutionizing neural language processing (NLP), replacing recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures in virtually all NLP tasks (Devlin et al. 2018; Radford et al. 2018). The work presented in this chapter is based on the seminal paper by Vaswani *et al.*, which introduces the self-attention mechanism that is key to the success of Transformers (Vaswani et al. 2017). Since I work on images, where input sequences tend to be long, I use Fast-Transformers, an efficient approximation of softmax self-attention, as introduced by Katharopoulos *et al.* (Katharopoulos et al. 2020).

4.1.1 Attention Is All You Need

Vaswani *et al.* were the first to introduce Transformers. More specifically, they introduced an encoder-decoder structure, where the encoder maps an input encoding $\mathbf{x} \in \mathbb{R}^{N \times F}$ into a continuous latent space $\mathbf{z} \in \mathbb{R}^{N \times F}$, with N corresponding to the number of input tokens and F representing the feature dimensionality per token. This latent space embedding $\mathbf{z}$ is then given to the decoder, which generates an M long output sequence $\mathbf{y} \in \mathbb{R}^{M \times F}$ iteratively, element by element. This auto-regressive decoding scheme means that the decoder generated the i-th output token while not only observing $\mathbf{z}$, but also all $i-1$ output tokens generated previously (Vaswani et al. 2017).

More formally, a Transformer is a function $T : \mathbb{R}^{N \times F} \rightarrow \mathbb{R}^{N \times F}$, represented by L Transformer layers

$$T_l(\mathbf{x}) = f_l(A_l(\mathbf{x}) + \mathbf{x}), \tag{4.1}$$

[1] https://github.com/juglab/FourierImageTransformer
[2] https://arxiv.org/abs/2104.02555

with A_l denoting a self-attention module and f_l being a simple feed forward network.

In the self-attention module, the input $\mathbf{x}$ is mapped to queries $Q = \mathbf{x}W_Q$, keys $K = \mathbf{x}W_K$ and values $V = \mathbf{x}W_V$ by matrix multiplication with learned matrices $W_Q \in \mathbb{R}^{F\times D}$, $W_K \in \mathbb{R}^{F\times D}$ and $W_V \in \mathbb{R}^{F\times F}$. The self-attention output is then computed by

$$A_l(\mathbf{x}) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V, \tag{4.2}$$

with the softmax-function being applied per row. Intuitively, the softmax-normalized similarity between computed keys and queries is used to obtain the weighted sum over the values.

Typically, instead of a single self-attention module, multi-head attention is being used. If that is the case, a transformer layer T_l learns multiple W_Q, W_K and W_V, allowing the layer to simultaneously perform multiple attention-based computations (Vaswani et al. 2017).

Since transformers do not explicitly encode the relative position between input tokens, positional encodings are required whenever specific input topologies need to be made accessible to the Transformer. In (Vaswani et al. 2017), a useful 1D positional encoding scheme was proposed. Later, Wang *et al.* (Zelun Wang and J.-C. Liu 2020) generalized this scheme to 2D topologies. In this work, I have adopted this encoding scheme, but use it not only to encode integer pixel-grid locations, but arbitrary real coordinates.

4.1.2 Fast-Transformers

While the advantage of transcending beyond CNN's localized receptive fields by introducing global attention proves beneficial for many learning tasks, the big downside is the computational cost associated with it. Due to the required matrix multiplication QK^T, the self-attention on an input sequence of length N requires $\mathcal{O}(N^2)$ memory and time. Katharopoulos *et al.* propose to generalize the self-attention for the i-th row to

$$A_l^i(\mathbf{x}) = \frac{\sum_{j=1}^{N}\text{sim}(Q_i, K_j)V_j}{\sum_{j=1}^{N}\text{sim}(Q_i, K_j)}, \tag{4.3}$$

with sim being a non-negative similarity function, which includes all kernels $k(x, y) = \langle\Phi(x), \Phi(y)\rangle$ with $k(x, y) : \mathbb{R}^{2\times F} \to \mathbb{R}^+$ where Φ is the feature mapping of the kernel. Using the feature mapping Φ Equation 4.3 can be written as

$$A_l^i(\mathbf{x}) = \frac{\Phi(Q_i)^T \sum_{j=1}^N \Phi(K_j) V_j}{\Phi(Q_i)^T \sum_{j=1}^N \Phi(K_j)}. \tag{4.4}$$

This reduces the time and memory complexity to $\mathcal{O}(N)$, because $\sum_{j=1}^N \Phi(K_j) V_j^T$ and $\sum_{j=1}^N \Phi(K_j)$ can be computed once and reused for each subsequent query computation (Katharopoulos et al. 2020).

Since the linearization of the softmax is infeasible Fast-Transformers use

$$\Phi(\mathbf{x}) = \text{elu}(\mathbf{x} + 1) \tag{4.5}$$

as feature mapping and show that their formulation performs on par with the original softmax-attention, while significantly reducing time and memory consumption.

4.1.3 Transformers in Computer Vision

The success of Transformers in NLP has naturally raised the question if computer vision tasks might as well benefit from transformer-like attention. Recent work on image classification (Dosovitskiy et al. 2020; Ramachandran et al. 2019) and pixel-by-pixel image generation/completion (M. Chen et al. 2020; Katharopoulos et al. 2020; Parmar et al. 2018) were among the first to successfully demonstrate the applicability of Transformers in the image-domain. In (Parmar et al. 2018), for example, the first n pixels of the flattened input image are used to condition a generative transformer setup that then predicts the remaining image in an auto-regressive manner. Since image-based applications have to deal with very long input sequences, the use of efficient transformer implementations (Bello 2021; Katharopoulos et al. 2020; Kitaev et al. 2020; S. Wang et al. 2020) is essential (see discussion above).

4.2 Methods

4.2.1 Fourier Domain Encodings (FDEs)

To compute the Fourier Domain Encoding (FDE) for an image $\boldsymbol{x}$, I first take the discrete Fourier transform (DFT), $\boldsymbol{X} = \mathcal{F}(\boldsymbol{x})$, resulting in the complex valued Fourier spectrum $\boldsymbol{X}$. The DC component of $\boldsymbol{X}$ is in its center-most location. Concentric rings of Fourier coefficients around the DC component are called Fourier rings. More central Fourier rings contain lower frequencies, coefficients further away from the DC component higher ones. Since I started with a real-valued image $\boldsymbol{x}$, I know that the coefficients to the left of the DC component are redundant to the ones on the right. Hence, I drop the left half of $\boldsymbol{X}$ and call the

remaining half $\boldsymbol{X}_h$. If one masks all Fourier ring coefficients up to a radius r and then back-transforms the result we receive $\boldsymbol{x}_{\mathbf{0}-\boldsymbol{r}} = \mathcal{F}^{-1}(\boldsymbol{X} \odot \boldsymbol{M}_{\mathbf{0}-\boldsymbol{r}})$, with $\mathbf{M}_{\mathbf{0}-\mathbf{r}}$ being a circular mask containing 1 at every location from the DC component up to the Fourier coefficients at distance r and 0's beyond that. The image $\boldsymbol{x}_{\mathbf{0}-\boldsymbol{r}}$ is a lower-resolution version of the original image $\boldsymbol{x}$, its effective resolution depending on the value of r, see Figure 4.1 for an example. Hence, if I create a sequence of Fourier coefficients starting from the DC component and followed by an unrolling of (half) Fourier rings from $\boldsymbol{X}_h$, I end up with a sequential image representation

$$\mathbf{S} = \text{unroll}(\boldsymbol{X}_h) = [c_1, c_2, \ldots, c_N]^T, \tag{4.6}$$

where c_i – the words of theS input sequence – are complex Fourier coefficients. The sequence $\mathbf{S}$ has the desired prefix-properties required for the super-resolution task, which will be introduced in Section 4.3.

In order to proceed, the complex Fourier coefficients c_i are converted into normalized amplitudes

$$a_i = \frac{2(|c_i| - a_{max})}{a_{max} - a_{min}} - 1 \tag{4.7}$$

and phases

$$\phi_i = \frac{\angle(c_i)}{\pi}, \tag{4.8}$$

where a_{min} and a_{max} are minimum and maximum amplitudes computed over all training images and the function $\angle$ returns the phase of a given Fourier coefficient. Hence, the complex sequence $\mathbf{S}$ is now the described by the normalized real-valued matrix

$$\mathbf{C} = \begin{bmatrix} a_1 & \cdots & a_N \\ \phi_1 & \cdots & \phi_N \end{bmatrix}^T, \tag{4.9}$$

with $\mathbf{C} \in \mathbb{R}^{N \times 2}$.

The final goal is to transform each word (a_i, ϕ_i) into an F-dimensional vector. To this end I feed $\mathbf{C}$ through a single trainable linear layer that increases the feature dimensionality from 2 to $\frac{F}{2}$, to which a $\frac{F}{2}$-dimensional 2D positional encoding is concatenated. The 2D positional encoding is an adapted version of (Zelun Wang and J.-C. Liu 2020), which accepts arbitrary (non integer) coordinates. This allows me to encode the original polar coordinates of the Fourier coefficient in the original 2D Fourier spectrum $\boldsymbol{X}_{\boldsymbol{h}}$ in the positional encoding. The final FDE image sequence is therefore $\boldsymbol{E} \in \mathbb{R}^{N \times F}$.

Predicted output words $\mathbf{Z} = [z_1, \ldots, z_k]$, with $z_i \in \mathbb{R}^F$, are fed through two linear layers that back-transform the F-dimensional encoding of z_i into predicted am-

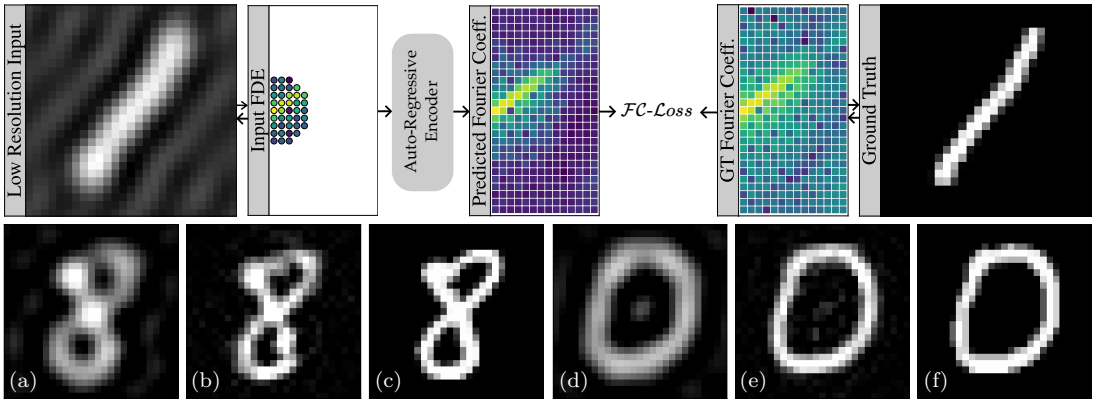

Figure 4.1: **FIT for super-resolution.** Low-resolution input images are first transformed into Fourier space and then unrolled into an FDE sequence, as described in Section 4.2.1. This FDE sequence can now be fed to a FIT, that, conditioned on this input, extends the FDE sequence to represent a higher resolution image. This setup is trained using an $\mathcal{FC}$-$\mathcal{L}oss$ that enforces consistency between predicted and ground truth Fourier coefficients. During inference, the FIT is conditioned on the first 39 entries of the FDE, corresponding to **(a, d)** 3× Fourier binned input images. Panels **(b, e)** show the inverse Fourier transform of the predicted output, and panels **(c, f)** depict the corresponding ground truth.

plitudes and phases $\hat{c}_i = (\hat{a}_i, \hat{\phi}_i)$, respectively. The output of the phase-predicting layer is additionally passed through the *tanh*-activation function to ensure that all normalized phases are in $[-1, 1]$.

4.2.2 Fourier Coefficient Loss

I train the Fourier Image Transformers (FITs) with a loss function consisting of two terms, (i) the amplitude loss

$$\mathcal{L}_{amp}(\hat{a}_i, a_i) = 1 + (\hat{a}_i - a_i)^2, \tag{4.10}$$

computed between the predicted amplitudes $\hat{a}_i$ and the target amplitudes a_i, and (ii) the phase loss

$$\mathcal{L}_{\angle}(\hat{\phi}_i, \phi_i) = 2 - cos(\hat{\phi}_i - \phi_i), \tag{4.11}$$

with $\hat{\phi}_i$ the predicted phase and ϕ_i the corresponding target phase.

The final Fourier coefficient loss $\mathcal{L}_{\mathcal{FC}}$ is the multiplicative combination of both individual losses, given by

$$\mathcal{L}_{\mathcal{FC}}(\hat{\mathbf{C}}, \mathbf{C}) = \frac{1}{N} \sum_{i=0}^{N} \mathcal{L}_{amp}(\hat{a}_i, a_i) \cdot \mathcal{L}_{\angle}(\hat{\phi}_i, \phi_i). \tag{4.12}$$

4.3 FIT for Super-Resolution

The FDE I described above is a sequential representation of an input image $\boldsymbol{x}$, for which each prefix *pre* encodes a reduced resolution image $\boldsymbol{x}_{pre}$. Hence, using FDE image sequences, I can train an auto-regressive Fourier Image Transformer that takes an encoded sequence $\boldsymbol{E} = [e_1, \ldots, e_{N-1}]$ as input and predicts the FDE sequence $\boldsymbol{Z} = [z_2, \ldots, z_N]$. This FIT for super-resolution (FIT: SRes) is trained w.r.t. the correct target sequence $\boldsymbol{C} = [c_2, \ldots, c_N]$, using the previously introduced Fourier coefficient loss $\mathcal{L}_{\mathcal{FC}}$, which is computed using the back-transformed predicted amplitude and phase values $\hat{\boldsymbol{C}} = [\hat{c}_2, \ldots, \hat{c}_N]$, where $\hat{c}_i = (\hat{a}_i, \hat{\phi}_i)$, as explained in Section 4.2.1.

Once the transformer is trained, a prefix of a complete FDE sequence $\boldsymbol{E}_{pre} = [e_1, \ldots, e_{|pre|}]$ is used to condition the transformer, which is then used in iterations to auto-regressively predict the missing part of the complete sequence $\boldsymbol{E} = [e_1, \ldots, e_N]$, *i.e.* filling in predicted high-frequency information not contained in $\boldsymbol{E}_{pre}$ (see Figure 4.1).

Note that the proposed super-resolution setup operates exclusively on Fourier domain encoded data. All final prediction images $\hat{\boldsymbol{x}}$ are generated by computing the inverse Fourier transform on predictions $\hat{\boldsymbol{C}}$, which are rearranged (rolled) into $\hat{\boldsymbol{X}}_{\boldsymbol{h}}$ and completed to a full predicted Fourier spectrum $\hat{\boldsymbol{X}}$, *i.e.* $\hat{\boldsymbol{x}} = \mathcal{F}^{-1}(\text{roll}(\hat{\boldsymbol{C}}))$.

4.3.1 Super-Resolution Data

MNIST (Y. LeCun and Cortes 2010): Cropped to 27×27 pixels with the default train-test split (in 60′000 and 10′000 images, respectively). The train images are further split into 55′000 samples for training and 5′000 validation images.

CelebA 128×128 (Z. Liu et al. 2015): Converted to gray scale and downscaled to 63×63 pixels. The images are randomly split into 20′000, 5′000 and 5′000 training, validation and test samples, respectively.

I evaluated all results using (*i*) Fourier Ring Correlation (FRC) (Van Heel et al. 1982) in Fourier space and (*ii*) the peak signal-to-noise ratio (PSNR) in image space.

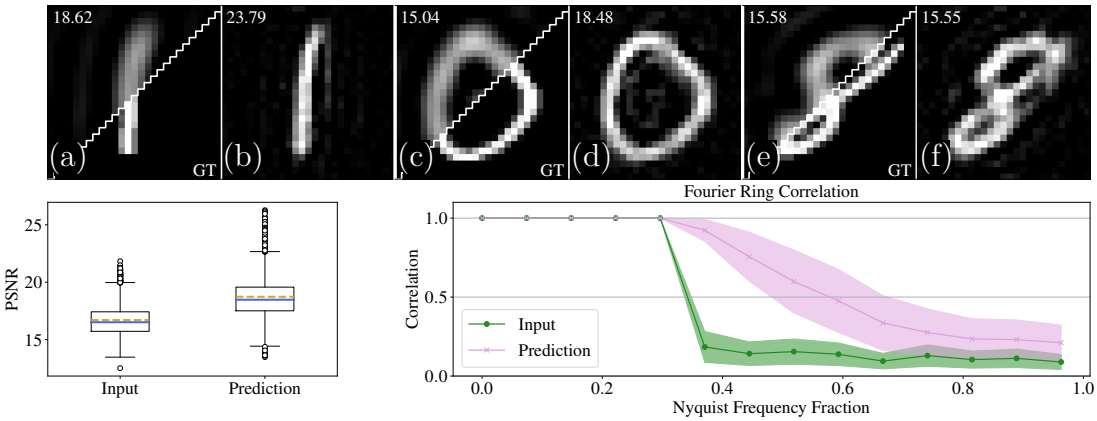

Figure 4.2: **Super-resolution results on MNIST.** The top left triangles of **(a, c, e)** show 3× binned (low-res) MNIST inputs. Conditioned by these inputs, our trained FIT auto-regressively generates results as shown in *(b, d, f)*. Inputs and predictions are labeled with the peak signal-to-noise ratio (PSNR) values computed w.r.t. ground truth images, shown in the lower-right half of (a, c, e). The examples in (a-f) correspond to the 98th, 50th and 2nd percentile in terms of obtained PSNR over all MNIST prediction results. Box-plots show the distribution of PSNR values of Fourier binned inputs and predicted outputs, respectively (mean in dashed gold and median in solid blue). The Fourier ring correlation plot shows how predicted Fourier coefficients are improving w.r.t. ground truth coefficients. Shaded areas correspond to +/-1 standard deviation. The correlation for the first 5 Fourier rings is 1 because these rings have been used as inputs to the FIT.

4.3.2 Super-Resolution Experiments

Super-Resolution Training

I used a $F = 256$ dimensional FDE, with the positional encoding being based on polar coordinates, *i.e.* Fourier coefficients of same frequency have the same radius. The FDE is passed to a *causal-linear* transformer (Katharopoulos et al. 2020) with 8 layers, 8 self-attention heads, a query and value dimensionality of 32, dropout of 0.1, attention dropout of 0.1, and a dimensionality of the feed-forward network of 1024.

This setup is trained auto-regressively, *i.e.* with a triangular attention mask. I used the rectified Adam optimizer (RAdam) (L. Liu et al. 2019) with an initial learning rate of 0.0001 and weight decay of 0.01 for 100 epochs. The batch size is 32 and the learning rate is halved on plateauing validation loss.

Super-Resolution Results

Quantitative results for all conducted super-resolution experiments on MNIST data are shown in Figure 4.2, where I show (i) 3× binned low-res MNIST input images corresponding to $\boldsymbol{E}_{pre} = [c_1, \ldots, c_{39}]$, also sketched in Figure 4.1, (ii) cor-

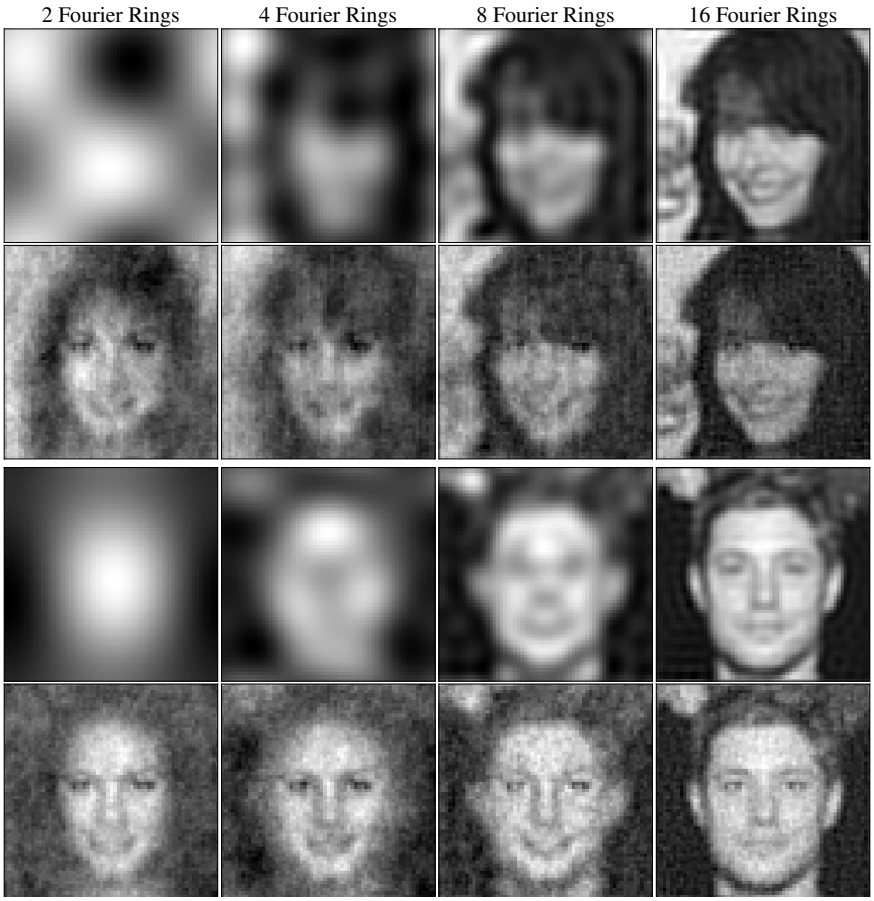

Figure 4.3: **Super-resolution results on CelebA.** On two input images (rows 1+2 and 3+4) we see predictions of a trained super-resolution FIT conditioned on 2, 4, 8 and 16 Fourier rings (columns). Upper rows show the iFFT of the input FDEs used to condition the FIT, lower rows depict the iFFT of the predicted results.

responding ground truth MNIST images, (*iii*) the predictions of the "FIT: SRes" network trained on the MNIST data as described in Section 4.3, (*iv*) two box-plots showing the distribution of PSNR values computed between the ground truth images and the downscaled inputs and predicted outputs, respectively, and (*v*) Fourier ring correlation plots showing the correlation between the predicted Fourier coefficients and the corresponding ground truth.

In Figure 4.3, I show two sequences of super-resolution results obtained with a FIT trained on the CelebA data. For each image, I conditioned the trained transformer on 2, 4, 8, and 16 Fourier rings, respectively. This corresponds to low-resolution images subject to 16×, 8×, 4×, and 2× binning in Fourier space, respectively.

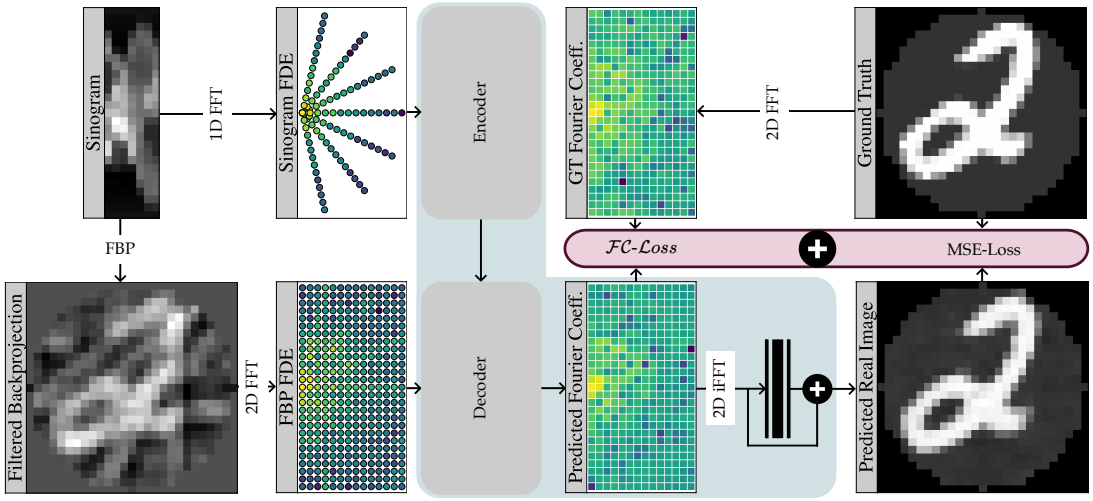

Figure 4.4: **FIT for computed tomography.** I propose an encoder-decoder based Fourier Image Transformer setup for tomographic reconstruction. In 2D computed tomography, 1D projections of an imaged sample (*i.e.* the columns of a sinogram) are back-transformed into a 2D image. A common method for this transformation is the filtered backprojection (FBP) (Kak et al. 2002; Ramesh et al. 1989). Since each projection maps to a line of coefficients in 2D Fourier space, a limited number of projections in a sinogram leads to visible streaking artefacts due to missing/unobserved Fourier coefficients. The idea of my FIT setup is to encode all information of a given sinogram and use the decoder to predict missing Fourier coefficients. The reconstructed image is then computed via an inverse Fourier transform (iFFT) of these predictions. In order to reduce high frequency fluctuations in this result, I introduce a shallow conv-block after the iFFT (shown in black). I trained this setup combining the $\mathcal{FC}$-$\mathcal{L}oss$, see Section 4.2.2, and a conventional MSE-loss between prediction and ground truth.

4.4 FIT for Tomography

My Fourier Image Transformer for tomograpic reconstruction (FIT: TRec) is based on an encoder-decoder transformer architecture as shown in Figure 4.4. As input to the encoder I use the Fourier Domain Encoding (FDE) of a raw sinogram $\boldsymbol{s}$. As described above, $\boldsymbol{s}$ consists of P pixel columns $[s_1, \ldots, s_P]$ of 1D projections of $\boldsymbol{x}$ at angles $[\alpha_1, \ldots, \alpha_P]$. The Fourier slice theorem states, see also Section 1.3, that the discrete 1D Fourier coefficients $\boldsymbol{C}_i = \mathcal{F}(s_i)$ coincide with the values of the 1D slice at angle α_i through the 2D Fourier spectrum $\mathcal{F}(\boldsymbol{x})$. To assemble the full FDE of a sinogram I need to combine all $\boldsymbol{C}_i$ with the adequate positional encoding (using polar coordinates) of all Fourier coefficients, as dictated by the Fourier slice theorem and sketched in Figure 4.4.

Hence, the encoder creates a latent space representation $\boldsymbol{Z}$ that encodes the full input sinogram $\boldsymbol{s}$. This latent space encoding is then given as input to the

decoder. The decoder is used to predict all Fourier coefficients $\hat{\boldsymbol{C}}$, such that the predicted reconstruction $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ can be computed by $\hat{\boldsymbol{x}} = \mathcal{F}^{-1}(\text{roll}(\hat{\boldsymbol{C}}))$, where roll arranges the 1D sequence back into a discrete 2D Fourier spectrum. This setup is called "FIT: TRec".

Additionally, I propose a variation of this procedure, called "FIT: TRec + FBP", where the decoder not only receives the latent space encoding $\boldsymbol{Z}$, but also FDEs of the Fourier coefficients $\boldsymbol{C}_{\text{FBP}} = \mathcal{F}(\text{FBP}(\boldsymbol{s}))$, where FBP denotes the function computing the filtered backprojection of a sinogram (see Figure 4.4). Note that the implementation of "FIT: TRec" coincides with "FIT: TRec + FBP", with FBP being replaced by a function ZERO which returns 0 for all inputs.

I train "FIT: TRec" and "FIT: TRec + FBP" using the $\mathcal{L}_{\mathcal{FC}}$-loss of Eq.4.12. Additionally, I introduced a residual convolution block consisting of two convolutional layers (3×3 followed by 1×1) with $d_{\text{conv}} = 8$ intermediate feature channels. This conv-block (conv) receives the inverse Fourier transform of the predicted Fourier coefficients $\hat{\boldsymbol{x}} = \mathcal{F}^{-1}(\text{roll}(\hat{\boldsymbol{C}}))$ as input and is trained using the MSE-loss between the predicted real-space image $\text{conv}(\hat{\boldsymbol{x}})$ and the known ground truth image $\boldsymbol{x}$. Hence, the full loss is the sum over $\mathcal{L}_{\mathcal{FC}}$ and the MSE-loss.

In order to speed up training, I start by feeding only a low-resolution subset of all Fourier coefficients $\boldsymbol{C}_i = \mathcal{F}(s_i)$ (and $\boldsymbol{C}_{\text{FBP}}$), and successively increase this subset over training until the full sets are used. This forces the FIT to first learn good low resolution features and later learn to add suitable high resolution predictions.

4.4.1 Computed Tomography Data

As described in Section 1.3, tomographic image reconstruction in 2D operates on a number of 1D projections of a given true object $\boldsymbol{x}$. I used a tomographic simulation process which is based on the work by Leuschner *et al.* (Leuschner et al. 2019). Furthermore, I chose the detector length to be equal to the width of the chosen object (*i.e.* ground truth image) to which the synthetic tomography pipeline is applied. To avoid spurious contributions to individual projections, I needed to set all pixel intensities outside the largest image-centered circle to 0 (hence, we see only circular images in Figures 4.5, 4.6 and 4.7).

MNIST (Y. LeCun and Cortes 2010): Data is split and preprocessed as described in Section 4.3.1. Additionally, for visualization purposes, I min-clipped all pixel intensities within the before-mentioned largest image-centered circle to 50. Finally, I used this data to compute $P = 7$ equally spaced projections which

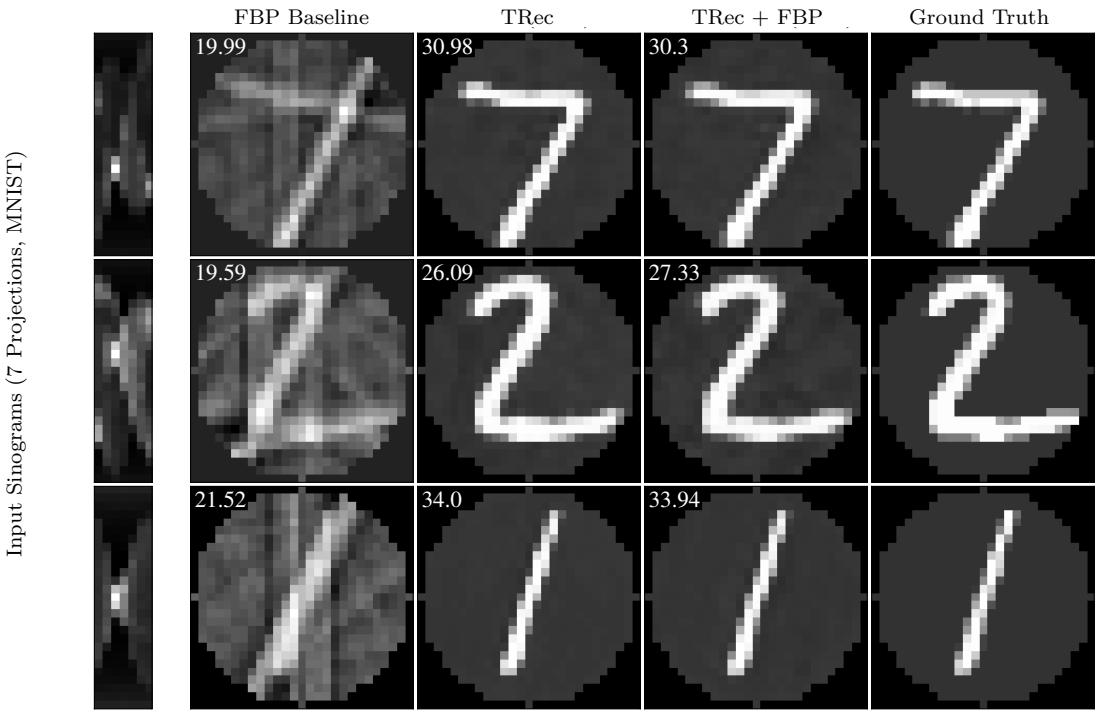

Figure 4.5: **Tomographic reconstruction results.** These are three qualitative results for the MNIST (Y. LeCun and Cortes 2010) dataset. From left to right, the input sinogram, reconstruction results obtained with filtered backprojection (FBP) (Kak et al. 2002; Ramesh et al. 1989), the results obtained with the "FIT: TRec" setup, the results obtained with the "FIT: TRec + FBP" setup, and the corresponding ground truth images are shown. In the top left corner of each reconstruction the peak signal-to-noise ratio (PSNR) with respect to the ground truth image is shown.

are assembled in sinograms $\boldsymbol{s}^j_{\text{MNIST}} = [s^{j,1}_{\text{MNIST}}, \ldots, s^{j,P}_{\text{MNIST}}]$.

Kanji (Clanuwat et al. 2018): Data is randomly split into 50′000 train, 5′000 validation and 5′000 test samples and all images are cropped to 63×63 pixels, which are otherwise processed as described for MNIST. Finally, I used this data to compute $P = 33$ equally spaced projections which are assembled in sinograms $\boldsymbol{s}^j_{\text{Kanji}} = [s^{j,1}_{\text{Kanji}}, \ldots, s^{j,P}_{\text{Kanji}}]$.

LoDoPaB (Leuschner et al. 2019): The original train- and validation-data is first reduced to 4′000 and 400 randomly chosen images respectively and for testing all 3′553 images are used. All selected images are downscaled to 111×111 pixels and I computed $P = 33$ equally spaced projections like for the Kanji data.

All tomographic reconstruction experiments with "FIT: TRec" and "FIT: TRec + FBP" and the FBP baseline are evaluated using peak signal-to-noise ratio (PSNR) w.r.t. available ground truth.

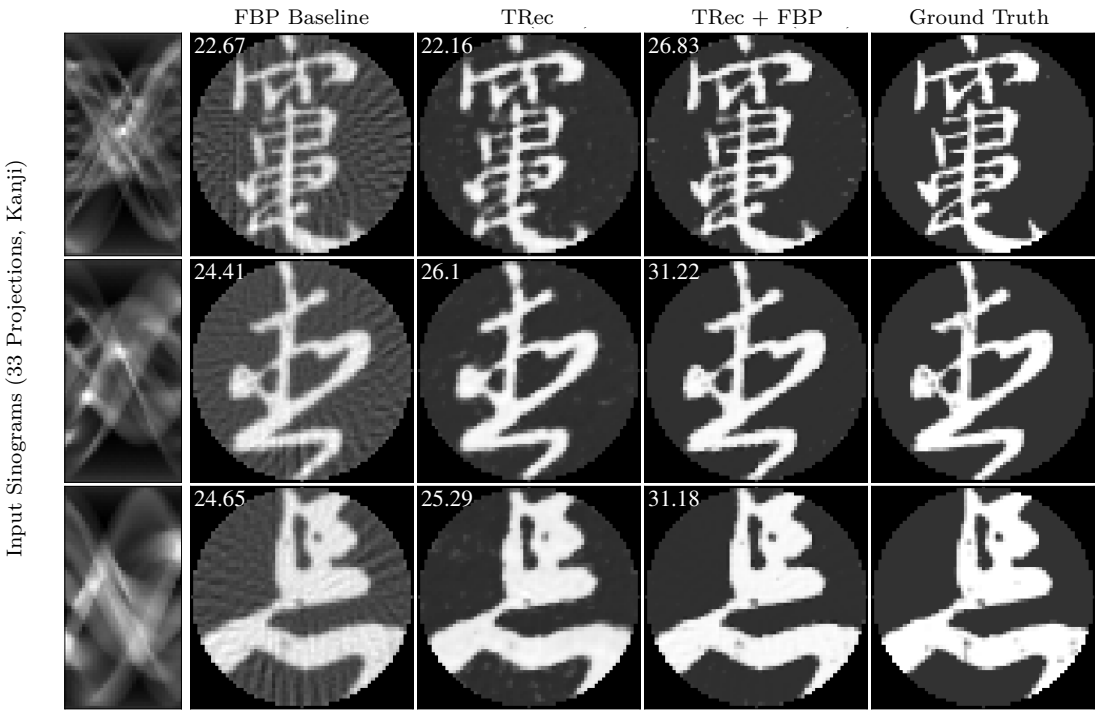

Figure 4.6: **Tomographic reconstruction results.** These are three qualitative results for the Kanji (Clanuwat et al. 2018) dataset. From left to right, the input sinogram, reconstruction results obtained with filtered backprojection (FBP) (Kak et al. 2002; Ramesh et al. 1989), the results obtained with the "FIT: TRec" setup, the results obtained with the "FIT: TRec + FBP" setup, and the corresponding ground truth images are shown. In the top left corner of each reconstruction the peak signal-to-noise ratio (PSNR) with respect to the ground truth image is shown.

4.4.2 Computed Tomography Experiments

Computed Tomography Training

Like before, I consistently used $F = 256$ dimensional FDEs, and employed the *linear* encoder and decoder method by Katharopoulos *et al.* (Katharopoulos et al. 2020). More specifically, I used 4 transformer layers, 8 self-attention heads per layer, a query and value dimensionality of 32, dropout of 0.1, attention dropout of 0.1, and a dimensionality of the feed-forward network of 1024. The residual conv-block has $d_{\text{conv}} = 8$ intermediate feature channels.

All networks are optimized using RAdam (L. Liu et al. 2019), with an initial learning rate of 0.0001 and weight decay of 0.01 for 300 (MNIST), 120 (Kanji), and 350 (LoDoPaB) epochs. The batch size is 32. The learning rate is halved on plateauing validation loss.

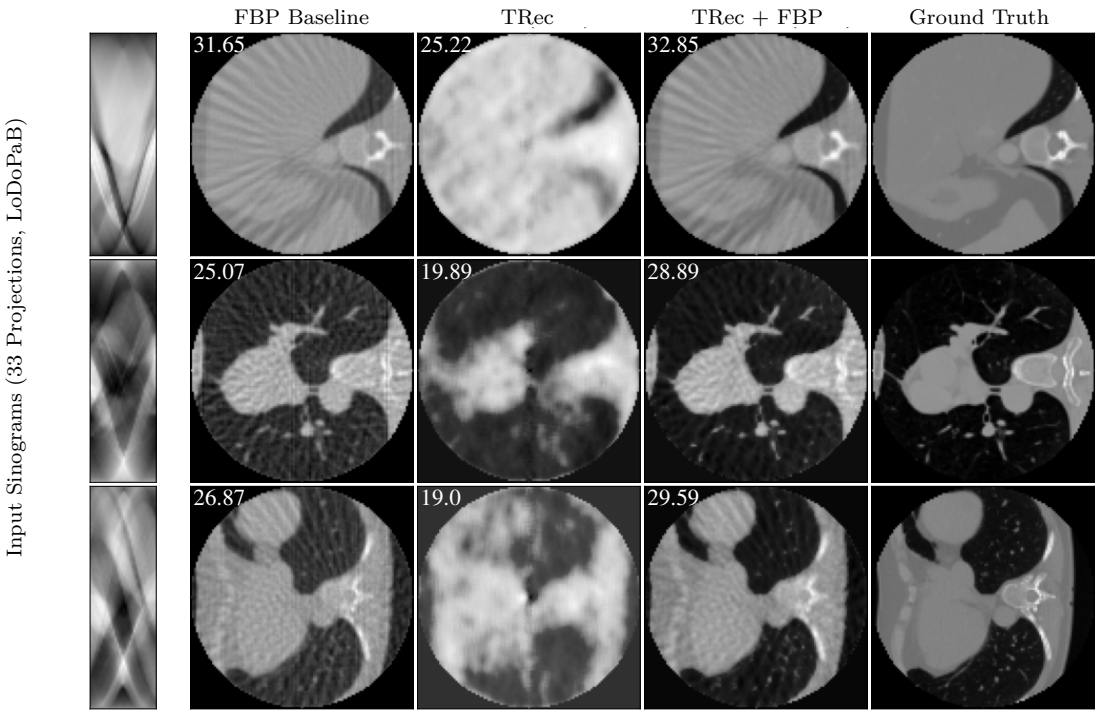

Figure 4.7: **Tomographic reconstruction results.** These are three qualitative results for the LoDoPaB (Leuschner et al. 2019) dataset. From left to right, the input sinogram, reconstruction results obtained with filtered backprojection (FBP) (Kak et al. 2002; Ramesh et al. 1989), the results obtained with the "FIT: TRec" setup, the results obtained with the "FIT: TRec + FBP" setup, and the corresponding ground truth images are shown. In the top left corner of each reconstruction the peak signal-to-noise ratio (PSNR) with respect to the ground truth image is shown.

Ablation Studies

I propose two ablation setups for all tomographic reconstruction experiments.

First, I ask what influence the encoded latent space information $\boldsymbol{Z}$, *i.e.* the output of the encoded sinogram, has on the quality of the overall reconstruction $\hat{\mathbf{x}}$. To that end, I performed ablation experiments for all 3 datasets, for which I do not feed $\boldsymbol{Z}$ to the decoder. Technically this is implemented by replacing the decoder by an encoder network (since only one input remains to be fed). I called these experiments "Only FBP".

The second ablation study asks, to what degree the conv-block contributes to the overall reconstruction performance, *i.e.* I want to verify that the convolution block alone is not sufficient to solve the task at hand. Hence, I trained the conv-block on pairs of images $(\mathrm{FBP}(\boldsymbol{s}), \boldsymbol{x})$, *i.e.* the filtered backprojection of sinograms $\boldsymbol{s}$ and their corresponding ground truth images $\boldsymbol{x}$. I labelled these experiments

Dataset	Method	PSNR
MNIST (Y. LeCun and Cortes 2010)	Baseline: FBP	17.87
	FIT: TRec + FBP (Ours)	27.85
	FIT: TRec (Ours)	**27.90**
	Ablation: Only FBP	26.89
	Ablation: Only Conv-Block	22.53
Kanji (Clanuwat et al. 2018)	Baseline: FBP	22.06
	FIT: TRec + FBP (Ours)	**30.72**
	FIT: TRec (Ours)	25.99
	Ablation: Only FBP	30.49
	Ablation: Only Conv-Block	26.92
LoDoPaB (Leuschner et al. 2019) (downscaled)	Baseline: FBP	26.89
	FIT: TRec + FBP (Ours)	**30.98**
	FIT: TRec (Ours)	21.90
	Ablation: Only FBP	30.74
	Ablation: Only Conv-Block	30.70

Table 4.1: **Quantitative tomographic reconstruction results.** I report the average peak signal-to-noise ratio (PSNR) with respect to ground truth, for each of the three used datasets. For each dataset, I compare the results of the "FIT: TRec + FBP" and "FIT: TRec" setups to results obtained with the filtered backprojection (FBP) (Kak et al. 2002; Ramesh et al. 1989) baseline, and the two ablation studies described in Section 4.4.2.

"Only Conv-Block".

For all ablation experiments all hyper-parameters not explicitly mentioned above are kept unchanged.

Tomographic Reconstruction Results

Qualitative tomographic reconstruction results for all three datasets I used are shown in Figures 4.5, 4.6 and 4.7. For each dataset, I show three input sinograms, the reconstruction baseline obtained via filtered backprojection (FBP), results obtained via "FIT: TRec"and "FIT: TRec + FBP", and the corresponding ground truth images. In Table 4.1, PSNR numbers for all three datasets using the FBP baseline, the "FIT: TRec + FBP" and "FIT: TRec" training setups, and both ablation studies are given. All code used to reproduce the reported results is available on GitHub[3].

[3] https://github.com/juglab/FourierImageTransformer

4.5 Discussion

I proposed the idea of Fourier Domain Encodings (FDEs), a novel sequential image encoding, for which each prefix represents the whole image at reduced resolution, and demonstrated the utility of FDEs for solving two common image processing tasks with Transformer networks, *i.e.* super-resolution and tomographic image reconstruction.

For the super-resolution task I showed that Fourier Image Transformer can be trained to, when conditioned on an FDE corresponding to a low-resolution input image, auto-regressively predict an extended FDE sequence that can be back-transformed into a higher resolution output. It is obvious, the information required to generate a higher resolution image must be stored in the trained network, and I have shown in Figure 4.3, how this learned prior completes very low to moderate resolution inputs in sensible ways. It is curious to see that eyes are the first high-resolution structures filled in by the trained FIT. I believe that this is a direct consequence of all training images being registered such that the eyes are consistently at the same location.

For the tomographic reconstruction task, I employed an encoder-decoder transformer that encodes a given FDE sequence corresponding to a given sinogram and can then be used to predict Fourier coefficients at arbitrary query locations. I used the decoder to predict all Fourier coefficients of a fully reconstructed image, which we could then visualize via inverse Fourier transformation (iFFT). I noticed that introducing a shallow residual convolution block after the iFFT reduces unwanted high frequency fluctuations in predicted results. While I see that this procedure leads to very convincing results on MNIST, for more complex datasets, results quickly deteriorate. Hence, I proposed to additionally feed Fourier coefficients obtained by filtered backprojection (FBP) into the decoder. This leads to much improved results that outperform the FBP baseline, showing that the FIT does contribute to solving the reconstruction task.

My results show that Transformers, currently the dominant approach for virtually all NLP tasks, can successfully be applied to complex and relevant tasks in computer vision. While this is encouraging, I see a plethora of possibilities for future improvements. For example, the Transformers I used are rather small. I believe that an up-scaled version of the training setups with more attention heads and more layers would already lead to much improved results. Still, I also believe that there is plenty of room for methodological improvements that do not require more computational resources, making this line of research also accessible to many other research labs around the globe.

Chapter 5

Conclusions and Outlook

With this thesis I introduced content-aware image denoising techniques to the field of electron microscopy (EM). I presented Noise2Void, the first self-supervised image denoising approach based on neural networks. Then I proposed novel ideas to image reconstruction based on Transformer networks. In the next paragraphs I will revisit the individual contributions and outline some possible future works.

In the introduction of this thesis we looked at EM as tool for bio-medical research. In scanning electron microscopy (SEM) a focused electron beam is used to scan a sample row by row (Collett 1970). Whole tissues are imaged with volumetric SEM methods, where the sample is slice-wise imaged with destructive (SBF-SEM, FIB-SEM) or non-destructive (array tomography or serial-section SEM) methods. Common to all SEM methods is that the imaging quality is dependent on the scanning speed, with slower scanning speeds resulting in higher SNR images. This results in long acquisition times which makes large connectomics projects expensive and time consuming. With SEM-CARE I showed how supervised image denoising methods from fluorescence microscopy can be translated to SEM image data. With carefully acquired pairs of slow and fast scanned SEM images we can train a supervised CARE network, which we can later apply to unseen noisy, fast scanned image data. This approach results in a potential 40- to 50-fold imaging speedup. Interesting followup work to SEM-CARE could combine the information from different SEM detectors. Such networks could pool information from backscattered and secondary electrons with elemental identification data obtained from measured X-rays.

Next, we looked at cryo transmission electron microscopy (cryo TEM) tomograms. In cryo TEM the acquisition of high quality images is impossible due to missing contrast agents and beam induced sample damage (Knapek and Dubochet

1980). The beam induced sample damage limits the total electron dose to which the sample can be exposed, hence any cryo TEM acquisition is dominated by Poisson noise. But modern cryo TEM detectors acquire short bursts of images (movies) to avoid motion blur, which are aligned and summed into a single observation. With cryoCARE I take advantage of this procedure and instead of summing up the aligned frames up into a single image I split them in two summed subsets. The first image contains all even movie frames and the second image contains all odd movie frames. This is done for each tilt-angle and results in an even- and odd-frames tilt-series. From these two tilt-series two tomograms are reconstructed containing the same signal but different independent noise contributions. Hence, the requirements for Noise2Noise (Lehtinen et al. 2018) training are fulfilled and these images can be used to train supervised CARE networks. After training both tomograms are denoised and voxel-wise averaged resulting in a high contrast and high SNR tomogram. These tomograms are used for manual data browsing and particle picking and I have shown that down stream processes benefit as well. I have integrated cryoCARE into Scipion, an EM image processing framework, which eases the usability of cryoCARE significantly. Furthermore, I am looking forward to future developments of cryoCARE like methods. I see potential in optimizing the current reconstruction algorithms jointly with cryoCARE or similar methods.

In Chapter 3, I have presented Noise2Void a novel training scheme, which only requires single noisy observations to train content-aware image denoising networks. Noise2Void introduces the concept of a blind-spot network, which has access to all pixels in its receptive field except for the center pixel. The idea is, that the signal in an image is not pixel-wise independent and the network is able to predict the missing pixel value by looking at the local neighborhood. For the noise the assumption is pixel-wise independence, in other words the noise contribution for a single pixel is not predictable by looking at the surrounding pixels. I proposed an efficient implementation of Noise2Void, which simulates the blind-spot by replacing single pixels in the input with random values and only computing the loss for these perturbed pixels. The implementation is publicly available as a Python package and additionally a one-click solution exists in Fiji (Schindelin et al. 2012). Personally, I find it interesting to see that current SOTA natural language processing (NLP) approaches like BERT (Devlin et al. 2018) use similar techniques *i.e.* they mask individual words in the input and train the network to predict them. I believe that research in self-supervised learning will move more into the spotlight in the following years. We already see high quality attention maps emerge from self-supervised Vision Transformer training (Caron et al. 2021) and I would not be surprised if similar methods will be used in the future to train self-supervised segmentation approaches. Regarding self-supervised image denoising,

a big challenge is still presented with structural noises and reconstruction artefacts from for example tomography. I am expecting that novel network architectures and loss functions will be developed for self-supervised training, such that we can eventually disentangle ground truth signal, structured noise and pixel-wise independent noise by accessing the latent space encodings of such neural networks. In this regard I am especially looking forward to followup work based on Mangal Prakashs recent publication (Prakash, Krull, et al. 2021).

In Chapter 4, we looked into Transformer networks and their application towards image restoration. Transformers work on 1D sequences, which makes them well suited for sequential data domains like text. However, the recent success of Transformers (Devlin et al. 2018; Radford et al. 2018) has inspired the computer vision community to take a closer look and apply Transformers to image classification (Dosovitskiy et al. 2020; Ramachandran et al. 2019) and pixel-by-pixel image generation (M. Chen et al. 2020). This pixel-by-pixel image generation feels artificial to me, because images almost never come cut in half and generating the second half is seldom the problem. A much more common problem is super-resolution, where we have access to a low-resolution image and would like to restore a corresponding high-resolution image. This lead me to the development of the Fourier Domain Encoding (FDE), which takes the Fourier transformation of an image and brings it in a 1D sequence where each prefix corresponds to a lower resolution version of the encoded image. I used these encodings to present a proof of concept for Fourier Image Transformer (FIT) trained for super-resolution. Then I considered the missing wedge artefacts in tomography. The missing wedge artefacts in tomographic imaging originate due to sparse-view imaging. Sparse-view imaging is used to keep the total exposure of the imaged sample to a minimum, by only acquiring a limited number of projection images. However, tomographic reconstructions from sparse-view acquisitions are affected by missing wedge artefacts, characterized by missing wedges in the Fourier space and visible as streaking artefacts in real image space. All methods dealing with these missing wedge artefacts do so in real image space *i.e.* they try to remove an artefact from a reconstructed image. However with Fourier Image Transformer for tomograpic reconstruction (FIT: TRec) I presented a method which aims at filling in the missing data directly in Fourier space preventing the artefacts of occurring in the first place. FIT for tomographic reconstruction is for now a proof of concept, which shows how Transformer architectures in combination with FDEs can be used to train tomographic reconstruction networks. Using such an approach would also be interesting for sub-tomogram averaging in cryo TEM. The closest work to this is cryoDRGN (Zhong et al. 2021), which uses a fully connected network to build sub-tomogram averages of molecules with multiple conformations.

Finally, I want to emphasize the important work done by the labs of

Anna Kreshuk with ilastik (Berg et al. 2019), Ricardo Henriques with ZeroCostDL4Mic (von Chamier et al. 2021), and Florian Jug with CSBDeep (Weigert et al. 2018). These frameworks and projects are crucial to make deep learning available to a wider community of users and enabling bio-image analysts to use SOTA deep learning tools on their own data. In my opinion, one of the most important collaborations to make deep learning reproducible in the future is the bioimage modelzoo project[1]. Personally, I feel extremely lucky to be part of such a great community which helps each other to develop great bioimage analysis tools and invests the time to bring them to our customers – the biologists in the labs around the world.

Thank you for reading.

[1] https://bioimage.io/#/

Bibliography

Abbe, E. (1873). Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung. *Archiv für mikroskopische Anatomie*, *9*(1), 413–468.

Adler, J., & Öktem, O. (2018). Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, *37*(6), 1322–1332.

Baena, V., Schalek, R. L., Lichtman, J. W., & Terasaki, M. (2019). Serial-section electron microscopy using automated tape-collecting ultramicrotome (atum). *Methods in cell biology*, *152*, 41–67.

Batson, J., & Royer, L. (2019). Noise2self: Blind denoising by self-supervision. *International Conference on Machine Learning*, 524–533.

Bello, I. (2021). Lambdanetworks: Modeling long-range interactions without attention. *International Conference on Learning Representations*. https://openreview.net/forum?id=xTJEN-ggl1b

Bepler, T., Kelley, K., Noble, A. J., & Berger, B. (2020). Topaz-denoise: General deep denoising models for cryoem and cryoet. *Nature communications*, *11*(1), 1–12.

Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A. J., & Berger, B. (2019). Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*. https://doi.org/10.1038/s41592-019-0575-8

Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M. et al. (2019). Ilastik: Interactive machine learning for (bio) image analysis. *Nature Methods*, *16*(12), 1226–1232.

Betzig, E., Patterson, G. H., Sougrat, R., Lindwasser, O. W., Olenych, S., Bonifacino, J. S., Davidson, M. W., Lippincott-Schwartz, J., & Hess, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, *313*(5793), 1642–1645.

Bhella, D. (2019). Cryo-electron microscopy: An introduction to the technique, and considerations when working to establish a national facility. *Biophysical reviews*, *11*(4), 515–519.

Bracewell, R. N. (1956). Strip integration in radio astronomy. *Australian Journal of Physics*, *9*(2), 198–217.

Bredies, K., Kunisch, K., & Pock, T. (2010). Total generalized variation. *SIAM Journal on Imaging Sciences*, *3*(3), 492–526.

Broaddus, C., Krull, A., Weigert, M., Schmidt, U., & Myers, E. (2020). Removing structured noise with self-supervised blind-spot networks. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI 2020)*.

Buades, A., Coll, B., & Morel, J.-M. (2005). A non-local algorithm for image denoising. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *2*, 60–65.

Buchholz, T.-O., Jordan, M., Pigino, G., & Jug, F. (2019). Cryo-care: Content-aware image restoration for cryo-transmission electron microscopy data. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 502–506.

Buchholz, T.-O., Krull, A., Shahidi, R., Pigino, G., Jékely, G., & Jug, F. (2019). Content-aware image restoration for electron microscopy. *Methods in cell biology*, *152*, 277–289.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Chambolle, A., & Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, *40*(1), 120–145.

Chen, H., Zhang, Y., Kalra, M. K., Lin, F., Chen, Y., Liao, P., Zhou, J., & Wang, G. (2017). Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, *36*(12), 2524–2535.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. *International Conference on Machine Learning*, 1691–1703.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). *Deep learning for classical japanese literature*. arXiv: cs.CV/1812.01718 [cs.CV].

Collett, B. M. (1970). Scanning electron microscopy: A review and report of research in wood science. *Wood and Fiber Science*, *2*(2), 113–133.

Crosby, K., Eberle, A. L., & Zeidler, D. (2016). Multi-beam sem technology for high throughput imaging. *MRS Advances*, *1*(26), 1915–1920.

Cybulski, J. S., Clements, J., & Prakash, M. (2014). Foldscope: Origami-based paper microscope. *PloS one*, *9*(6), e98781.

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, *16*(8), 2080–2095.

de la Rosa-Trevín, J., Quintana, A., del Cano, L., Zaldívar, A., Foche, I., Gutiérrez, J., Gómez-Blanco, J., Burguet-Castell, J., Cuenca-Alba, J., Abrishami, V., Vargas, J., Otón, J., Sharov, G., Vilas, J., Navas, J., Conesa, P., Kazemi, M., Marabini, R., Sorzano, C., & Carazo, J. (2016). Scipion: A software framework toward integration, reproducibility and validation in 3d electron microscopy. *Journal of Structural Biology*, *195*(1), 93–99. https://doi.org/https://doi.org/10.1016/j.jsb.2016.04.010

Denk, W., & Horstmann, H. (2004). Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biol*, *2*(11), e329.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dobro, M. J., Melanson, L. A., Jensen, G. J., & McDowall, A. W. (2010). Plunge freezing for electron cryomicroscopy. *Methods in enzymology*, *481*, 63–82.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dubochet, J. [Jacques], Adrian, M., Chang, J.-J., Homo, J.-C., Lepault, J., McDowall, A. W., & Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. *Quarterly reviews of biophysics*, *21*(2), 129–228.

Dumoulin, V., & Visin, F. (2016). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.

Finnoff, W., Hergert, F., & Zimmermann, H. G. (1993). Improving model selection by nonconvergent methods. *Neural Networks*, *6*(6), 771–783.

Fischer, A. H., Henrich, T., & Arendt, D. (2010). The normal development of platynereis dumerilii (nereididae, annelida). *Frontiers in zoology*, *7*(1), 1–39.

Frangakis, A. S., & Hegerl, R. (2001). Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *Journal of structural biology*, *135*(3), 239–250.

Gan, L., & Jensen, G. J. (2012). Electron tomography of cells. *Quarterly reviews of biophysics*, *45*(1), 27–56.

Gilbert, P. (1972). Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of theoretical biology*, *36*(1), 105–117.

Golding, C. G., Lamboo, L. L., Beniac, D. R., & Booth, T. F. (2016). The scanning electron microscope in microbiology and diagnosis of infectious disease. *Scientific reports*, *6*(1), 1–8.

Gordon, R., Bender, R., & Herman, G. T. (1970). Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *Journal of theoretical Biology*, *29*(3), 471–481.

Guerra, J. M. (1995). Super-resolution through illumination by diffraction-born evanescent waves. *Applied physics letters*, *66*(26), 3555–3557.

Hauptmann, A., Lucka, F., Betcke, M., Huynh, N., Adler, J., Cox, B., Beard, P., Ourselin, S., & Arridge, S. (2018). Model-based learning for accelerated, limited-view 3-d photoacoustic tomography. *IEEE transactions on medical imaging*, *37*(6), 1382–1393.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*, 770–778.

Hell, S. W., & Wichmann, J. (1994). Breaking the diffraction resolution limit by stimulated emission: Stimulated-emission-depletion fluorescence microscopy. *Optics letters*, *19*(11), 780–782.

Heumann, J. M., Hoenger, A., & Mastronarde, D. N. (2011). Clustering and variance maps for cryo-electron tomography using wedge-masked differences. *J Struct Biol*, *175*(3), 288–299. https://doi.org/10.1016/j.jsb.2011.05.011

Heymann, J. A., Hayles, M., Gestmann, I., Giannuzzi, L. A., Lich, B., & Subramaniam, S. (2006). Site-specific 3d imaging of cells and tissues with a dual beam microscope. *Journal of structural biology*, *155*(1), 63–73.

Honzátko, D., Bigdeli, S. A., Türetken, E., & Dunbar, L. A. (2020). Efficient blind-spot neural network architecture for image denoising. *2020 7th Swiss Conference on Data Science (SDS)*, 59–60.

Horstmann, H., Körber, C., Sätzler, K., Aydin, D., & Kuner, T. (2012). Serial section scanning electron microscopy (s 3 em) on silicon wafers for ultrastructural volume imaging of cells and tissues. *PloS one*, *7*(4), e35172.

Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J., & Patwardhan, A. (2016). Empiar: A public archive for raw electron microscopy image data. *Nature methods*, *13*(5), 387.

Jain, V., & Seung, S. (2009). Natural image denoising with convolutional networks. *Advances in Neural Information Processing Systems*, 769–776.

Jin, K. H., McCann, M. T., Froustey, E., & Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, *26*(9), 4509–4522.

Jonić, S., Sorzano, C., & Boisset, N. (2008). Comparison of single-particle analysis and electron tomography approaches: An overview. *Journal of Microscopy*, *232*(3), 562–579.

Kak, A. C., Slaney, M., & Wang, G. (2002). Principles of computerized tomographic imaging.

Kapuscinski, J. (1995). Dapi: A dna-specific fluorescent probe. *Biotechnic & Histochemistry*, *70*(5), 220–233.

Kasthuri, N., Hayworth, K. J. [Kenneth Jeffrey], Berger, D. R., Schalek, R. L., Conchello, J. A., Knowles-Barley, S., Lee, D., Vázquez-Reina, A., Kaynig, V., Jones, T. R. et al. (2015). Saturated reconstruction of a volume of neocortex. *Cell*, *162*(3), 648–661.

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. *Proceedings of the International Conference on Machine Learning (ICML)*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In Y. Bengio & Y. LeCun (Eds.), *Iclr*. http://dblp.uni-trier.de/db/conf/iclr/iclr2014.html#KingmaW13

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Klein, S., Staring, M., Murphy, K., Viergever, M. A., & Pluim, J. P. (2009). Elastix: A toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, *29*(1), 196–205.

Knapek, E., & Dubochet, J. (1980). Beam damage to organic material is considerably reduced in cryo-electron microscopy. *Journal of molecular biology*, *141*(2), 147–161.

Köhler, A. (1893). Ein neues beleuchtungsverfahren für mikrophotographische zwecke. *Zeitschrift für wissenschaftliche Mikroskopie und für Mikroskopische Technik*, *10*(4), 433–440.

Kremer, J. R., Mastronarde, D. N., & McIntosh, J. R. (1996). Computer visualization of three-dimensional image data using imod. *Journal of structural biology*, *116*(1), 71–76.

Kriss, T. C., & Kriss, V. M. (1998). History of the Operating Microscope: From Magnifying Glass to Microneurosurgery. *Neurosurgery*, *42*(4), 899–907. https://doi.org/10.1097/00006123-199804000-00116

Krull, A., Buchholz, T.-O., & Jug, F. (2019). Noise2void-learning denoising from single noisy images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2129–2137.

Krull, A., Vičar, T., Prakash, M., Lalit, M., & Jug, F. (2020). Probabilistic noise2void: Unsupervised content-aware denoising. *Frontiers in Computer Science*, *2*, 5. https://doi.org/10.3389/fcomp.2020.00005

Laine, S., Karras, T., Lehtinen, J., & Aila, T. (2019). High-quality self-supervised deep image denoising. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/2119b8d43eafcf353e07d7cb5554170b-Paper.pdf

LeCun, Y., & Cortes, C. (2010). MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/. http://yann.lecun.com/exdb/mnist/

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient backprop. *Neural networks: Tricks of the trade* (pp. 9–48). Springer.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., & Aila, T. (2018). Noise2Noise: Learning image restoration without clean data. *ICML*, 2965–2974.

Leuschner, J., Schmidt, M., Baguer, D. O., & Maaß, P. (2019). The lodopab-ct dataset: A benchmark dataset for low-dose ct reconstruction methods. *arXiv preprint arXiv:1910.01113*.

Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., Agard, D. A., & Cheng, Y. (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em. *Nature methods*, *10*(6), 584.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of International Conference on Computer Vision (ICCV)*.

Maier, A., Steidl, S., Christlein, V., & Hornegger, J. (2018). Medical imaging systems: An introductory guide.

Mao, X., Shen, C., & Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 2802–2810.

Möckl, L., Lamb, D. C., & Bräuchle, C. (2014). Super-resolved fluorescence microscopy: Nobel prize in chemistry 2014 for eric betzig, stefan hell, and william e. moerner. *Angewandte Chemie International Edition*, *53*(51), 13972–13977.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Icml*.

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P. M., Grigoras, I. T., Malinauskaite, L., Malinauskas, T., Miehling, J. et al. (2020). Single-particle cryo-em at atomic resolution. *Nature*, *587*(7832), 152–156.

Nicastro, D., Schwartz, C., Pierson, J., Gaudette, R., Porter, M. E., & McIntosh, J. R. (2006). The molecular architecture of axonemes revealed by cryoelectron tomography. *Science*, *313*(5789), 944–948. https://doi.org/10.1126/science.1128618

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), 62–66.

Palovcak, E., Asarnow, D., Campbell, M. G., Yu, Z., & Cheng, Y. (2020). Enhancing the signal-to-noise ratio and generating contrast for cryo-em images with convolutional neural networks. *IUCrJ*, *7*(6).

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. *International Conference on Machine Learning*, 4055–4064.

Pluk, H., Stokes, D., Lich, B., Wieringa, B., & Fransen, J. (2009). Advantages of indium–tin oxide-coated glass slides in correlative scanning electron microscopy applications of uncoated cultured cells. *Journal of microscopy*, *233*(3), 353–363.

Prakash, M., Krull, A., & Jug, F. (2021). Fully unsupervised diversity denoising with convolutional variational autoencoders. *International Conference on Learning Representations*. https://openreview.net/forum?id=agHLCOBM5jP

Prakash, M., Lalit, M., Tomancak, P., Krul, A., & Jug, F. (2020). Fully unsupervised probabilistic noise2void. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 154–158.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Akad. Wiss.*, *69*, 262–277.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/3416a75f4cea9109507cacd8e2f2aefc-Paper.pdf

Ramesh, G., Srinivasa, N., & Rajgopal, K. (1989). An algorithm for computing the discrete radon transform with some applications. *Fourth IEEE Region 10 International Conference TENCON*, 78–81.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 234–241.

Roth, S., & Black, M. J. (2005). Fields of experts: A framework for learning image priors. *CVPR*, *2*, 860–867.

Roth, S., & Black, M. J. (2009). Fields of experts. *International Journal of Computer Vision*, *82*(2), 205.

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, *60*(1-4), 259–268.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Rust, M. J., Bates, M., & Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nature methods*, *3*(10), 793–796.

Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S.-y., Hayworth, K. J. [Kenneth J], Huang, G. B., Shinomiya, K., Maitlin-Shepard, J., Berg, S. et al. (2020). A connectome and analysis of the adult drosophila central brain. *Elife*, *9*, e57443.

Scheres, S. H. (2012). Relion: Implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, *180*(3), 519–530.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B. et al. (2012). Fiji: An open-source platform for biological-image analysis. *Nature methods*, *9*(7), 676–682.

Shami, G. J., Cheng, D., & Braet, F. (2019). Chapter 2 - expedited large-volume 3-d sem workflows for comparative microanatomical imaging. In T. Müller-Reichert & G. Pigino (Eds.), *Three-dimensional electron microscopy* (pp. 23–39). Academic Press. https://doi.org/https://doi.org/10.1016/bs.mcb.2019.03.012

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Sines, G., & Sakellarakis, Y. A. (1987). Lenses in antiquity. *American Journal of Archaeology*, *91*(2), 191–196. http://www.jstor.org/stable/505216

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., & Ludtke, S. J. (2007). Eman2: An extensible image processing suite for electron microscopy. *Journal of structural biology*, *157*(1), 38–46.

Tappen, M. F., Liu, C., Adelson, E. H., & Freeman, W. T. (2007). Learning gaussian conditional random fields for low-level vision. *CVPR*, 1–8.

Tegunov, D., & Cramer, P. (2019). Real-time cryo-electron microscopy data preprocessing with warp. *Nature methods*, *16*(11), 1146–1152.

Tegunov, D., Xue, L., Dienemann, C., Cramer, P., & Mahamid, J. (2021). Multi-particle cryo-em refinement with m visualizes ribosome-antibiotic complex at 3.5 å in cells. *Nature Methods*, *18*(2), 186–193.

Thevenaz, P., Ruttimann, U. E., & Unser, M. (1998). A pyramid approach to subpixel registration based on intensity. *IEEE transactions on image processing*, *7*(1), 27–41.

Ulman, V., Maška, M., Magnusson, K. E., Ronneberger, O., Haubold, C., Harder, N., Matula, P. [Pavel], Matula, P. [Petr], Svoboda, D., Radojevic, M. et al. (2017). An objective comparison of cell-tracking algorithms. *Nature methods*, *14*(12), 1141.

Van Heel, M., Keegstra, W., Schutter, W., & Van Bruggen, E. (1982). Arthropod hemocyanin structures studied by image analysis. *Life Chem. Rep. Suppl*, *1*, 69–73.

van Zuylen, J. (1981). The microscopes of antoni van leeuwenhoek. *Journal of microscopy*, *121*(3), 309–328.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008.

von Chamier, L., Laine, R. F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E., Lerche, M., Hernández-Pérez, S., Mattila, P. K., Karinou, E. et al. (2021). Democratising deep learning for microscopy with zerocostdl4mic. *Nature communications*, *12*(1), 1–18.

Voss, N., Yoshioka, C., Radermacher, M., Potter, C., & Carragher, B. (2009). Dog picker and tiltpicker: Software tools to facilitate particle selection in single particle electron microscopy. *Journal of structural biology*, *166*(2), 205–213.

Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D. et al. (2019). Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. *Communications biology*, *2*(1), 1–13.

Wang, S., Li, B., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Wang, Z. [Zelun], & Liu, J.-C. (2020). Translating math formula images to latex sequences using deep neural networks with sequence-level training. *International Journal on Document Analysis and Recognition (IJDAR)*, 1–13.

Wang, Z. [Zhou], Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., Schmidt, D., Broaddus, C., Culley, S., Rocha-Martins, M., Segovia-Miranda, F., Norden, C., Henriques, R., Zerial, M., Solimena, M., Rink, J., Tomancak, P., Royer, L., ... Myers, E. W. (2018). Content-aware image restoration: Pushing the limits of fluorescence microscopy. *Nature Methods*. https://doi.org/10.1038/s41592-018-0216-7

Xie, Y., Wang, Z., & Ji, S. (2020). Noise2Same: Optimizing a self-supervised bound for image denoising. *Advances in Neural Information Processing Systems*, *33*, 20320–20330.

Zernike, F. (1942). Phase contrast, a new method for the microscopic observation of transparent objects. *Physica*, *9*(7), 686–698. https://doi.org/https://doi.org/10.1016/S0031-8914(42)80035-X

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155.

Zheng, S., Palovcak, E., Armache, J.-P., Cheng, Y., & Agard, D. (2016). Anisotropic correction of beam-induced motion for improved single-particle electron cryo-microscopy. *bioRxiv*. https://doi.org/10.1101/061960

Zhong, E. D., Bepler, T., Berger, B., & Davis, J. H. (2021). Cryodrgn: Reconstruction of heterogeneous cryo-em structures using neural networks. *Nature Methods*, *18*(2), 176–185.

List of Publications

The following publications contain material presented in this thesis:

1. **Buchholz, T. O.**, Jordan, M., Pigino, G., and Jug, F. (2019, April). "Cryo-care: content-aware image restoration for cryo-transmission electron microscopy data". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (pp. 502-506).*

2. **Buchholz, T. O.**, Krull, A., Shahidi, R., Pigino, G., Jékely, G., and Jug, F. (2019). "Content-aware image restoration for electron microscopy". In: *Methods in cell biology, 152, 277-289.*

3. Krull, A., **Buchholz, T. O.**, and Jug, F. (2019). "Noise2void-learning denoising from single noisy images". *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2129-2137).*

The following publication is currently under review and available as arXiv preprint:

1. **Buchholz, T. O**. and Jug, F. (2021). "Fourier Image Transformer". *arXiv preprint arXiv:2104.02555.*

Selbstständigkeitserklärung

1. Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

2. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten: (keine)

3. Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

4. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und ist auch noch nicht veröffentlicht worden.

5. Ich bestätige, dass ich die geltende Promotionsordnung der Fakultät Informatik der Technischen Universität Dresden anerkenne.

	Dresden, 10. Juni 2021
Unterschrift des Doktoranden	Ort, Datum