Online Super-Resolution For Fibre-Bundle-Based Confocal Laser Endomicroscopy

Agnieszka Barbara Szczotka

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Wellcome / EPSRC Centre for Interventional and Surgical Sciences (WEISS)

Dept. Medical Physics and Biomedical Engineering

University College London

November 15, 2021

Lead Supervisor

Prof. Tom Vercauteren

Co-Supervisors

Dr. Matthew J. Clarkson

Dr. Dzhoshkun Ismail Shakir

Clinical Supervisor

Prof. Stephen P. Pereira

I, Agnieszka Barbra Szczotka, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

”... much more interesting about what we experience

is how we experience it. ”

Adam Łabaza

rest in peace my friend, 22:37 28.05.2021

Abstract

Probe-based Confocal Laser Endomicroscopy (pCLE) produces microscopic images enabling real-time *in vivo* optical biopsy. However, the miniaturisation of the optical hardware, specifically the reliance on an optical fibre bundle as an imaging guide, fundamentally limits image quality by producing artefacts, noise, and relatively low contrast and resolution. The reconstruction approaches in clinical pCLE products do not fully alleviate these problems. Consequently, image quality remains a barrier that curbs the full potential of pCLE.

Enhancing the image quality of pCLE in real-time remains a challenge. The research in this thesis is a response to this need. I have developed dedicated online super-resolution methods that account for the physics of the image acquisition process. These methods have the potential to replace existing reconstruction algorithms without interfering with the fibre design or the hardware of the device.

In this thesis, novel processing pipelines are proposed for enhancing the image quality of pCLE. First, I explored a learning-based super-resolution method that relies on mapping from the low to the high-resolution space. Due to the lack of high-resolution pCLE, I proposed to simulate high-resolution data and use it as a ground truth model that is based on the pCLE acquisition physics. However, pCLE images are reconstructed from irregularly distributed fibre signals, and grid-based Convolutional Neural Networks are not designed to take irregular data as input. To alleviate this problem, I designed a new trainable layer that embeds Nadaraya-Watson regression. Finally, I proposed a novel blind super-resolution approach by deploying unsupervised zero-shot learning accompanied by a down-sampling kernel crafted for pCLE.

I evaluated these new methods in two ways: a robust image quality assessment and a perceptual quality test assessed by clinical experts. The results demonstrate that the proposed super-resolution pipelines are superior to the current reconstruction algorithm in terms of image quality and clinician preference.

Impact Statement

In this thesis, novel processing pipelines for enhancing the image quality of pCLE are proposed. The technical contributions all address the challenges faced in clinical practice and provide solutions that can be immediately applied to clinical hardware.

The technical contributions have a direct impact on the research in the domain of medical imaging and computer vision. It is the first attempt to address the challenge of online SR in endomicroscopy based on deep learning (DL). Particularly, I demonstrated that very common in medical imaging lack of ground truth images can be overcome, and thereby enables DL supervised training. Second, I showed that irregular signals from fibre-bundle imaging devices could be directly used as input data for the Convolutional Neural Network (CNN) without prior reconstructions. Third, I proposed a Zero-Shot DL network trained in an unsupervised manner that allows for building patient, case, and fibre specific models giving the pipeline robustness to external information, which may be considered necessary for medical devices. These contributions were shared during scientific conferences and published in high-impact journals to inform future research. This dissemination may be a starting point for further research not only in image enhancement for fibre-bundle-based imaging devices but any modalities in need of proposed solutions.

I proposed new solutions that improve the image quality of pCLE and have the potential to replace reconstruction algorithms within current clinical devices. This has a direct impact on any industry producing fibre-based imaging devices, specifically our collaborator and sponsor Mauna Kea Technologies to incorporate innovative and effective solutions to bring value to their products and make a positive impact on their clients. The proposed solutions were demonstrated to improve the image quality of the enodmicroscopies produced by their flagship product Cellvizio, and might be translated into clinical devices as the new feature and allow for a new generation of Artificial intelligence (AI)-driven endomicroscopy. It also encourages further investment in AI research and development as it brings an edge to the market offerings.

In addition to the quantitative results that confirm the power of these methods, I demonstrated, through image-quality surveys, that clinical experts prefer DL-based reconstructions over current pCLE images. In particular, experts found the enhanced images easier to interpret and saw potentially valuable diagnostic information in the enhanced structural details. These surveys provide encouraging preliminary data that would support further clinical studies to assess clinical impact.

The potential clinical impact of this research lays in improving diagnostic procedures. Not only does the improved image resolution enhance visible details; but it might also reveal hitherto invisible structures. Thus, this research has the potential to drive endomicroscopy towards outperforming histology, the current diagnostic gold standard, in precision and accuracy. The improvement in image quality would ultimately help pCLE become a widespread tool that increases patient comfort by reducing the time and uncertainty of the diagnostic procedure. In addition, such improvements in image quality would also enable treatment and surgical procedures relying on instant microscopic feedback, thereby reducing costs.

List of Contributions

The contributions described in details in this thesis are largely based on the following peer-revived publications:

Journal Articles

Agnieszka Barbara Szczotka*, Daniele Ravì*, Dzhoshkun Ismail Shakir, Stephen P Pereira, and Tom Vercauteren. “Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction”. In: *International Journal of Computer Assisted Radiology and Surgery* 13.6 (2018), pp. 917–924. DOI: 10.1007/s11548-018-1764-0.

Agnieszka Barbara Szczotka, Dzhoshkun Ismail Shakir, Daniele Ravì, Matthew J Clarkson, Stephen P Pereira, and Tom Vercauteren. “Learning from irregularly sampled data for endomicroscopy super-resolution: a comparative study of sparse and dense approaches”. In: *International Journal of Computer Assisted Radiology and Surgery* 15 (2020), pp. 1167–1175. DOI: 10.1007/s11548-020-02170-7.

Agnieszka Barbara Szczotka, Dzhoshkun Ismail Shakir, Matthew J Clarkson, Stephen P Pereira, and Tom Vercauteren. “Zero-shot super-resolution with a physically-motivated downsampling kernel for endomicroscopy”. In: *IEEE Transactions on Medical Imaging* (2021), pp. 1–1. DOI: 10.1109/TMI.2021.3067512.

* Daniele Ravì and Agnieszka Barbara Szczotka have contributed equally to this work.

Co-authored Publications

Daniele Ravì, **Agnieszka Barbara Szczotka**, Stephen P Pereira, and Tom Vercauteren. "Adversarial training with cycle consistency for unsupervised super-resolution in endomicroscopy". In: *Medical Image Analysis* 53 (2019), pp. 123–131. DOI: `10.1016/j.media.2019.01.011`.

Acknowledgements

This research is funded by Mauna Kea Technologies, Paris, France.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 The rise of optical biopsy

Medical imaging is widely used to provide the clinician with structural, functional and pathological information about the human body. Amongst imaging systems such as magnetic resonance imaging (MRI), computed tomography (CT) and ultrasonography, the lower limit of achievable image resolution remains at the level of tens of micrometres which only allows for observing body systems, organs and tissues. However, this resolution is not detailed enough to observe life on a cellular level, which is often a key to diagnosing conditions such as cancer. To acquire images of the cells, microscopy is commonly used, and microscopy may reach even atomic imaging resolution [46].

Microscopes have been widely used to support diagnosis and treatment for years. They became a core modality for providing cellular level information, especially needed in the early recognition of conditions such as cancer. The early identification of many mucosal diseases has a tremendous effect on the recovery and survival of a patient. To make a cellular-based diagnostic decision, typically a tissue sample is taken from the patient and examined under a microscope in a laboratory. This histopathological *ex vivo* test is called a biopsy. It is the “gold standard” method used in the pathology of dead tissue to determine the presence or degree of disease.

There are several steps in the standard protocol for a biopsy. First, a sample of the tissue is extracted from the area of interest. Next, the sample is transported to the laboratory. After the pathologist receives the sample, they process it, often with the use of chemicals and dyes, to be able to see specific elements of the cell during examination under the microscope. The pathologist analyses the prepared sample by inspecting it visually. The results of the biopsy are sent back to the clinician, who uses the results to support diagnosis, treatment and monitoring of a patient.

The biopsy protocol is time-consuming and will introduce a delay in the flow of crucial diagnostic information as a part of a complex medical procedure. The biopsy result may be available in a few days depending on several factors such as the urgency of the case or local healthcare policy [231]. In some cases, such as during an operation, to avoid additional delays, the pathologist may need to be available to consult on tissue margins using fresh frozen biopsies [73], which often is suboptimal in regards to sample preparation. For example, to examine brain lesion during neurosurgery, a stereotactic brain biopsy is typically performed in the operating room.

However, the biopsy procedure is prone to error during the extraction of a tissue sample. To take a biopsy, a clinician needs to target the region of interest in the examined tissue. This step may also require needle insertion. For instance, in patients with cystic pancreatic lesions, fine-needle aspiration with the support of endoscopic ultrasound is performed. Although the introduction of high-definition and high-magnification endoscopes were shown to reduce the rate of missed diagnoses and allow for targeted biopsies [120], their imaging resolution is on the macroscopic level and provides only images of the examined surface. The precise tissue sampling is still prone to errors [21]. As tissue extraction is often performed blindly in regards to the precise knowledge about the location of the pathology, the extraction may be repeated several times to increase the probability of targeting affected tissue. Thus, it is not uncommon that the biopsy sample does not contain tissue that may be affected by the disease [13]. Another common issue is that the cells in the sample are destroyed or contaminated with another tissue. The sampling error consequently leads to missed diagnosis as a false negative. This negatively affects the specificity and sensitivity of the biopsy. Biopsies may be non-conclusive, and as an effect, the testing has to be repeated, which may cause a significant burden to the patient and disturb treatment delivery.

As biopsy tests may fail to provide conclusive information to the clinician, there is an opportunity to improve the biopsy workflow by developing more real-time biopsy techniques. One solution is to insert a miniaturised microscope inside the patient. Introducing an *optical biopsy* performed as *in vivo* microscopy has emerged as a promising tool to address some of the crucial issues related to the standard biopsy procedure. It may bring a shift in how diseases are detected and managed [86]. The imaging guidance on a cellular level has the potential to alleviate not only sampling errors of the standard biopsy but also provide novel image-based microscopic feedback to the clinician.

An optical biopsy is a non-invasive medical imaging modality based on utilising light-tissue interactions to visualise and assess the epithelium and has been used in both humans and animals. It has been implemented with the use of several techniques such as fluorescence endoscopy, optical coherence tomography, confocal microendoscopy, spectroscopy, molecular imaging, and so forth [30]. Although several implementations of the optical biopsy were proposed, the most commonly used, due to its availability and clinical approval, is fibre-bundle endomicroscopy, also known as probe-based confocal laser endomicroscopy (pCLE). Within this thesis, the focus is on pCLE, and the term endomicroscopy is used when talking about the pCLE-based procedure.

In this Chapter, the reader is presented with background knowledge covering image quality and image enhancement topics, as would be relevant for optical biopsies generated by pCLE. First, the clinical application of endomicroscopy is covered. Next, the physical basis of the design of an endomicroscopy probe is covered in detail to explain the factors affecting the creation and quality of images. Subsequently, image reconstruction algorithms are covered, as they lie at the core of image reconstruction from raw and distorted data. It will become apparent how current hardware designs and reconstruction algorithms can negatively influence the image quality in pCLE. Finally, a summary will highlight the need for further image enhancement algorithms to maximise the efficacy of pCLE.

1.2 Endomicroscopy in the clinic

The optical biopsy permits histopathology-like examination *in situ* (in the place of interest, e.g. a tumour) and *in vivo* (inside a living organism) via instantaneous generation of structural and functional representations of the tissue with microscopic resolution. It is a relatively new technique that allows for cellular-level information provided immediately to the clinician conducting a medical procedure. Thus, it is evolving as an alternative method of performing a biopsy. As many research studies have reported, optical biopsy creates images of tissue that return information similar to that available through histology [141]. The tissue type and the pathology can be recognised by detecting comparable features to those visible under optical microscopy. Classical histology uses stains for better visualisation of tissue producing coloured images (RGB). In optical microscopy, fluorescence substance is used to create images mainly presented by luminescence. In an optical biopsy, a visualised cross-section of tissue is parallel to the surface. While in histology both the transverse and parallel cross-section of tissue can be visualised. The optical biopsy has also a small field of view and reduced resolution, due to the use of a miniaturised optical objective. In contrast, the standard histology provides a large view of tissue samples often without compromising resolution and image quality, as it uses powerful microscopes.

Standard pCLE protocol

This design of the optical biopsy, particularly using probes, makes it easy to implement into existing clinical procedures. The most common application of pCLE is to perform it together with endoscopes [33]. pCLE serves as an augmentation of endoscopy, delivering unprecedented access to on-the-spot microscopic images of cells in the tissue, in addition to the standard imaging view of the surface obtained from the endoscope camera. This new modality relies on the miniaturisation of the microscope objective lens and carrying light via optical fibre bundles. The pCLE probes are designed to be inserted into a working channel or be incorporated into the endoscope [122]. During endoscopy, a clinician positions the tip of the probe in direct contact with the tissue to be tested. The laser light transported by the optical fibres interacts with tissues and then is gathered back. The images are created based on a tissue signal, generated by fluorescent light, which is collected by probe objective. As the pCLE is based on the principle of fluorescence occurring in the scanned tissue, just before the procedure, a patient is administered an intravenous fluorescent agent. The exception is lung examination as the tissues exhibit autofluorescence. The fluorescing substance helps to achieve better contrast of the generated optical biopsies. These reconstructed images are displayed on the screen of the device in real time during the procedure. An example of a commercial endomicroscopy device is shown in Figure 1.1.

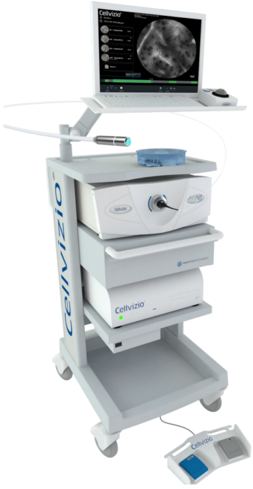


Figure 1.1: Commercially available probe-based confocal laser endomicroscopy Cellvizio (Mauna Kea Technologies, Paris, France), image taken from [232].

pCLE is shown as an up-and-coming way of performing real-time optical biopsy. The successful application of pCLE imaging has been addressed in a variety of studies, and this maturing technology was revived in multiple surveys and books [223, 45, 141, 122]. The imaging modality is valid independently, but it is often used to refine the already commonly used, frequently indiscriminate, endoscopic procedures. pCLE has been shown to be effective in various circumstances, such as in detecting cell abnormalities, cancer and inflammation [141], and pCLE imaging use differs across structures such as the oesophagus, stomach, colon, bile and pancreatic duct and lungs. In the following paragraphs, pCLE classification systems are introduced and examples of the application of pCLE recognised in the clinic and research fields are outlined.

Diagnostic standardisation in pCLE

For the pCLE community, it is essential to follow standard protocols on how to interpret images to ensure systematic clinical outcomes. In 2009, users of endomicroscopy agreed on a standard classification system called the *Miami Classification*. The system standardises the terminology and indications used to diagnose tissue malignancy based on pCLE images [92]. Guidelines have been developed for the endoscopic imaging of the gastrointestinal (GI) tract, which is currently the main application of pCLE [35]. Following the development of the Miami Classification for GI, efforts have been made to improve it as the imagery of other structures may require special treatment. New criteria for benign inflammatory conditions of biliary strictures have been established under the name *Paris Classification* [107]. Many other sets of indications have been suggested as an extension of the *Miami Classification* for findings in, including but not limited to, pancreatobiliary strictures [99], gastric pit patterns and vessel architectures [150] and duodenal epithelial tumours [152].

Example applications of pCLE to specific anatomical structures

Cancer Early-stage cancer is often indicated by different types of epithelial cancer cells. These abnormal cells are distinguishable with the use of endomicroscopy, and pCLE generates images facilitating detailed analysis of cellular and sub-cellular structures [141, 112]. Although pCLE is not used as a routine examination yet, it was confirmed to be a specific and accurate clinical tool in early cancer detection. The direct and immediate investigation of potential precancerous tissue gives pCLE technology a unique advantage with respect to standard biopsy. It might offer improved treatment in a very early stage of cancer, which is crucial to the success of therapy and the comfort of the patient. It also has the potential for detecting specific cancer markers and screen tissue, otherwise not examined in standard biopsies [114].

Barrett's oesophagus The main application for pCLE is the diagnosis, treatment and monitoring of patient suffering from Barrett's oesophagus (BE) [141]. The precise and accurate diagnosis of BE is possible because pCLE allows the detection of dysplasia and neoplasia in real-time improving results yield with high-definition white-light endoscopy [90]. Additionally, pCLE can be used to guide surgical intervention for Barrett's oesophagus. After treatment, it plays the role of a monitoring tool during follow-up endoscopy. It allows checking the effects of the treatment and examines for malignant tissue. Dubnart et al. have shown that when pCLE is used in surveillance of BE, the need for standard biopsies can be greatly reduced [56]. Overall, *in vivo* biopsies reach a high diagnostic accuracy of around 97% [38]. Moreover, patients who have a high risk of developing BE are ones with gastroesophageal reflux disease (GERD), a condition caused by chronic reflux. Applying pCLE for these patients showed the diagnostic applicability of pCLE to GERD [106], enabling very early indication in diagnosing BE.

Polyps Conventional colonoscopy is a standard way of detecting polyps, which are recognised as the most important risk factor in colon cancer. pCLE gives intraprocedural histology information, which is not available by colonoscopic examination alone and can be used in combination with standard or enhanced endoscopy to identify target tissue and aid detection of colon cancer [117]. pCLE has been shown to outperform it in the differentiation of colorectal polyps [111]. Without waiting for the results of histology, diagnosis can be made during the single endoscopic procedure based on images provided by pCLE. If needed, they can perform any surgical procedure related to colorectal lesions without delay, avoiding unnecessary re-intervention [104]. Polyp classification and assessment after polypectomy is becoming more efficient and less expensive [66].

Inflammatory bowel disease Neumann et al. found a high agreement of pCLE with histology findings related to inflammatory bowel disease (IBD) [102]. They demonstrated that IBD could be diagnosed by pCLE, specifically when pCLE is used for characterisation of IBD associated lesions [155]. A benefit of pCLE was in reducing the time of taking surveillance biopsies in IBD. Moreover, the optical biopsy reveals specific features of mucosal inflammation, which makes pCLE a good predictor of clinical outcomes associated with IBD including Crohn's disease [172]. Kiesslich et al. have validated the utility of pCLE in the detection of the cell shedding and barrier loss and shown that it can serve as an indication to recognise relapse of IBD [96]. pCLE have been tested in the retrospective analysis and a prospective study for identification of intramucosal bacteria, which are more common and more spread in patients with IBD, showing promising results [88].

Biliary tract The low sensitivity of current methods applied to detect malignant tissue in a biliary duct was a motivation for trying pCLE as a better way to support the observation of indeterminate biliary strictures. It proved to have a modest impact on the cancer diagnosis [87]. A pCLE probe can visualise biliary strictures passed through a colonoscope or catheter, providing real-time microscopic images of the biliary epithelium [122]. This new information advances current imaging techniques used in detecting cancer within biliary strictures. They proposed four descriptive criteria to differentiate benign inflammation in biliary strictures in the *Paris classification* [107].

Pancreatic cysts Endoscopic ultrasound (EUS) is used to examine structures accessible from the GI tract, such as the pancreas, allowing for visualisation of the macroscopic characteristics that can be extended with the optical biopsy. The combination of endomicroscopy and EUS, called EUS-guided needle-based confocal laser endomicroscopy (nCLE), is possible using the nCLE mini-probe passed through an endoscopic needle during EUS [129]. This newly combined imaging modality has been adapted as a tool for *in vivo* real-time microscopy of a pancreatic mass. The optical biopsy has been shown to improve differentiation of various types of pancreatic lesions [136]. The clinician can also assess any changes that might have occurred in the pancreas by evaluating cellular images produced by nCLE in a safe manner [203].

Lungs An improved endoscopic procedure with pCLE also appears to be promising in the diagnosis of many lung diseases [77]. Bronchoscopy was improved by inserting nCLE inside lung nodules and lymph nodes, providing access to feedback information for diagnostic, staging and treatment procedures in lung cancer [217]. It has been demonstrated that the visualisation of malignant cellular structures such as dark enlarged pleomorphic cells, dark clumps, and directional streaming is feasible and safe [211]. The introduction of nCLE has made it possible to diagnose malignancy with high accuracy to 95%. There is a correlation between results obtained with standard histology [137] and chest CT [207] and pCLE.

Urology pCLE has been shown to have the potential to enhance conventional cystoscopy and ureteroscopy thanks to its capability of visualising differences among normal urothelium, low-grade tumours and high-grade tumours [61]. Wu et al. has proposed diagnostic criteria allowing diagnosis of urinary tract pathology, mainly bladder cancer visualised with pCLE in conjunction with white light cystoscopy (WLC) [93]. Another study has shown that pCLE outperform WCL for cancer diagnosis if the practitioner has expertise in working with endomicroscopy and learning of the non-specialist requires only short training [109]. At present, the urothelial carcinoma of the upper tract (UTUC) lacks a specific diagnostic technique. It was demonstrated that endomicroscopy produced results that corresponded with standard histopathological results [178]. We may improve conservative disease management thanks to safe real-time histology of UTUC lesions.

Initial studies There are also numerous primary studies showing the first applications of pCLE in surveillance, treatment and post-treatment monitoring. For example, initial research has found that pCLE is sensitive in detecting specific lymphocytes for diagnostic of coeliac disease [67]. Preliminary studies also suggest the application of pCLE in a description of gastritis related to Helicobacter pylori [79], but prohibitive cost compared with a simple Clo-test or a faecal Helicobacter antigen test. pCLE showed a higher specificity than other techniques in research aimed at finding real-time histology for head and neck cancer, e.g. diagnostic of Lugol's voiding lesions [134]. Tanbakuchi et al. reported that pCLE could be used safely for real-time histology of ovarian epithelium [78]. There are also pCLE applications in various research laboratories, from pre-clinical studies to immunochemistry [47], and animal studies [91].

Advantages and disadvantages of pCLE

The optical biopsy is superior to the standard biopsy in several domains [63]. It is a non-invasive technique that provides real-time details to allow the target to be determined immediately during endoscopic imaging procedures, making it less time-consuming than conventional biopsy procedures. It increases diagnostic confidence [135] by providing an alternate way for clinicians to make more educated and accurate medical decisions based on a real-time visual image of the living mucosa in their native environment. Moreover, a biopsy is conducted on dead tissue without an option for dynamic functional information; with optical biopsy, observing living tissue has become possible [48]. As an imaging guide, it allows for localisation of the affected tissue of interest during a medical procedure such as surgery. Thanks to the ability to manoeuvre precisely in the human body, targeted care and tissue removal can become more robust when imaging input is available, e.g. more successful removal of cancer tissue around the nerves during prostate surgery [151]. Additionally, optical biopsy accompanying tissue sampling decreases the likelihood of mistakes and enhancing biopsy susceptibility. It can prevent hazards such as unnecessary treatments and traditional biopsy costs [89]. Using pCLE reduces the number of biopsies needed while increasing diagnostic yield [121]. As shown by the multiple initial studies, pCLE also opens up new opportunities to develop or build new diagnostic approaches where microscopic details are used to make more informed decisions. This gives the potential for experimental bio-markers and fluorescent image-tagged molecular agents to be used [63].

Although optical biopsy provides new or improved possibilities for discrimination of pathological tissue, it suffers from several limitations. The technology remains safe, yet some complication such as allergic reaction to fluorescent agent or bleeding related to needle insertion, which is used to conduct pCLE imaging, may occur rarely. pCLE is a small-field imaging device that is therefore only suitable to be used at an area already detected by normal or optically improved endoscopes [63]. The assessment of the images produced by an optical biopsy requires specialist knowledge, and often images can be challenging interpreting. The clinician has to be trained to understand the image content, or consultation with a histologist is needed [89], and the learning curve for pCLE experts varies between applications. Endomicroscopy is not a widespread tool. Compared to other established image modalities, the popularity of optical biopsy is still low, making the community of clinicians and researchers using the technique small as well. The growing clinical research using optical biopsy has helped develop novel treatments, yet its clinical potential is not completely exploited. Even though pCLE allows redesigning some endoscopic procedures to be efficient and sometimes less expensive, pCLE may have a high cost for many applications. The optical probes are costly and require special treatment, such as specific disinfection. These factors make the procedure less accessible to the patient.

Beyond the limitations related to integrating the technique into clinical workflow, there are drawbacks created by the technical implementation of this modality as this modality relies on the miniaturisation of the microscope objective and transporting light via optical fibres. The physics of the acquisition through fibres constitute a severe limitation to image quality. As a consequence, the image quality is decreased by comparison to typical bench-top microscopes used to conduct a biopsy. With the disadvantages mentioned above, the limited image quality often interferes with its potential status as the “gold standard” test.

1.3 The physical and technical basis of pCLE

There have been many research works on the specific design of technology implemented into endomicroscopy, and a good review on technologies used in endomicroscopy is provided by [223, 95]. Commercial solutions of endomicroscopy, available and approved for clinical use, were developed based on different light technologies such as optical coherence tomography used in the NinePoint Medical (Bedford, Massachusetts, USA) platform, white light microscopy implemented into products of Olympus Medical Systems Co (Tokyo, Japan), and confocal laser microscopy put into effect in devices developed by Pentax Medical (Tokyo Japan), Optiscan (Melbourne, Australia), Carl Zeiss Meditec (Jena, Germany), Mauna Kea Technologies (Paris, France).

The primary focus of the research in this thesis is around pCLE. Some of the pioneering work on the development of pCLE was undertaken in the early 90s by [7, 12], and the continuous advances in the field brought a real-time confocal scanning method to biomedical applications [29, 19, 28]. The core of the technology is a single fibre bundle comprising tens of thousands of optical fibres, each acting as the equivalent of a single-pixel detector. This bundle is used to transport bidirectional light collected by scanning the proximal end of the fibre. This particular technology is currently implemented into only one commercially available endomicroscope called Cellvizio developed by Mauna Kea Technologies.

In this thesis, research is centred around the Cellvizio system. Cellvizio has a good market position thanks to its wide distribution and uses in clinical practice and research, relative to other systems. Although the research presented in this thesis may apply to any fibre-bundle-based imaging system, it is mainly validated on data acquired by the Cellvizio, as no other clinical data were available from the other pCLE system. Thanks to the close collaboration with the Cellvizio manufacturer, they made it possible to access internal specifications related to the device implementation and design of fibres crucial in developing this research. The following paragraphs dive into the technical details of pCLE implementation based on Cellvizio and describe all the system elements and their role in the image acquisition process.

The implementation of pCLE hardware is based on the physical principle of classical confocal microscopy. Confocal microscopy is based on the optical sectioning of fluorescence light coming from the focal plane of the lens emitted by a sample [17, 4], which is prepared by slicing and dying with a specific fluorophore. A diagram of the trajectory of the light in a confocal microscope is presented in Figure 1.2. A light beam (red arrows) from the light source is focused on the sample by a microscope objective. When the light beam illuminates the sample, the dye particles in the sample are excited, and fluorescence occurs. The fluorescence light (blue arrows and grey dots) is emitted back from the sample in the objective lens direction. A beam splitter (mirror) is designed to reflect only the light produced by fluorescence in one focal plane. The aperture rejects the light from the out-of-focus

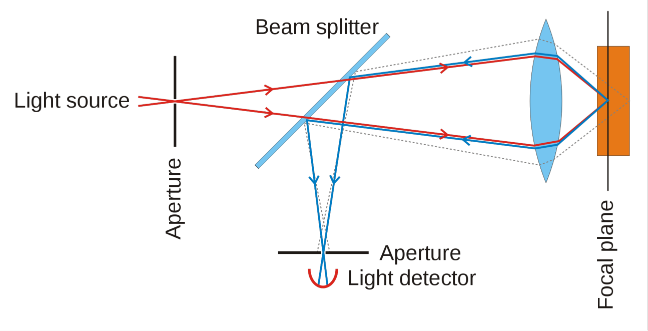


Figure 1.2: The concept of a confocal microscope configuration, taken from [233].

plane (grey dots), only light rays coming from the in-focus plane (blue arrows) are selected by the aperture, also known as pinhole. Finally, a light detector detects the sectioned beam light; the beam is scanned using oscillating mirrors over the regular pattern across the sample to create images. The primary attributes of confocal microscopy can block light from out-of-focus planes, providing high-resolution image reconstruction. It also allows imaging a thick sample since a confocal microscope can gather light coming just from the in-focus plane, which may be set deeper in the sample than it was possible with traditional light microscopes.

A schematic design of a pCLE system is shown in Figure 1.3. This pCLE system is built with a distal optical bundle composed of optical fibres, a laser scanning unit (LSU) that encapsulates a device's optics, a laser source and an image processing unit.

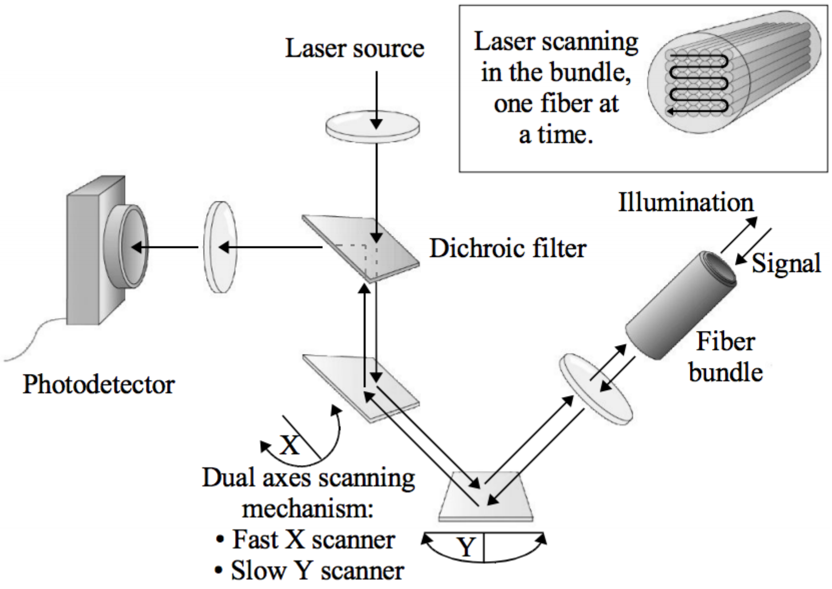


Figure 1.3: pCLE system, image taken from [28]

The light is generated by a laser diode which typically emits coherent blue light of a wavelength of 488 nm. This photon energy is compatible with the fluorescent excitation energy of a specific fluorophore (e.g. Fluorescein), enabling high fluorescent contrast for imaging biological tissue. The light travel through the systems is similar to confocal microscopy. In particular, the laser light is directed from the source in the LSU through the optical bundle to the tissue that is in touch with the optical head of the bundle. The photons interact with a fluorophore in the tissue, and fluorescence occurs. The fluorescence light radiated from the tissue is transported back through the optical bundle from the optical head to the LSU.

The optical bundle is used to transport light in both directions: from the proximal laser unit to the distal end of the fibre and from the tissue to the LSU. It is manufactured as a long and flexible probe that can be easily inserted into an endoscope's working channel. An optical probe has the optical head used as a microscope objective on one side and the connector used to attach and position the bundle into the LSU on another side. The probe's adaptation enables the movement of an objective of the microscope inside the human body and makes distant imaging feasible. The optical bundle allows for 2-dimensional imaging of structures in touch with and parallel to the tip of the bundle head. There are several probes distributed for the Cellvizio system characterised by different confocal depths, length, resolution, a field of view (FoV) and application [1]. The bundle is built with tens of thousands of optical fibres. Every single fibre in the bundle acts as a single-pixel detector. The number of optical fibres assembled into a bundle defines its diameter and consequently FoV. Amongst all offered probes, two main types of bundles can be distinguished: a bigger one with a diameter compatible with the working channel of a scope (pCLE), and a smaller one with a diameter allowing insertion inside a needle (nCLE). The FoV depends on their diameter and ranges from $240 \times 240\mu m$ to $600 \times 600\mu m$. Optical fibres have a specific imaging depth set on $0 - 70\mu m$. The user cannot tune the imaging depth, as it is a characteristic of the probe. The length of the probe is defined mainly by its application. The probes are compatible for sterilisation or disinfection, which for safety is allowed to be exercised about 10-20 times before disposal because of specific properties of the head construction and fibres.

The LSU key role is to process the light transported back by the probe, using optics. The scanning unit point-by-point scans the light beam coming out of the bundle's proximal part to produce an image. The scanning system acts as a pinhole,

[1] http://www.cellvizio.net/learn/technical-training/5-how-to-use-the-miniprobes

which allows for receiving light only from one fibre in time. The scanning unit comprises two mirrors: first (*Y*) oscillating horizontally with a frequency of 4 kHz, and second (*X*) 9-18 Hz galvanic mirror scanning frame. Customised hardware synchronises mirrors that reflect light to the filter. The dichroic filter selects only fluorescence light, and the pinhole rejects light coming out of the focal plane of the optical head of fibre. The scanning system collects from 9 to 18 frames per second. The filtered light is transformed into a voltage signal by the pixel detector and finally digitised. The digitised signal is translated into raw data, which is later reconstructed by dedicated algorithms implemented in the image processing unit. The image processing unit is a computer with dedicated software responsible for calibration, reconstruction, and restoration of the final pCLE video sequences.

1.4 Reconstruction of pCLE images

During the scanning procedure, the surface of a bundle's profile is acquired. This profile represents collections of signals coming from fibres. Each fibre provides a single source of information. The image acquisition performs signal oversampling, creating a bigger number of pixels than an actual number of fibres assembled into a bundle to allow to distinguish individual fibres [53]. An example image of the bundle surface in shown in Fig 1.4, and it is referred to as the raw image.

The acquisition design has a negative effect on image creation. The raw image is distorted with several artefacts such as a honeycomb pattern caused by irregularly assembled fibres into a bundle, varying pixel intensities due to variability amongst fibres' properties, and a skewed image shape because of a geometric distortion during scanning. The raw image is not suitable for the end-users direct use as it is adversely affected by these artefacts making images hard to interpret.

The honeycomb pattern effectively shows the fibres' positions within the profile of a bundle. Each fibre is visible as a small dot in the image surrounded by interconnected boundaries seen as a darker background in the pixelated image. The fact that the bundle is constructed with thousands of individual fibres causes the effect. They have variable size and shape and are irregularly distributed across the

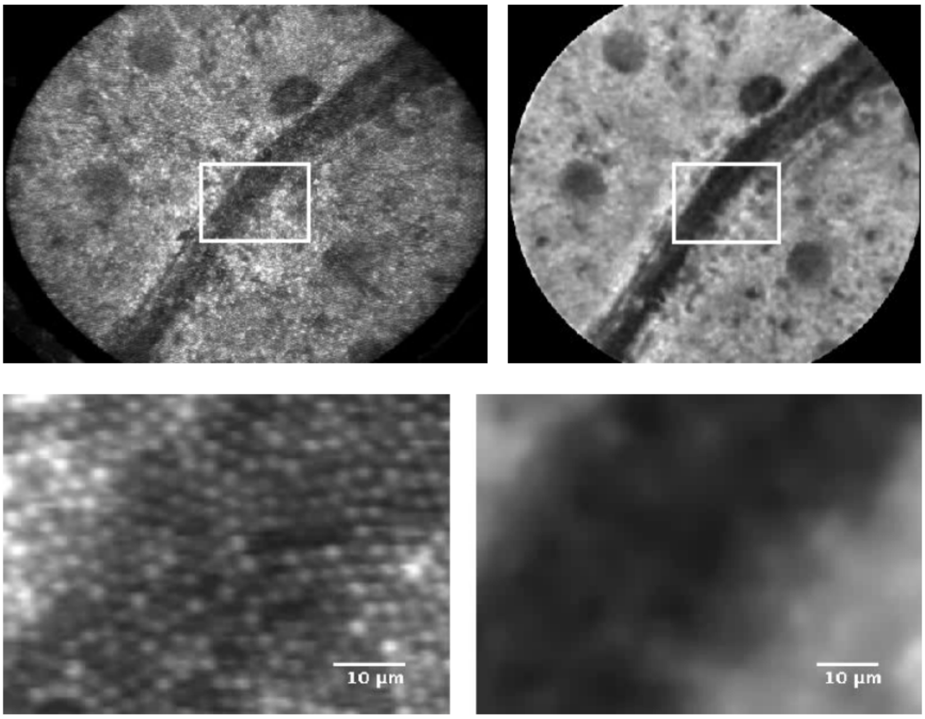


Figure 1.4: "Autofluorescence FCM images of a Ficus Benjamina leaf. Left: Raw data. Right: Reconstructed images. Top: Complete images. Bottom: Zoom on rectangle". The image and the description are taken from [53].

bundle's FoV. Although fibres are packed very tightly, there is a minimal space between them that does not transport light. The ensemble of fibres, specifically their cladding, produces dark surroundings around each fibre virtually visible as the distinctive semi-regular honeycomb pattern.

Fibre intensities in each cell of the honeycomb pattern represent individual and non-calibrated fibre signals. Each fibre has individual light transition properties, including coupling efficiency and inter-core coupling spread, and has a spatio-temporally variable auto-fluorescent response. As a result, each fibre's intensity in the raw image represents tissue signal indirectly, and it is depended on the specific transfer function of each fibre.

Geometric distortions are related to different scanning speeds of the mirrors in the LSU. The laser beam moves on a sinusoidal scan trajectory with varying speeds depending on its position within a bundle allowing for signal oversampling. This causes scanning distortion, affecting the raw image relative fibre position within the bundle's FoV. As a result, the raw image is elongated in the shape of an ellipse instead of having a circular shape of a bundle section.

There have been multiple reconstruction algorithms proposed to alleviate the mentioned artefacts. Perperidis et al. published a detailed review presenting developments in reconstruction algorithms dedicated to images acquired with fibre-bundle endomicrocopes [223]. Shinde and Matham [128] surveyed algorithms to remove honeycomb patterns specifically in application to endomicroscopy. The research presented in this thesis is based on data produced by Cellvizio devices. The algorithm implemented in those devices is a "gold standard" method used in the clinic for reconstructing pCLE images from the raw signal. The following paragraphs give the reader an understanding of that algorithm and present its limitations regarding the image quality of pCLE reconstructions.

The pCLE reconstruction algorithm is based on the research proposed in [67, 103, 53]. It consists of three steps: fibres' position and the signal is estimated, signals are calibrated, and finally, signals are interpolated onto the Cartesian grid to reconstruct the pCLE image.

Estimation In the first step, the position of fibres in the bundle's FoV is estimated. The fibre positions are detected by threshold-based segmentation explicitly designed for the bundle [53]. The segmentation allows extracting image spaces in each honeycomb cell and builds a map with fibres. The segmented fibre positions are distorted due to the bundle's scanning specifics, which uses a sinusoidal movement of the laser beam driven by mirrors. Therefore, the mapping has to compensate for that distortion. It corrects the position of the fibres in the raw image space to their actual position within an optical bundle based on the acquisition parameters steering mirrors.

Next, the segmented image is used to estimate the fibres' signal. Each segmented region represents a fibre position in an oversampled raw image. Typically, each segment spans over 15-50 raw pixels, but each fibre provides only one sampling point on the tissue. All pixel values in each segment are averaged to obtain an estimated individual signal of the fibre.

Calibration During the second step, calibration is performed. Every fibre acts as a mono-detector characterised by its parameters, resulting in a characteristic fibre response. This response is dependent on the transfer function of the fibre that needs to be modelled independently of other fibres. The fibre photometric model proposed in [67] represents the effect of the transfer function of a fibre as proportional to the concentration of fluorophore seen by a fibre as a function of time. We can interpret this model as a linear combination of a fibre's constant gain and offset. As a result of the advancements, that assumption was rejected by Savoire et al. who proposed the online blind calibration to estimate the fibres' gain and offset as slowly time-varying variables form several frames for pCLE video [103]. Once calibration coefficient parameters are estimated, fibre signals can be normalised. After calibration, there is no more difference in response between fibres.

Interpolation In the third step, the image's reconstruction from calibrated fibre signals to a Cartesian image is performed. The pCLE signals produced by each fibre are non-uniformly distributed over the bundle FoV in a characteristic irregular and a quasi-hexagonal geometrical pattern, where each fibre produces a single-pixel signal. Since the fibres in the bundle's FoV are distributed irregularly, they have to be cast onto the Cartesian grid by their corrected position, and the gaps between fibres have to be filled. To do so, the signal from a fibre is interpolated based on its position within the pattern. The interpolation is performed based on a Delaunay triangulation and linear interpolation constructed from the fibre pattern in such a way that signals from three neighbouring fibres are used to obtain seven pixels in the Cartesian image.

The estimation of the bundle properties is a part of the device calibration process, which runs when the platform is initialised, and the probe is inserted into the device. Once the probe is calibrated, the reconstruction includes signal calibration and interpolation only, and that allows producing live pCLE video stream in real time. The reconstructed image is shown in Figure 1.4. The "gold standard" reconstruction removes the honeycomb pattern effectively, compensates for the difference in signals transmitted by individual fibres and handles geometric distortion well. The Delaunay triangulation-based interpolation allows performing interpolation to deliver pleasant-looking pCLE reconstructions.

A bundle built with 25k fibres captures 25k unique scattered fibre signals. Those signals are used to reconstruct an image on a Cartesian grid. The interpolation does not recover high-frequency information from raw data but populates the grid. Consequently, the oversampled pCLE reconstruction with around 175k pixels (excluding any black borders) carries only 1/7 of informative pixels. This reconstruction produces images with minimal information recovery from the raw data. Additionally, reconstruction introduces edge artefacts caused by linear interpolation across the edges of the underpinning Delaunay triangulation [53].

While this resampling allows compensating for artefacts such as the honeycomb pattern produced by non-Cartesian fibre arrangements, it does not have any denoising properties. The pixel signals contain both tissue signal and noise. Current pCLE reconstruction algorithms interpolate noisy pCLE signals onto an oversampled Cartesian grid without compensation for noise. There are two types of pCLE noise: temporally constant and dynamic. The temporally constant noise comes from fibre non-linearity and calibration errors and is partly removed by the reconstruction algorithm. The temporally dynamic noise comes from the random photo-detection. The distinction of noisy fibres is not trivial, and the current reconstruction algorithm treats all signals equally. The effect of noise on the image quality is amplified because noise is reconstructed by Delaunay interpolation. It means that 21 pixels are affected due to interpolation with neighbours by a single noisy fibre signal which is visible as the "blob-like" structure spanning over several pixels at the image, especially evident on the triangulation edges.

1.5 The need for quality improvement of pCLE

Clinical motivation

The clinical applications of pCLE, summarised briefly in Section Section 1.2 "Endomicroscopy in the clinic", adds to the growing body of evidence that pCLE can complement or replace classical biopsy, which is established as the "gold standard" method for tissue testing. Besides the demonstrated impact of introducing pCLE into the clinic as a new imaging modality, it is apparent that information yield from pCLE-acquired images is a subject of discussion among specialist using pCLE in their practice. Amongst them, the authors of [141] reviewed pCLE as an imaging modality in application to GI and measured average specificity and sensitivity of pCLE. They have shown that it still does not provide diagnostic specificity and sensitivity sufficient to become a standard tool in some applications. More explicitly, Krishna et al. [218] has stated that it is essential to monitor the image quality of endomicroscopy; as such, it is critical to interpreting images correctly. As a result, lower accuracy in endomicroscopy-based diagnosis is linked to its limited image quality that requires special care such as the input of a clinical advisor and analysis of multiple images to increase the likelihood of capturing more details [216]. In the study establishing the Paris Classification, the authors have shown a 21% decrease in accuracy of interpretation for images with poor quality versus images of at least good quality. Chang et al. suggested strategies for collecting pCLE images of the urinary tract and addressed their limitation in image quality, indicating the need for the development of new generation probes capable of improving the quality of pCLE images [110]. To sum up, the image quality of pCLE is a direct limiting factor influencing the reliability of this imaging technique, and there is a strong message coming from the clinical environment that it is desirable to improve pCLE image quality.

The clinical need for an upgrade of pCLE images provides the opportunity to research a new and relatively unexplored area of improving the quality and the resolution of endomicroscopy. As specificity and sensitivity depend on the quality of the pCLE images, enhanced quality of images could improve pCLE images' interpretability. The increase in image resolution will enhance and increase the number of visible details in images. Potentially it might reveal hitherto invisible structures allowing for more readable pCLE imaging-guided feedback.

The clinical impact of this research might be seen in improving the diagnostic procedure. The enhancement of the pCLE might bring a more reliable source of information, improve the pCLE diagnosis, and further benefit patients. As such, this research has the potential to drive real-time endomicroscopy towards outperforming histology - the current diagnostic gold standard - in precision and accuracy. Besides decreasing the cost of the current diagnostic procedure, this will also reduce the time of the diagnostic procedure and increase patient comfort.

Technical constraints

As apparent in the current move to high-definition endoscopic detectors, the general trend for image sensor manufacturers is to increase the resolution. Recently introduced 4K endoscopes provide 8M pixels, a difference to pCLE of 2-to-3 orders of magnitude. The high-definition probes for pCLE will not be available soon due to size constraints. We argue that there is an unmet need to improve the resolution of endomicroscopic images without interfering with the hardware design.

pCLE provides immediate real-time information and allows clinicians to make more informed and efficient diagnostic decisions. Based on pCLE images seen during the endoscopic procedure, further surgical procedures can be performed without delay, avoiding unnecessary re-intervention. Additionally, thanks to the instant image-based feedback loop, clinicians can target treatment delivery more precisely than is possible using the standard procedures lacking that information. As pCLE is allowing for imaging instantaneously, any proposed enhancement of its image quality has to deliver seamless online augmentation. Ideally, pCLE enhancement has to be available in real-time to avoid offline analysis, which would slow down and complicate the process and would not fit the standard clinical workflow required in this context.

Building on the idea that online pCLE image enhancement is desirable, there is a scope for the development of a software augmented solution enriching pCLE images. In response to that need, this thesis proposes algorithmic resolution augmentation methods, which are more agile and straightforward to implement than hardware engineering solutions and do not interfere with exiting pCLE workflow. The focus is on delivering an approach that has the potential to replace currently used “gold standard” reconstruction. In Section 1.4 “Reconstruction of pCLE images”, the limitation of the reconstruction was shown. Alleviating problems arising from the reconstruction algorithm, such as noise interpolation, image oversampling, and edge artefacts, is the basis for advancing the pCLE reconstruction.

1.6 Research objective, questions and hypotheses

The limited image quality of pCLE is the research problem tackled in this dissertation. The primary research objective of this PhD project is to improve the image quality of endomicroscopy with an online super-resolution algorithm. The solution that might use prior information available for real-time image enhancement is considered. The main constrain for solving the limited quality problem is the lack of high-definition ground truth pCLE, as they are not available for pCLE.

This thesis answers research questions as follows:

- What is prior information in the context of endomicroscopy?
- Is the use of prior information possible, and does it benefit the image quality of endomicroscopy?
- Is prior information available for real-time processing?
- How can the prior information be exploited to improve the image quality of endomicroscopy?

The main research hypothesis are:

- The current probe-based confocal laser endomicroscopy reconstruction is suboptimal and does not use any prior information. Using prior information during the reconstruction process is possible and allows for a better representation of reconstructed endomicroscopies.

The specific research hypotheses are:

1. Prior information is available and allows for real-time processing of pCLE.
2. A temporal redundancy in the pCLE video sequence allows for the fusion of the temporally aliased high-frequencies to reconstruct detailed frames.
3. Accounting for the fibre pattern of the pCLE bundle in the reconstruction algorithm benefits the quality of the reconstructed endomicroscopies.
4. A physical model of pCLE inserted into the reconstruction algorithm drives quality improvement of the reconstructed endomicroscopies.

1.7 Thesis organisation and contributions

The thesis is organised as follows:

In Chapter 1 “Introduction”, I give a detailed description of endomicroscopy as a medical imaging modality allowing for performing an optical biopsy by providing details on its clinical applications. Next, the technical principles and design of pCLE are described with further information on a reconstruction algorithm in the context of image quality. Finally, the need for image quality improvement in pCLE is justified, and the main research hypothesis is stated.

In Chapter 2 “Background on super-resolution”, the reader can find a critical review of the methodology essential to the research presented in this thesis. First, the background on super-resolution (SR) and image quality assessment is given. Then, I present developments in SR, including older 1st generation techniques and current state-of-the-art methods exploiting deep learning to solve the SR problem. I also outline existing attempts to improve the image quality of pCLE. This Chapter gives background knowledge to position the contributions presented in this thesis in the context of the image SR used to improve pCLE image quality.

In Chapter 3 “Single-image super-resolution for endomicroscopy”, Chapter 4 “A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks” and Chapter 5 “Zero-shot super-resolution for endomicroscopy”, I present three contributions developed to meet the objective of this thesis by answering the research questions.

Chapter 6 “Conclusions, limitations and future research lines” is a summary of my contributions. I provide a critical discussion of the research results presented in this thesis, focusing on its novelty and limitation. Finally, I discuss potential new avenues opened thanks to the conducted research.

Contributions

My first contribution, presented in Chapter 3 “Single-image super-resolution for endomicroscopy”, is on leveraging an exemplar-based super-resolution (EBSR) to address the challenge of pCLE image enhancement. EBSR allows for improving image quality by profiting from a dataset of aligned pairs of low-resolution (LR) images and high-resolution (HR) images; yet, those HR data are not available in pCLE. To tackle the lack of HR images, I propose a novel synthetic data generation approach that unlocks EBSR for pCLE.

At the core, I develop a novel simulation that approximates the HR images. The HR images are estimated by exploiting the temporal information contained in a sequence of LR images with a video registration technique. The LR images are generated with the forward model of the fibre bundle. Pairs of estimated HR images and realistic synthetic LR are used to train EBSR models.

I analysed the performance of three different state-of-the-art deep learning (DL) models that were trained with data generated by the proposed simulation. The super-resolved reconstructions obtained from a dataset of pCLE video sequences were validated through an extensive image quality assessment. The validation considers several quality scores, including a Mean Opinion Score (MOS) survey conducted on a group of pCLE experts. The findings suggest that the proposed novel simulation unlocks the training of EBSR and produces models capable of recovering pCLE images with enhanced quality. In this chapter, I demonstrate that obtained SR of pCLE produces a compelling and convincing boost in the quality by incorporating prior information on pCLE physic acquisition and without the reliance on real ground truth data.

My second contribution on addressing pCLE enhancement, presented in Chapter 4 “A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks”, is focused on how to adjust convolutional neural networks (CNNs) to specificities of the pCLE image acquisition process. In pCLE, the fibre bundles produce irregular fibre signals that need to be reconstructed to Cartesian images. In Chapter 3 “Single-image super-resolution for endomicroscopy”, I show that CNNs improve pCLE image quality. Yet, this exploits already reconstructed pCLE images, as classical CNNs may be suboptimal regarding irregular data.

In Chapter 4 “A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks”, I take under investigation how to handle irregularly sampled pCLE signals with CNNs directly, with no prior reconstruction. The main target of this chapter is an ablation study that aims to compare the sparse and dense approaches in pCLE image reconstruction and the CNN-based SR methods. For a comparison study, I propose a novel sparse approach allowing the input of fibre signals into CNN. That allows looking into taking irregularly sampled or reconstructed pCLE images as the input of CNNs.

As the main methodology, I design a novel CNN layer allowing to embed of a trainable generalised Nadaraya–Watson (NW) kernel regression into the CNN framework. Using the novel layer, I construct DL-based architectures allowing for reconstructing high-quality pCLE images directly from the irregularly sampled input data.

To test the proposed method, I created synthetic sparse pCLE images allowing for reference-based assessment. I validated the results through an image quality assessment based on a combination of the following metrics: peak signal-to-noise ratio and the structural similarity index. The conducted analysis shows that both dense and sparse CNNs outperform the reconstruction method currently used in the clinic. I also show the applicability of models to the original pCLE domain. Therefore, the proposed solution enables a principled adaptation of CNNs for processing sparse data in application to SR.

The third contribution, presented in Chapter 5 “Zero-shot super-resolution for endomicroscopy”, brings together both the power of CNNs-based SR methods and accounting for a physical pCLE model. I have shown that SR methods can be successfully employed to improve the quality of endomicroscopy imaging. Yet, the inherent limitation of research on SR in endomicroscopy remains the lack of ground truth HR images, commonly used for both supervised training and reference-based image quality assessment (IQA). Therefore, I explore alternative methods, such as unsupervised SR.

The main methodology is a design of a novel zero-shot super-resolution (ZSSR) approach that relies only on the endomicroscopy data to be processed in a self-supervised manner without the need for ground-truth HR images. I tailored the proposed pipeline to the idiosyncrasies of endomicroscopy by introducing both: a physically motivated Voronoi downscaling kernel accounting for the endomicroscope's irregular fibre-based sampling pattern and realistic noise patterns. I also took advantage of video sequences to exploit a sequence of images for self-supervised zero-shot image quality improvement.

I run ablation studies to assess my contribution in regards to the downscaling kernel and noise simulation. I validate the method proposed in this Chapter on both synthetic and original data. Synthetic experiments were assessed with reference-based IQA, while the results for original images were evaluated in a user study conducted with both expert and non-expert observers. The results demonstrated superior performance in image quality of ZSSR reconstructions compared to the baseline method. The ZSSR is also competitive compared to supervised single-image SR, especially being the preferred reconstruction technique by experts.

Chapter 2

Background on super-resolution

Contents

In Chapter 1 “Introduction”, the reader was presented with an in-depth introduction to endomicroscopy and its limitations in image quality that come from the image formation process. The current interpolation-based reconstruction method is recognised as suboptimal. The reconstruction algorithm distributes fibres on a Cartesian grid, creating a very sparse image with non-informative spaces between fibre cores. In pCLE, irregularly spaced signals on the image grid impose the need for interpolating the missing signals. It is known that interpolation techniques, such as bicubic upsampling or Delaunay triangulation used in pCLE, do not increase information in the image, but repeat pixel values in the image space [74] itself introducing artefacts.

Techniques capable of recovering missing high-frequency information at the image are called super-resolution (SR) [23, 164, 14]. Not only does SR help to recover degraded image information from a LR signal, but most importantly increases the number of informative pixels in the image and essentially enlarges the image with improved quality and content compared to its LR source. In pCLE, the enlargement is related to increasing the informative pixel rate at the image, and not to directly upsizing an already oversampled images. The developments presented in this thesis aim to replace pixels interpolated into the inter-fibre spaces with a more informed estimation of the lost high-frequency signal. Therefore, SR can be considered under the scope of methods capable of effectively improving pCLE image quality by serving as a novel reconstruction algorithm.

In the following sections, the fundamentals of SR methods are introduced. First, image quality assessment for evaluating SR techniques is presented, and then, the literature on SR is reviewed. The review gives a background on SR and provides a critical view of how those methods may improve image quality in pCLE. The list of the approaches proposed to tackle the SR task with a focus on deep learning in application to natural images is presented, and special attention is given to developments that are the most immediately relevant for pCLE. Finally, the limitation of applying SR to the pCLE domain are discussed.

2.1 Image quality assessment

Image quality (IQ) is a term describing the appearance of any flaws in an image [8, 62] and is affected by several factors, such as contrast, blur, noise and artefacts, etc. The performance of image processing techniques improving IQ, such as SR, can be measured by the change of the degradation in the reconstructed image and its effect on the perceptual quality of an image [62]. Therefore, Image Quality Assessment (IQA) aims to quantify the end-user experience related to observed images, and is used to benchmark SR algorithms [42].

IQA is categorised into two groups: the *objective assessment* that measures the IQ automatically with computational models designed to correlate with the human visual sensitivity, and the *subjective assessment* that involves a human to evaluate the quality of the images [42]. When a high-quality reference image is not available, images can be assessed with *non-reference* methods; otherwise, either *full-reference* or *reduced-reference* methods are used to test the difference between a ground truth image and a test image.

Objective IQA Over the years, many researchers have contributed objective measurements to the field of IQA [85], including but not limited to: PSNR [51], SSIM [31], MS-SSIM [25], UQI [20], VIF [39], MAD [72], FSIM [94], GCF [34], NIQE [101], BRISQUE [100] and LIPIPS [195]. However, due to the task's complexity, which is to mimic human perceptions, the challenge remains open [108]. For SR validation, SSIM and PSNR provide complementary information [190, 229], and they serve as baseline metrics when a reference ground truth is available.

The most commonly used metric to evaluate IQ is the peak signal-to-noise ratio (PSNR) [50]. PSNR is based on mean-square-error (MSE) and derives from the pixel difference between a test image $\hat{I}$ and a reference image I:

$$\mathrm{PSNR} = 10 * \log_{10}(\frac{l^2}{\frac{1}{k}\sum_{i=1}^{k}(I(i) - \hat{I}(i))^2}), \tag{2.1}$$

where l is a maximum pixel value, $\hat{I}$ is a test image, I is a reference image, k is number of pixels and i represents a pixel. PSNR measures the power of the signal at a pixel level. This metric correlates well with corruption by noise, and we expect the SR image to have a higher PSNR score than the LR image. Although PSNR performs well as a measure of denoising, it is not robust to structural changes, so any misalignment, colour, and luminance transformation between target and reference add bias to the quality estimation [68]. Ledig et al. demonstrated that images preferred by humans might not correspond to a high PSNR score [148]. Thus, the correlation of PSNR with perceived quality is not sufficient to be used on its own for evaluating the performance of SR algorithms [50].

To compensate for the limitation of PSNR, the Structural SIMilarity (SSIM) metric was proposed [31]. SSIM is based on structural similarity and the degradation of structural information. It is defined with a luminance μ_I and a contrast σ_I, which for image with NxM pixels are estimated as the mean and the standard deviation of image intensities, respectively. These measures are used to define comparison functions of the luminance C_l, the contrast C_c, the structure C_s between the samples taken as various windows from two images and given as:

$$C_l(I,\hat{I}) = \frac{2\mu_I\mu_{\hat{I}} + \mathrm{C}_1}{\mu_I^2 + \mu_{\hat{I}}^2 + \mathrm{C}_1}, \tag{2.2}$$

$$C_c(I,\hat{I}) = \frac{2\sigma_I\sigma_{\hat{I}} + \mathrm{C}_2}{\sigma_I^2 + \sigma_{\hat{I}}^2 + \mathrm{C}_2}, \tag{2.3}$$

$$C_s(I,\hat{I}) = \frac{\sigma_{I\hat{I}} + C_3}{\sigma_I\sigma_{\hat{I}} + \mathrm{C}_3}, \tag{2.4}$$

$$\sigma_{I\hat{I}} = \frac{1}{(NxM) - 1}\sum_{i=1}^{NxM}(I(i) - \mu_I)(\hat{I} - \mu_{\hat{I}}), \tag{2.5}$$

where $\hat{I}$ is a test image, I is a reference image, C_1, C_2, C_3 are constants for numerical stability and $\sigma_{I\hat{I}}$ is the covariance between I and $\hat{I}$. These three function compose to form the SSIM:

$$\mathrm{SSIM}(I,\hat{I}) = [C_l(I,\hat{I})]^{\alpha}[C_c(I,\hat{I})]^{\beta}[C_s(I,\hat{I})]^{\gamma}, \tag{2.6}$$

where α, β, γ are control parameters for adjusting their relative importance. Practically, SSIM is applied locally, as image statistic is spatially nonstationary and capturing the effect of the image distortions might be localised. The final score is the mean value of SSIM scores calculated per each window modified by a circular-symmetric Gaussian weighing function to avoid blocking artefacts. SSIM was designed based on the assumption that human visual perception is highly adapted for extracting structural information from a scene [31]. As a perception-based model, SSIM can mask luminance and contrast of pixel intensities. Minor structural changes and noise do not affect the metric score and allows measuring changes in local patterns on the image robustly.

Subjective IQA Human evaluation studies are preferred in general to the objective test as they allow us to compare the effectiveness of SR algorithms based on authentic psychometric measures [126]. An individual observer can determine the quality of images easily by judging their resolution, amount of noise, and artefacts etc. Images can be assessed as a single stimulus when only one image is displayed or a double stimulus when a test and a reference image are shown jointly [98, 126, 18]. Although human studies are the most accurate and reliable test of perceptual preferences, they are costly to complete and may suffer from biases such as observer's mood and alertness or electronic viewing devices used [98].

Mean Opinion Score (MOS) is one of the preferred ways to make systematic IQA questionnaires using a rating scale [84]. The user opinion can be matched up to the grading scale, which allows a participant to rate the quality of the image quantitatively. It allows querying humans about their subjective judgement on the quality of images and summarises their opinion with an arithmetic mean score. Questions used in MOS surveys can be specific and evaluate the user experience of noise, visibility of specific structures, or contrast enhancement etc. A MOS survey gives direct feedback to the end-user but is also slow, inconvenient and expensive. This evaluation is subjective and prone to variability coming from an individual perception.

The forced-choice pairwise comparison was shown to be the most time-efficient and accurate tool to assess images [98]. It is because a cognitively simple procedure generates the most consistent results. In the forced-choice test, observers need to compare only two images at a moment and make a quick binary decision with no rating scales. The limitation is that it requires a bigger number of comparisons to achieve statistical significance of the results, and the lack of a grading scale makes the analysis less straightforward [16].

The concept of IQ in medical imaging goes beyond the presented measures as it is necessary to check how improved images serve a clinician when making a diagnosis [139]. Each medical imaging modality provides anatomical or functional information that requires a qualified radiologist to interpret and to determine acceptable image quality [65]. The specific artefacts, such as chemical shift artefacts in MRI or atypical pCLE noise, seen in the medical field, differ from natural images. Thus, standard IQA should be adjusted to the specifics of the imaging modality limitation [139].

Improving the clinical experience of images is closely linked to the quality of the images [139, 65]. In this thesis, the primary focus of the investigation on pCLE SR algorithms is to verify whether SR reconstructions are interesting clinically and whether they improve the pCLE experience. The standard reference-based IQA metrics support medical image evaluation partly [49], which also applies to pCLE. Besides quantitative metrics, the surveys, such as MOS, have to be conducted with experts. The experts need to check if the SR images do not contain artefacts, and enhanced structures are valid clinically. It is also essential to ensure that image enhancement will not burden the experience of using the modality by introducing artefacts.

2.2 The concept of super-resolution

An imaging camera only ever permits taking a representation of a scene with limited information compared to the actual scene, because of digital sampling and sometimes the degrading effects of the sensor and manufacturing limitations of the optics. The image captured with the camera is a measurement of an undiscovered ground truth image as illustrated by the Nyquist theorem [23]. The relationship of low-resolution (LR) observed imaging data to high-resolution (HR) signal is interpreted via the observational model of the camera that describes the image degradation generated by motion, sensor and optical blur, aliasing effects, etc.

Transcending the resolution limitations of an imaging system with post-processing methods is known as super-resolution. Restoring missing high-frequency components and removing the degradation caused by the LR image acquisition process is challenging. The key idea behind SR is that using aliased HR information in combination with prior knowledge about the image space that is present in LR image signals allows reconstruction of an SR image. The SR process aims to retrieve an HR approximation $\widehat{I}_{HR}$ of the ground truth HR image from a single or several LR images based on the redundancy of information in the explored data and prior knowledge of the degradation process and HR properties.

Formally, the HR image I_{HR} is related to the LR image I_{HR} by the degradation process D defined as:

$$D(I_{HR}, \rho) = I_{HR} \downarrow_{s,h,g} + n, \{s,h,g\} \subset \rho, \tag{2.7}$$

where $\downarrow$ is the downsampling operation, and ρ are parameters of the degradation process, including but not limited to scaling s, the blurring h, the motion g factors and the noise n inherent from the imaging system. The degradation parameters shown in Eq. (2.7) are unknown, and hence they need to be estimated from LR images, which is a computationally complex and numerically ill-posed problem, as there are multiple estimates of high-frequency components, which are lost due to degradation. Those components are under-determined and as such cannot be recovered without injecting prior information.

We can solve Eq. (2.7) as the transform of a HR image degradation by warping, blurring, and sub-sampling operators contributing to produce an LR image. The degradation can be modelled either fully or partially depending on the approaches employed to generate super-resolved images [23]. The estimation of a HR output from LR is indicated as an inverse of Eq. (2.7):

$$\widehat{I_{HR}} = F(I_{LR}; \theta, D(\rho)), \tag{2.8}$$

where F is the SR process and θ denotes the parameters of F. Therefore, the aim of SR is to find F by minimising the cost function *loss* between the generated $\widehat{I_{HR}}$ and the test image I_{HR}, given as:

$$\widehat{\theta} = \arg\min_{\theta} loss(\widehat{I_{HR}}, I_{HR}) + \lambda \omega(\theta), \tag{2.9}$$

where $\omega(\theta)$ is the regularisation term which captures prior knowledge, and λ is the trade-off parameter.

The concept of signal sampling is dated back as early as 1915 [1]. With the digital sampling of the analogous signal, the concept of compression was introduced. Due to the occurring loss of information, the enhancing resolution of the digital imaging sensors came as the natural consequence. In 1984, Tsai and Huang proposed the initial work on SR [3] for images, and since that date, SR research has become widespread in the computer vision community [23]. There has been considerable progress in SR for natural images in recent years [127], and the state-of-the-art methods are in the majority based on DL [228]. There exists a very extensive literature on SR with many surveys [229, 49, 225, 9, 10, 105, 40, 60, 52, 23, 27, 22] and books [74, 14, 44] published and as well as regular conferences and workshops dedicated to the subject, such as NTIRE [171, 222, 190], PIRM [177], AIM [200, 205]. As the literature on this matter is vast and easy to find, in this Chapter, the main types of SR techniques are summarised and followed by a comprehensive survey of the most relevant works to the pCLE domain.

2.2.1 The first generation of super-resolution algorithms

Initial SR approaches were based on *single-image super-resolution* (SISR) and exploited signal processing techniques, including kernel methods applied to the input image [58]. SISR generates details based on prior information, such as a camera model or external image examples. The branch of SR that extracts and fuses LR information from the multiple frames to reconstruct details at HR image is called *multiple image super-resolution* (MISR) [10].

We can perform the estimation of degradation operators in the *frequency or spatial domain*. The frequency techniques, based on properties of the Fourier transform, employ a model handling multiple shifted LR images [75, 3, 36], yet these are limited in handling complex degradation models [10]. The *spatial domain* methods have received the most interest and includes methods that aim to recover SR pixels on HR grids including Bayesian methods [64, 24, 118], registration [5, 11, 69, 41] and sparse-based approaches [81, 70, 83].

We can divide SR methods as reconstruction-based and learning-based algorithms [105]. Reconstruction algorithms estimate fidelity functions with regularisation, employing prior knowledge to solve the ill-posed problem of estimating the HR image, and use multiple images in the spatial domain. Learning-based algorithms use training images to gain information for SR reconstruction, and most methods are SISR techniques that exploit machine learning, neighbours or sparse representation to obtain a mapping from LR to HR space.

Bayesian methods The reconstruction process can be regarded stochastically as an optimisation problem aiming at the optimal SR reconstruction. The degradation operators and LR image can be given as stochastic variables; thus we can formulate SR reconstruction in a fully Bayesian framework [74]. There are several methods to solve a problem and they vary by fundamental assumptions to estimate the degradation matrix, prior term, and statistical inference methods.

Registration is the process of aligning several slightly shifted images of the same view [23]. Irani et al. [5] gave rise to the use of registration-based SR by demonstrating how their simplified registration framework with only translation and rotation could help the resolution improvement. The theory behind applying registration to SR lies in assuming that LR inputs contain HR information in the sub-pixel space. The LR image is a unique aliased and sub-sampled versions of the HR image [74]. Thus, the sub-pixel alignment generating a dense mosaic image revealing more details and reconstructing the image with richer content thanks to a higher sampling rate and increased signal to noise ratio. The quality of SR reconstruction relies on the precise alignment of LR frames [10]. We can implement image registration in numerous ways by exploring intensity- and feature-based similarity between images, and transformation describing motion and deformation of the images [32]. It can be performed in a spatial and a frequency domain [26].

2.2.2 Deep learning-based super-resolution

Deep learning (DL) is a type of machine learning based on representation learning that has established itself as state-of-the art in a variety of computer vision and image understanding tasks [97]. There is an exponentially increasing interest in the use of convolutional neural network (CNN) architectures to tackle the SR problem [229, 81, 171, 222, 190, 177, 200, 205]. The main interests in CNN-based SR are in finding effective ways to train networks, developing novel training strategies, improving loss functions and perfecting CNN architectures.

Training scheme The most research interest has been around exemplar-based super-resolution (EBSR), which leverages the power of supervised training over an enormous collection of corresponding HR and LR images. EBSR methods currently represent the state-of-the-art for the SR task [171]. However, collecting pairs of realistic LR and HR is difficult and generally, data for the training process are created by downsampling HR images with bicubic kernels and anti-aliasing [175]. The rise of Generative Adversarial Networks (GANs) has allowed for weakly supervised learning that utilises unpaired images. The fundamental idea behind a GAN-based approach is to capture complex degradation from data with game theory-based training that uses a generator to create images from noise and a discriminator to establish whether generated results and randomly sampled images comes from the target distribution [179]. Evolving from GANs, image-to-image translation [176] brought the idea of cycle-GANs to be used for general SR. Yuan et al. trained an architecture with several generators and discriminators using losses, such as adversarial, cycle consistency and identity losses to learn the mapping in a loop by splitting training into stages from noisy LR to clean LR, and then from clean LR to clean HR [193]. Introducing unsupervised CycleGAN [176] allowed SR models to be robust to a lack of HR images, unknown downscaling including noise and blur and still produce state-of-the-art results compared to supervised models [193, 206]. Besides GANs, another training strategy that does not rely on extensive training data is a zero-shot (ZS) framework. It uses the internal statistics from one image only and a downscaling kernel to train a network blindly, achieving competitive SR reconstructions compared to supervised methods [189]. This approach can be extended with the use of GANs to estimate image specific kernels that can be plugged into the ZS framework enabling improved SR [196]. Although supervised SR with deep learning is still maturing, its limitation is the use of synthetic data. The research on blind SR is emerging, and yet still does not offer state-of-the-art results.

Loss function in SR The choice of a loss function strongly drives the behaviour of CNNs. In SR, there is a prevalent choice for losses based on pixel dissimilarity between images [229]. Although the most popular choice of the loss is a mean squared error (MSE), aka. L2, it was shown that L1 outperforms MSE in SR task helping model convergence [175]. Zhao et al. [175] proposed SSIM based loss that enables training models to achieve higher quality SR reconstruction without pixelation artefacts when compared to either L1 or L2. These losses steer the SR models towards the reconstitution of HR images with high PSNR and SSIM. However, they do not necessarily contribute good quality reconstructions in the subjective human opinion as they are often lacking high-frequency details or have over smooth details [148, 144, 31].

Since the objective of an ideal optimiser is to mimic the perceptual capabilities of the human visual cortex, perceptual metrics have been shown to outperform the distortion-aimed losses [148, 144, 193, 166]. For example, generative approaches employ custom adversarial loss based on extracted features that classify SR images as fake or true, helping to recover images preferred by humans regardless of their PSRN score [148]. Motivated by style transfer research [131, 142] texture loss using adversarial training has been proposed [166]. It was shown to contribute to the reconstruction of images with enhanced and more realistic textures, yet prone to generating elements not matching ground truth. Zhang et al. have shown that features extracted by any deep network can be adapted as a perceptual similarity metric [195] that can be a loss function as well. They designed a framework called LIPIPS that calibrates three deep networks by scoring the images given by raters. The LIPIPS outperforms the widely used IQA metrics. CNN-based losses have produced a break-through performance for the SR task, allowing reconstructing sharper images that contain high-frequency details, yet they are often prone to generating artificial information [229].

CNN architectures Super-Resolution Convolutional Neural Network (SRCNN) was the first proposed CNN architecture to solve SR task via deep learning [124], followed by improved Fast SRCNN [140]. F/SRCNN is based on a sparse-coding technique and is trained in an end-to-end manner to learn the correspondence between LR and HR patches. The authors showed that the method outperforms classical SR techniques. As the result, it became the baseline reference method to compare new SR models.

The pioneering work of Dong et al. [124, 140] created enormous interest in investigating novel architectures. Kim et al. inspired by VGG-net proposed a residual network VDSR [147] that learns residuals between LR and HR images, resulting in much faster convergence and even better reconstruction accuracy than SRCNN. Following the success of residual mapping, Enhanced Deep Residual Network for Single Image Super-Resolution (EDSR) based on SRResNet architecture[163] was proposed. The authors removed batch normalisation modules, justifying it with the fact that the SR task does not require feature normalisation, which actually lowers the performance of the SR model. A unique approach was showed with Pixel recursive super-resolution architecture based on PixelCNNs [160]. It is built as a probabilistic deep model that learns local pixels dependencies. This model can reconstruct plausible images for large scale factors but has considerable computational complexity with the run-time of the reconstructions being significant. Beyond supervised models, a Generative adversarial network for image super-resolution (SRGAN) was the first work utilising the power of the generative approach in SR task [148]. They trained a model with a selective loss function to classify HR images into super-resolved images and ground-truth HR images.

Video super-resolution (VSR) can be achieved by applying SISR techniques to the video sequence directly. However, the improvement in resolution can be even better when a spatio-temporal relation of frames in a video sequence is considered [9]. A video sequence comprises image frames that share similar content under some transformations, and to utilise that information, typical VSR embodies the alignment of frames through motion estimation, extraction and fusion of features via degradation models and restoration algorithms to reconstruct video. The SRCNN inspires vSRnet that extends convolution from 2D to 3D to be trained on both a spatial and a temporal dimension [145]. Detail-revealing Deep Video Super-resolution is a CNN framework for video SR which incorporates three-stage network motion estimation which is used to estimate inter-frame motion on the sub-pixel grid, SPMC Layer capable of compensating motion and aligning frames and Detail Fusion Net based on encoder-decoder architectures and performs reconstruction from low-level features [169]. The authors demonstrated that precise alignment of frames in video and compensation of the motion is an essential step in SR reconstruction of the sequence. Cabellero et al. used a spatio-temporal sub-pixel network named VESPCN exploiting fusion of frames from sequences with three types of joint processing such as early fusion, slow fusion and 3D convolutions [138]. Bidirectional recurrent convolutional network (BRCN) was proposed to perform efficient multi-frame SR reconstruction based on a fully convolutional framework [161].

Typically in a DL-based solution for classification, the flow of information through the network aims at shrinking image space by extracting meaningful and complex features. SR task requires expanding transformations, and CNNs are not capable of direct upscaling the input image by changing the number of pixels. Initially, a bicubic interpolation which upscales LR image to the size of HR image was used as the first non-trainable layer of the network [124]. As an alternative, the Sub-pixel convolutional layer was proposed to provide similar quality reconstruction to architectures using bicubic layers [154]. It learns an up-scaling filter, which allows for efficient expansions in the network.

To sum up, DL provides state-of-the-art machine learning algorithms capable of improving image quality by increasing the number of informative pixels in the SR process. Many methodologies were developed to bring image quality improvement such as supervised training when data are available as well as blind SR where data are harder to access. In order to train models, multiple loss functions were proposed starting from straightforward choice of MSE for very complex CNN-based feature extractors. Although the majority of the state-of-the-art solution is based on DL, non-DL developments are progressing substantially as well [212, 153]. All those developments were mainly exploited on natural images, yet there is the potential to bring them to medical imagining as well.

2.3 Image enhancement for fibre-bundled devices

Following extensive developments in SR for natural images over the last few years, which have been thoroughly studied and well documented in the literature as presented in Section 2.2 "The concept of super-resolution", there has been great interest within the medical imaging community in improving the image quality of imaging modalities [133]. The earliest proposed methodologies found in the literature in the majority are focused on the translation of SR techniques from natural images to the medical imaging domain [49]. The state-of-the-art SR methods are in the majority based on DL [228], and many studies have examined DL-based SR in medical imaging as well [219]. In the context of image restoration and denoising, GANs drew considerable attention [213]. Overall, many strategies have been proposed in the literature to deal with SR in imaging modalities such as MR [168, 76, 214], CT [158, 191], ultrasonography [59, 180], and PET [37, 226].

A review of the literature conducted shows that SR in endomicroscopy has been discussed infrequently compared to other imaging modalities. It can be attributed to the fact that fibre-based imaging is relatively new in the clinic. Most of the related works are concerned with image reconstruction. These aimed primarily at removing typical artefacts seen in fibre bundle imaging, such as honeycomb pattern, rather than actual super-resolution. The next sections review the SR techniques that have been implemented to improve image resolution of endomicroscopy, up to the current day. The Section 2.3.1 “Restoration and the first generation super-resolution” outlines early SR techniques applied to pCLE and Section 2.3.2 “The deep learning for super-resolution endomicroscopy” reports on existing DL-based solutions for tackling SR for images acquired with fibre-based bundles.

2.3.1 Restoration and the first generation super-resolution

Generic classical registration techniques, summarised in Section 2.2.1 “The first generation of super-resolution algorithms”, are not robust to typical distortions occurring in pCLE. These distortions are caused by irregular image grid, and non-rigid deformations caused by imagining a soft tissue. Researchers employ two major ways to adapt registration techniques to pCLE: a purely computational approach that aims to create wide FoV mosaics and accounts for pCLE specificity [41, 54, 69], and using hardware-based solutions to create many images with controlled shifts to facilitate high-quality image fusion leading to oversampling pCLE signal [123, 71, 113, 224, 192].

In pCLE, the work proposed by Vercauteren et al. [41] presents a demons registration-based framework. They implemented similarity matching adapted to align pCLE frames globally by using Riemannian manifolds. They also compensate for scanning-induced motion distortion and non-rigid tissue deformations. To adjust for pCLE specificity, they take into consideration irregularly sampled pCLE images. They proved this method improves the reconstructed endomicroscopic images and even reveals details that were not visible initially by augmenting spatial information when sufficient overlap of frames exist. They extended their work to enable real-time preview, showing its clinical relevance as a source of new information for clinicians examining patient [54].

Hu et al. have created mosaics robust to deformation occurring in organs and soft tissues and improve their quality by incorporating a SR method [69]. The core of the proposed pipeline is feature-based registration extended to handle potential outliers using the RANSAC algorithm. The registration explores the mapping between frames with a graph technique, and uses an iterative bundle adjustment that minimises misalignment error. Thanks to the abundance of multiple viewpoints available during registration, authors also incorporate maximum a posterior (MAP) estimation to combine information from different images leading to SR.

One major drawback of approaches proposed in [41, 69] is that they remain prone to residual misalignment. Since the SR task requires ideal alignment in HR image space, registration often fails in providing artefact-free SR reconstruction. The quality of registration is crucial to achieving SR, but the alignment of sequences of frames is not trivial because of deformations and artefacts in the images. Misalignment leads to incorrect fusion and consequently generates artefacts, such as ghosting. Further, registration is computationally expensive. It is a correlation method and requires many iterations and inverse operations to be solved, making this approach unsuitable and limited for real-time applications. Registration is affected by the motion during image acquisition, and as such is non-deterministic. Thus, image quality improvement depends on the probe's direction, speed, and coverage of imaging space, leading to only partial success in super-resolving images.

Based on the success in improving spatial resolution with mosaicking algorithms, Cheon et al. proposed a partial registration algorithm capable of real-time performance on data acquired with robotic system [123]. In the first step they filtered fibre signals using thresholding, to remove cladding artefacts. That step allowed them to discard uninformative background, and as a result, they obtained sparse frames. Those sparse images align with a reference image based on the entropy measure of a region of the image. The author demonstrated that their method allows creating a compound image with an improved resolution and tested their method on synthetic and real data examples. The major drawback of that method is that it assumes a random but small shift of the transverse probe motion, leading to little effectiveness in real-case scenarios where probe motion is more diverse. Additionally, they also suggest using a mechanical motor to induce probe motion suitable for the method requirements.

Several techniques are described in the literature to develop proof-of-concept platforms that complement registration algorithms to bring the most effective improvement in resolving images reconstructed for fibre bundled devices with oversampling techniques. As the downside of registration is to find the exact spatial transformation that allows for precise alignment of frames, several works propose incorporating hardware solutions, allowing for the acquisition of LR frames with controlled shifts. Those solutions cannot be easily translated to the clinic, because they would require either hardware changes or rigorous movement of the probe performed by the clinician which is not workable in practice. Nonetheless, there is a clear trend among researchers to not rely on computational approaches alone, thus some of those works are reviewed in the next paragraph.

Kyrish et al. demonstrated that increased sampling could be achieved thanks to aligning multiple images and fusing them into a single image [71]. They used a platform that allowed to pre-program a relative shift between probe and sample based on several trajectories. Authors have shown improvement in resolution and signal-to-noise (SNR) of the resulting image for both USAF target and biological sample. They suggest how to mimic the best sampling trajectory with a custom electromechanical actuator attached to the distal end of a bundle. Lee and Han [113] proposed a conceptually similar restoration method to [71]. The authors acquire four slightly shifted images in a square pattern and averaged them, allowing for the signal's superposition. They achieved the superposition thanks to the acquisition of images based on highly precise lateral movement by spatially dithering the probe. Nevertheless, both works did not study real-case scenarios when the robotic systems do not control motion between frames and the probe's movement pattern is random. K. Vyas et al. tackle the problem by proposing an imaging system with a piezoelectric transducer generating probe displacement to capture a signal with precise micro-shifts [192]. They extract those LR signals in the calibration stage and reconstructed with Delaunay triangulation-based interpolation, allowing for SR reconstructions. The recent work of C. Renteria et al. used a calibrated rotational mount to rotate the fibre bundle to acquire several images of the same scene from a different location that are co-registered with phase-based affine registration algorithm to get the final image with improved quality [224]. They acknowledge several limitations of their registration framework, such as a lack of robustness to honeycomb patterns that have to be filtered before registration, computational complexity, and the imperfection of registration, lowering the output quality.

Fibre bundle imaging system used in a high radiation environment suffer from similar imaging issues to pCLE. Zheng at al. propose an algorithm which removes the honeycomb pattern and improves the spacial resolution of images acquired with fibre-bundles [159]. They used thresholding, which is also used in pCLE [53], to create a fibre map and extract signals for reconstruction. They employ an in-painting algorithm based on improved non-local means (NLM) [130] utilising reoccurring patches within an image to fill the regions between fibres and generate continuous reconstructions with improved quality in comparison to baseline solution.

Eldaly at al. proposed a deconvolution framework and applied it to raw, noisy and irregularly spaced data acquired with coherent fibre bundle optical microscopy without a strong reliance on a fibre pattern [182]. They test two methods for estimation in the proposed hierarchical Bayesian model, such as Markov chain Monte Carlo and variational Bayes. To limit computational complexity, authors do not recreate the full image, but only the fibre signal. Both synthetic and real analysis have demonstrated that the proposed method improves the quality of the reconstruction from restored fibre signals.

Shao et al. proposed a framework that uses a forward model of the bundle to restore HR images from its LR counterparts [188]. They measure the PSF of each fibre and construct a fibre mapping embedded into their model as a fibre core operator and a geometric mapping matrix computed with an enhanced correlation coefficient image alignment technique. They employ scaled conjugate gradient to to compute the maximum a-posteriori solution given the prior of the basic Laplacian smoother. They evaluated their work on synthetic and real images, showing improvement of image quality, which increases with the number of LR images. The reconstructed images did not exhibit a honeycomb pattern, or any artefacts related to the reconstruction method itself.

Dumas et al. proposed a compressed sensing computational framework based on sparse-recovery algorithms handling multiple LR images acquired with using different simulated optical masks [181] in application to endomicroscopy [198]. Their approach enables recovery of a higher density of signals than one determined by the number of fibres in the bundle. This is achieved by oversampling thanks to intensity modulation implemented as parallel point scanning. They extended that work by proposing a snapshot spectral coding that uses a prism to capture shifted projections of the sample allowing imaging spaces normally lost in a single view due to cladding in between fibres [198].

Shin et al. demonstrated a compressed sensing (CS) approach for confocal endomicroscopy that uses spatially encoded speckle patterns to collect fibre signals for reconstruction with a total variation (TV) algorithm [167]. Han et al. designed a non-parametric formulation using the compressed sensing technique based on an iterative method for reconstruction of fibre bundle-based endoscopic optical coherence tomography that can be extended to any images acquired with fibre bundles [132]. Liu et al. formulate pCLE reconstruction as an estimation of HR from noisy and blurred LR based on a forward model with two algorithms: the two-step iterative shrinkage thresholding [43] algorithm with the TV regularisation presented in [220], and a fast iterative shrinkage-thresholding algorithm [55] proposed in [221].

The efforts in research on pCLE restoration were in majority focused on the first generation SR techniques, and are of limited use. The proposed methods do not perform well in terms of accuracy and are prone to artefacts. The methods based on the use of registration or Bayesian framework tend to be computationally expensive due to the inverse operations involved. Those first-generation methods fall short in providing a solution for real-time SR pCLE reconstruction.

2.3.2 The deep learning for super-resolution endomicroscopy

Several attempts at improving resolution with algorithmic solutions, in some cases supported by the advancement of hardware, were presented in Section 2.3.1 “Restoration and the first generation super-resolution”. The presented solutions mainly follow ideas known from first-generation SR algorithms, such as signal fusion utilising registration or statistical estimation of the forward model. The current state-of-the-art SR methods are based on DL and outperform first-generation algorithms significantly [171, 190, 222]. The next paragraphs outline the very few existing DL-based solutions presented in the literature specifically for tackling SR for pCLE. The main techniques for addressing this problem are focused on contending with the limitations of the availability of ground-truth HR data and generating a substitute for HR pCLE.

The authors of [208] built a custom optical system to acquire HR data for the training of their DL-based generative adversarial restoration neural model. They used a dual-sensor system to capture two images from the same light source by splitting the light beam in two. One path of the beam goes through a fibre bundle, and the other path is directed to the high definition (HD) optical system. In this way, the system captures both images simultaneously: LR images acquired via a fibre bundle and HR images captured via an optical microscope. Thanks to the precise image acquisition protocol, collected LR-HR image pairs have high accuracy geometrical alignment. Such a system is technically complex to set up and, at best, challenging to use in a clinical scenario. Clinical pCLE has high data variability, as data is collected from different devices, fibres, tissues and patients. While such a system may be suitable for lab testing, it would not integrate well within the clinical workflow and would thus be unsuitable for large-scale representative data collection.

Following the success of their single frame SR pipeline [208], Shao et al. proposed a deep learning method that handles multiple-frames [209]. They combine a motion estimation network that estimates motion parameters between consecutive raw frames with GT images' support. Then, frames in a sequence are aligned to the reference frame and input into 3D CNN. As the results aliased spatial-temporal information is used to estimate high-frequency signals. The research work is limited to sequences generated synthetically by the custom pCLE system, supported by the motion stage to shift the sample randomly. Both signals, LR fibre-based and

HR GT images are captured. Although such data are almost impossible to acquire in a real clinical scenario, they have shown very promising proof-of-concept on how to utilise the potential of temporal aliasing in DL-based image enhancement in endomicroscopy.

Our team investigated the application of a generative adversarial networks (GANs) to pCLE [206]; the work is coauthored but not reported as a contribution to this thesis. In contrast to Shao et al. [208] who used HR images acquired with their custom system, our work uses the potential of adversarial unsupervised training to get SR models without relying on HR pCLE images. The paired images HR-LR are unnecessary for training GANs. Instead, natural images were a source of HR data for the training. This work added a physically inspired cycle consistency loss to ensure that the proposed GAN architecture generates meaningful pCLE-like reconstructions. This constraint relies on the same reconstruction algorithm, which utilised the fibre pattern for the Delaunay triangulation-based interpolation technique presented in [187]. We demonstrated that the GAN generates promising SR pCLE reconstructions with improved image quality compared to baseline and state-of-the-art methods.

Xu et al. proposed conditional a GAN architecture that aims in recovering synthetic pCLE-like without honeycomb-like patterns [230]. They use a fibre bundle mask and a raw pCLE-like image as an attention channel, allowing the network to recover fibre regions better. Additionally, they use multi-scale losses and perceptual loss to achieve artefact-free reconstructions. The limitation of that study is that it uses basic simulated images that do not account for realistic signal loss related to an individual fibres' FoV. The performance of their GAN was not tested on realistic images.

pCLE images are very similar to CLE images, and that has attracted some research efforts in developing SR techniques [202, 186]. The fundamental similarity between CLE and pCLE is carrying the same view of tissue from the clinical perspective, yet there is a difference in their acquisition. The CLE uses a rather slow, distally-scanned single optical fibre with adjustable depth of the imaging plane, and pCLE uses a comparatively fast, proximally-scanned fibre bundle with the fixed plane [95]; thus, properties of collected signal such as noise and fibre responses are different for those modalities. The core difference, in the context of this thesis, is that pCLE requires reconstruction from irregularly sampled raw data, whereas CLE does not. The work presented in this thesis is focused on solutions for irregularly sampled data produced by bundles constructed with thousands of fibres, not one fibre like with CLE.

In [186] the authors designed a densely connected CNN based on DenseNet, and in [202] lightweight CNN introduced a novel weighting scheme based on attention modules using bilinear pooling. They have shown excellent properties of their CNN architectures being able to recover HR images from LR. Yet in both cases, they use a synthetic study where improvement for the original CLE images is not explored. The authors used synthetic LR images in both works and train networks against original CLE images as HR images.

Moreover, their synthetic experiments are limited to using a bicubic kernel as the downscaling method to generate LR images from the CLE images. Yet, the CLE relies on a different acquisition geometry, and this kernel is unsuitable for pCLE. Besides SR, the authors proposed a solution performing blind denoising that improves SNR of CLE without any reliance on GT data [201]. The authors designed an auto-encoder architecture trained with novel loss functions, such as auto-correlation and stationary losses, that allow distinguishing noise from the underlying signal, thereby outperforming state-of-the-art methods.

The DL methods presented in this thesis to improve pCLE image quality by estimation of high-resolution endomicroscopy are fundamentally different to the approaches proposed in the literature. In most works, researchers introduced the concept of SR to improve the spatial resolution of endomicroscopy concerning raw data. In this thesis, removing typical pCLE artefacts, such as honeycomb pattern, are not considered; SR solutions are applied to already restored images by the gold standard reconstruction algorithm. While the reviewed methods remove artefacts effectively, it becomes hard to assess how they improve the image's content beyond just restoration of the signal. Thus, it is difficult to compare those works directly with the development presented in this thesis. Among the proposed solutions, no method exists that gives a satisfactory solution to perform blind real-time SR that takes into account the physics of acquisition via pCLE fibre bundle from restored images. The lack of a relevant solution justifies the search for a novel approach that focuses on improving pCLE quality by estimating high-frequency information to augment endomicroscopies without reliance on GT HD data.

2.4 The challenges of super-resolution in endomicroscopy

SR is an ill-posed problem, and although many SR approaches have been proposed to solve that task, it is still an open challenge to improve image quality and resolution. There are bottlenecks related to the task fundamentals as well as the inherent limitations of methods solving SR task [225]. Specifically, in pCLE, the central challenges that have to be solved when developing the SR pipeline are optimising the execution time of the SR algorithm, a lack of ground truth HR data and adapting to the specifics of the acquisition that are very different to classical images.

Online performance The application of SR to enhance video streaming demands real-time performance. With pCLE, we seek the SR pipeline that integrates seamlessly into existing devices, allowing for online SR reconstruction of videos capturing 9-15 frames per second. Methods already considered for enhancement of pCLE such as Bayesian inference [188] that involves solving an inverse problem or registration [41, 69] that computes many inverse operations are suitable for the proof-of-concept or offline analysis but not optimal to be translated into a clinical tool because of their computational complexity. The pCLE SR algorithm needs either lower computational complexity [123] or optimisation [54] to provide on-demand, high-quality enhancement. DL-based techniques are not only giving state-of-the-art outcomes in image improvement but also meeting the criteria for *online* processing time. Although building DL models takes significant time and resources, since their training takes days using GPUs, the inference time of many DL-based

SR methods were shown to be close to real time [229]. Thus, DL are the right group of methods to examine for designing pCLE enhancement.

Ground truth HR data Predictive SR models rely on the availability of training data. Mainly, EBSR uses pairs of LR-HR images to learn the mapping from LR to HR space to improve the input LR image. It is inherently difficult to capture well aligned high-quality HR images and their LR counterparts in realistic scenarios. The requirement to miniaturise probes directly causes the limited quality of pCLE, and HD probes do not exist. The acquisition of high-quality pCLE images was made possible by designing the custom platform [208, 209]. Yet, that solution imposes changes in the hardware of endomicroscopy. Effectively, such a system becomes expensive and has little scalability as it is not appropriate for collecting a wide range of data from the clinic.

In the majority of works based on supervised training of DL networks, LR images are generated synthetically from high-quality HR images by downsampling with a known kernel [229]. Although such a solution may be sufficient when used on large, virtually noise-free HD natural images [210], it becomes impossible to apply in the pCLE domain. Foremost, since HR pCLE data does not exist, using the original noisy pCLE data as a substitute for ground truth HR data is limited. Such an approach was proposed for CLE in [185, 202], but it allows building only proof-of-concept SR models. As that approach does not improve the quality of the original images, but only tests the synthetic case with the downsampled original images. The availability of prominent noise and artefacts at the original pCLE images impedes the DL-model's effectiveness, as training works the best only with the high-quality HR reference images.

Another way to tackle the lack of HR pCLE images is to re-purpose data from another domain, e.g. natural images, to target the pCLE domain. It was made possible by using pCLE-specific constraint loss to train GANs [206] and simulating pseudo endomicroscopies with the process encapsulating a simple geometric model of the fibre bundle [230]. As concerns regarding the use of GANs in medical imaging have been voiced. The core principle behind the power of GANs is their capability to generate images, often using noise as input to the generative network, mimicking target data distribution. Due to the nature of the generation of images, GANs may create artificial images that match the learnt distribution but introduce non-realistic elements. Even though many successful works on GAN have been published [213, 206, 230], there is still uncertainty about whether GAN-based models generate results that deviate from the target domain.

There is clearly scope for developing a simulation of realistic HR pCLE data and an SR DL algorithm that can be trained blindly without ground truth data.

Downscaling kernel The most popular kernel choice considered in the SR literature is the bicubic kernel [229]. It has been shown that a bicubic kernel, often chosen as the default approach for simulating LR images, fails considerably when applied to real images [189, 171]. Researchers have shown that the reconstructions obtained from models trained on the known realistic kernels have higher image quality than the images reconstructed from models trained with the default kernel choice [171, 189]. In reality, the downscaling kernel is typically not known. When choosing one, several important technical considerations need to be taken, such as the optical model and the noise model. Given the strong association between the downscaling kernel and the quality of the SR reconstruction, it is essential to estimate the pCLE kernel as realistically as possible by taking into account the acquisition physics, such as irregularly distributed signals and noise patterns.

DL modelling SR models are expensive to train, as they require both optimisations of the architecture and training schema. The critical issue with very complex models is the convergence speed of the training. Several solutions were proposed to tackle this problem in training DL models, such as using a very high learning rate for network training [147], and removing batch-normalisation modules [163]. Not only does training optimisation play an essential role for modelling in SR, but also employing specific techniques in DL modelling can improve its performance. Specifically, in DL-based SR, Timofte et al. reported several factors that impact reconstruction quality, such as augmentation of data, use of large dictionaries with efficient search structures, cascading, image self-similarities, back-projection refine-ment, enhanced prediction by consistency check, and contextual reasoning [157]. Authors bench-marked the influence of those techniques using SRCNN [140] and showed positive outcomes for super-resolved images.

In pCLE, the main focus of this work is to design a lightweight network that can perform close to real-time enhancement. Thus, densely connected networks with many skip connections such as [202], and very deep and convoluted architectures such as [160] are not suitable for online inference. CNNs perform convolutions on a regular grid and are not designed to handle sparse data well [173]. Since pCLE fibre signals are irregularly distributed to be input to the CNN network, they should be first interpolated. Moreover, the reconstruction of pCLE signals to Cartesian images is highly non-linear, since it is based on triangulation, and adding back-projection would be not trivial to implement with sub-pixel precision.

To sum up, compared to natural images, pCLE brings unique needs that require special attention when developing a suitable SR solution. DL methods may allow for real-time state-of-the-art image enhancement, yet there need to be adaptation to pCLE. This research work will focus the adjustments on creating a downscaling kernel that mimics pCLE signal loss authentically, handling the irregularity of the pCLE signal, working around the lack of ground truth HR data and focusing on leveraging the power of CNNs and reducing their computational overhead.

Chapter 3

Single-image super-resolution for endomicroscopy

This Chapter is based on the research work **“Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction”** by Agnieszka Barbara Szczotka*, Daniele Ravì*, Dzhoshkun Ismail Shakir, Stephen P Pereira, and Tom Vercauteren published in *International Journal of Computer Assisted Radiology and Surgery*, 13.6 (2018), pp. 917–924.

Contents

3.1 Motivation: algorithmic improvement of pCLE image quality

In the Chapter 1 “Introduction”, we presented pCLE showing multiple clinical applications of this state-of-the-art imaging technique used as *in situ* and real-time *in vivo* optical biopsy. Despite the apparent success of pCLE in clinical practice, there are several constraints of the system affecting the quality of pCLE. As discussed previously, physics of the pCLE hardware, particularly using a fibre bundle for image acquisition and the current reconstruction algorithm, contributes to lower image quality by limiting its resolution and generating artefacts. The pCLE image quality impacts directly its performance in the clinical workflow [141]. Thus, it is clear that an improvement in the image quality of pCLE is vital for this imaging device to become a widely established diagnostic tool.

The improvement in pCLE image quality needs to meet the prerequisites of the system and be seamlessly implemented into the clinical workflow without any changes to its hardware. pCLE success is hugely attributed to its capability to provide microscopic images, and the fact that technology generates images in real time, instantly providing imaging feedback to the clinicians for the diagnosis and treatment. Building on the idea that online improvement in the resolution of pCLE images is desired, the main focus of this Chapter is to investigate an algorithmic solution for improving the quality of pCLE that has capabilities of the real-time performance and could be implemented into existing pCLE devices.

In this Chapter, we explore advanced single image super-resolution (SISR) techniques that can contribute to effective improvement in image quality. Although SISR for natural images is a relatively mature field, this work is the first attempt to translate these solutions into the pCLE context. Beyond SISR, video registration techniques [41] have been proposed to increase the resolution of pCLE. Such methods provide a baseline SR technique, but suffer from artefacts and are computationally too expensive to be applied in real time.

Because of the recent success of deep learning for SISR on natural images [171], this work focuses on exemplar-based super-resolution (EBSR) deep learning techniques. In principle, EBSR encapsulates data-driven methods that rely on learning a mapping from examples of LR to HR images. However, the translation of these methods to the pCLE domain is not straightforward, notably due to the lack of ground-truth HR images required for the training. There is indeed no equivalent imaging device capable of producing higher-resolution endomicroscopic imaging, nor any robust and highly accurate means of spatially matching microscopic images acquired across scales with different devices. Furthermore, in comparison with natural images, currently available pCLE images suffer from specific artefacts introduced by the reconstruction procedure that maps the tissue signal from the irregular fibre grid to the Cartesian grid. Despite the shortcoming arising from lack of training ground-truth data, and notably due to the potential of the data-driven EBSR, there is clear scope for adapting deep SISR directly to pCLE.

The contribution presented in this Chapter is threefold. First, three deep learning models for SISR are examined on the pCLE data. Second, to overcome the lack of ground-truth LR/HR image pairs for training purposes, a novel pipeline that generates pseudo-ground-truth data by leveraging an existing video registration technique [41] is proposed. Third, in the absence of a reference HR ground truth, to assess the potential clinical impact of our approach, a Mean Opinion Score (MOS) study was conducted with experts (1-10 years of experience), each assessing images according to three different criteria. To our knowledge, this was at the time of publication the first research work to address the challenge of SISR reconstruction for pCLE images based on deep learning. It generates pCLE pseudo-ground-truth data for training of EBSR models and demonstrates that pseudo-ground-truth trained models provide convincing SR reconstruction.

The organisation of this Chapter is as follows: the Section 3.2 “Novel simulation of realistic synthetic endomicroscopies” presents the proposed methodology of realistic pseudo-ground-truth generation; the Section 3.3 “Implementation details” presents the state-of-the-art for SISR and the implementation of effective training of the super-resolving models; Section 3.4 “Results” gives information on the evaluation of our approach using a quantitative IQA and a MOS study; Section 3.5 “Discussion and conclusions" summarises our contribution to pCLE SISR presented in this Chapter.

3.2 Novel simulation of realistic synthetic endomicroscopies

3.2.1 Dataset

In order to simulate synthetic endomicroscopies for the training of the SISR architectures, we exploit the Smart Atlas database [82]. The dataset is a collection of 238 anonymised pCLE video sequences acquired with Cellvizio. The atlas gathers videos presenting tissues of the colon and oesophagus with both non-neoplastic and neoplastic pathology types. Original data were acquired with 23 unique probes of the same pCLE probe type. Each video contains a different number of frames, on average 250 frames. The images have a diameter of approximately 500 pixels that corresponds to a field of view of 240 µm.

We use these video sequences, which are considered as original LR (LR_{org}) images, to generate pseudo HR images, as described in Section 3.2.2 "Pseudo-ground-truth image estimation based on video registration". The simulated HR images are used in our proposed simulation framework to generate synthetic LR (LR_{syn}) detailed in Section 3.2.3 "Generation of realistic synthetic pCLE data". The pipeline showing the entire simulation of realistic synthetic endomicroscopies is presented in Figure 3.1; next sections give details on each stage of the proposed simulation.

3.2.2 Pseudo-ground-truth image estimation based on video registration

In Chapter 2 "Background on super-resolution", we reviewed SR, which has received a lot of interest from the computer vision community in the recent decades [23]. Initial SR approaches were based on SISR and exploited signal processing techniques applied to the input image. In this Chapter, we exploit an alternative to SISR, which is multi-frame image SR, which is based on the idea that HR image can be reconstructed by fusing many LR images. Ideally, the combination of

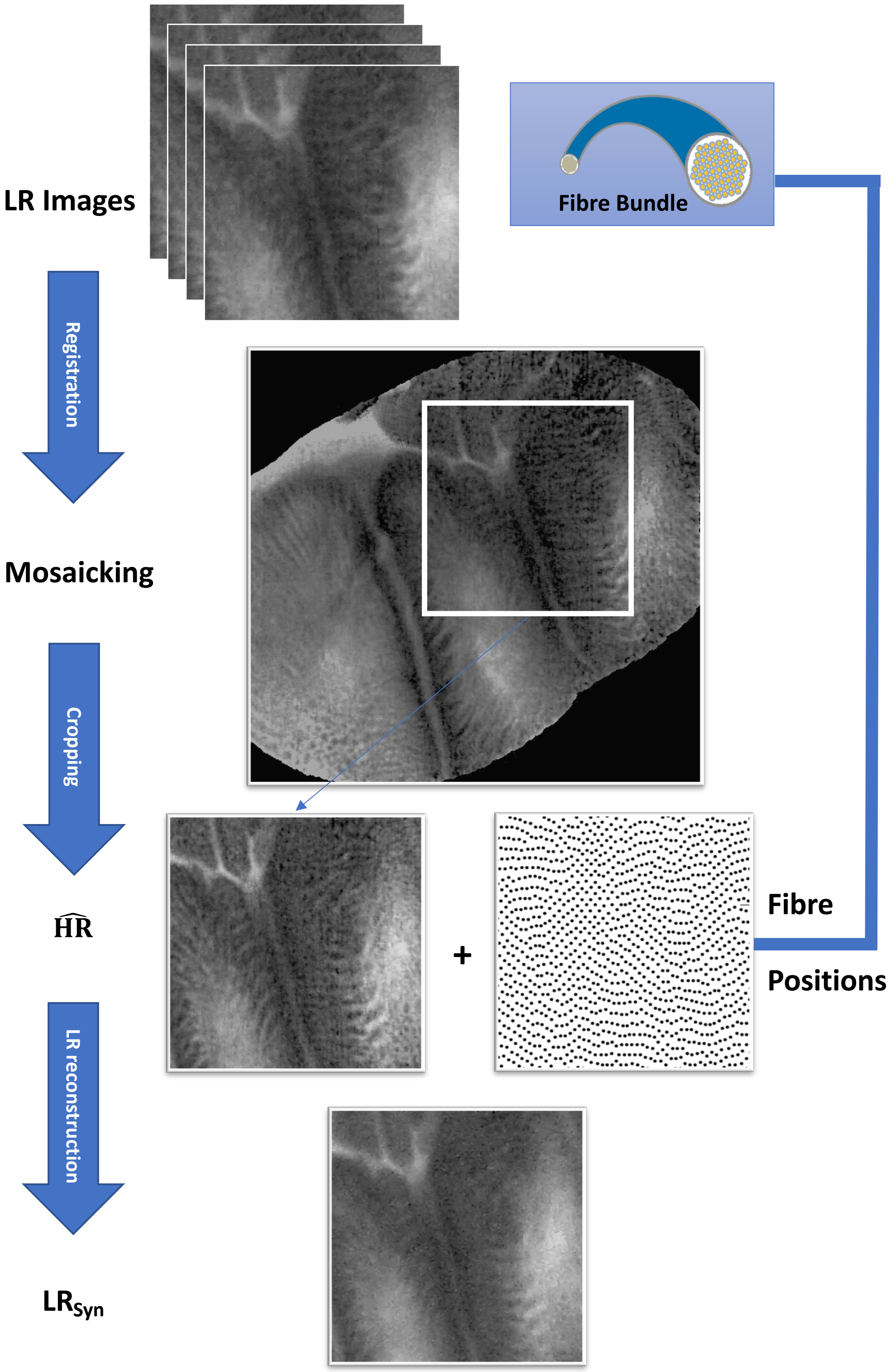


Figure 3.1: Pipeline used to generate LR synthetic images (LR_{syn}). The original pCLE video sequences (LR_{org}) are used to create mosaics, which next are cropped to pseudo HR images to finally be used for reconstruction of synthetic LR images (LR_{syn})

several LR image sources enriches the information content of the reconstructed HR image and contributes to improving its quality. Registration can be used to merge LR images acquired at slightly shifted field-of-views into a unified HR image.

In the specific context of pCLE, the work proposed by Vercauteren et al. [41] presents a video registration algorithm that can augment spatial information of the reconstructed pCLE image if sufficient overlap of fused frames occurs, and reveals details that were not visible initially. The quality of the registration is a key step to the success of the SR reconstruction. To achieve high-quality registration, we preselected subsequences from Smart Atlas videos. We have ensured that consecutive frames that only capture one scene represent continuous tissue signal without disrupted optical flow. The pre-selection of sequences allowed to get optimal results with registration proposed in [41]. The alignment of images captured at different times is not trivial, thus any discontinuity of the optical flow in the video sequence makes frames fusion with the registration challenging. Residual misalignment leads to incorrect fusion and generates artefacts, such as ghosting. Moreover, registration is a computationally expensive technique, making this approach unsuitable for realtime purposes.

In this work, to compensate for the lack of ground-truth HR pCLE data, we used the registration-based mosaicking technique [41] to estimate HR images. We refer to this algorithm in Section 2.3.1 “Restoration and the first generation superresolution”, where we give its in-depth critical review. The implementation of the mosaicking algorithm, including mosaic-to-image inverse transformation, was provided as proprietary software by Mauna Kea Technologies. The diagram of the mosaicing algorithm is presented in Figure 3.2, and the in-depth description is in [41]. Mosaicking acts as a classical SR technique and fuses several registered input frames by averaging the temporal information. The mosaics were generated for the selected sequences from the Smart Atlas database and used as a source of HR frames (Figure 3.1).

Since mosaicking generates a single large field-of-view mosaic image from a collection of input LR images, it does not directly provide a matched HR image

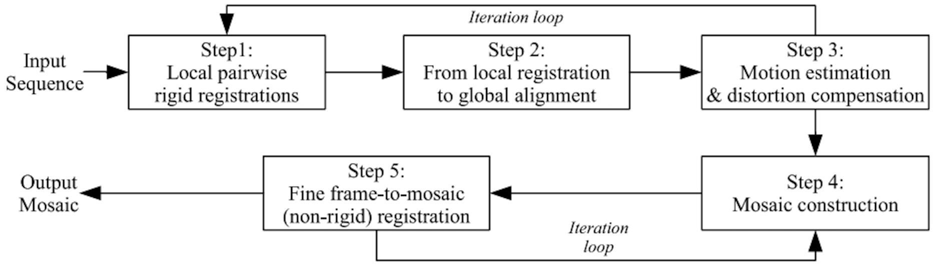

Iteration loop
Input Sequence
Step1: Local pairwise rigid registrations
Step 2: From local registration to global alignment
Step 3: Motion estimation & distortion compensation
Output Mosaic
Step 5: Fine frame-to-mosaic (non-rigid) registration
Step 4: Mosaic construction
Iteration loop

Figure 3.2: ”Block diagram of the mosaicing algorithm” taken from [41].

for each LR input. To circumvent this, we used the mosaic-to-image diffeomorphic spatial transformation resulting from the mosaicking process, implemented based on work [41] in the provided proprietary software, to propagate and crop the fused information from the mosaic back into each input LR image space. The image sequences resulting from this method are regarded as estimates of HR frames. These estimates will be referred to as $\widehat{HR}$ in the text.

The image quality of the mosaic image heavily depends on the accuracy of the underpinning registration, which is a difficult task. The corresponding pairs of LR and $\widehat{HR}$ images generated by the proposed registration-based method suffer from artefacts, which can hinder the training of the EBSR models.

Specifically, it can be observed that alignment inaccuracies occurring during mosaicking were a source of ghosting artefacts, which in combination with residual misalignments between the LR and $\widehat{HR}$ images, creates unsuitable data for the training. Sequences with obvious artefacts were manually discarded. However, even on this selected dataset, training issues were observed. To address these, we simulated LR-HR image pairs for training EBSR algorithms while leveraging the registration-based $\widehat{HR}$ images as realistic HR images.

3.2.3 Generation of realistic synthetic pCLE data

Currently available pCLE images are reconstructed from scattered fibre signal. Every fibre in the bundle acts as a single-pixel detector. To reconstruct pCLE images on a Cartesian grid, Delaunay triangulation and piecewise linear interpolation are used. The simulation framework developed in this study mimics the standard pCLE reconstruction algorithm and starts by assigning to each fibre the average of the signal from seven neighbouring pixels [28]. In the standard reconstruction algorithm, the fibre signal, which includes noise, is then interpolated. Similarly, the noise was added to the simulated data to produce realistic images and avoid creating a wide domain gap between real and simulated pCLE images.

Despite some misalignment artefacts, the registration-based generation of $\widehat{HR}$ presented in Section 3.2.2 "Pseudo-ground-truth image estimation based on video registration" produces images with fine details and a high signal-to-noise ratio. Our simulation framework (Figure 3.1) uses these $\widehat{HR}$ and produces simulated LR images with a perfect alignment.

The proposed simulation framework relies on observed irregular fibre arrangements and corresponding Voronoi diagrams. Each fibre signal was extracted from an $\widehat{HR}$ image, by averaging the $\widehat{HR}$ pixel values within the corresponding Voronoi cell.

To replicate realistic noise patterns on the simulated LR images, additive and multiplicative Gaussian noise (a and m respectively) is added to the extracted fibre signals fs to obtain a noisy fibre signal nfs as:

$$nfs = (1+m) * fs + a \tag{3.1}$$

The standard deviation of the noise distributions was chosen based on visual similarity between LR_{org} and LR_{syn} and between their histograms. Sigma values were 0.05 and 0.01*$(max\ fs - min\ fs)$ for multiplicative and additive Gaussian distribution respectively. In the last step, Delaunay-based linear interpolation was performed, thereby leading to our final simulated LR images.

LR and $\widehat{HR}$ images were combined into two data-sets: 1. Original pCLE ($pCLE_{org}$) built with pairs of LR_{org} taken from sequences of Smart Atlas database and $\widehat{HR}$ images, and 2. synthetic pCLE ($pCLE_{syn}$) built by replacing the LR_{org} images with LR_{syn} images. The purpose of these two datasets is to investigate their usability in the training of EBSR. Two types of experiments were run: using $pCLE_{org}$ to train models on original pCLE sequences, and using $pCLE_{syn}$ to train models on synthetic pCLE sequences. The performance of those models was evaluated to check how original or synthetic training data can be used to achieve SR of original pCLE sequences. The results of the investigation are presented in Section 3.4 “Results”.

3.3 Implementation details

Although many research groups have worked on deep-learning-based SR for natural images, and although CNNs are currently widely used in various medical imaging problems [165], only recently have CNNs been used for SR in medical imaging [168]. Currently, there are no established architectures for SR task in pCLE, and this is the first work on using SR for pCLE. We chose recently developed state-of-the-art architectures developed for natural images and train them for pCLE SR. In this study, our main aim is to show that synthetic data are suitable for pCLE training. Thus, we selected three methodically different approaches to test their suitability for generating effective SR models train on synthetic pCLE data. The next Section gives a short overview of selected architectures. We refer the reader to the original works for the visual presentation of the architectures and their implementation details.

3.3.1 Super-resolving networks

We used three CNNs networks for SR: sparse-coding based FSRCNN [140], residual based EDSR [163], and generative adversarial network SRGAN [148]. Every network was trained with the two pCLE datasets original $pCLE_{org}$ and synthetic $pCLE_{syn}$ presented in Section 3.2.3 “Generation of realistic synthetic pCLE data”.

Following the success of SRCNN [124], the first neural network for SR in natural images, authors proposed improved SRCNN architecture called Fast Super-Resolution Convolutional Neural Networks (F/SRCNN) [140]. F/SRCNN is based on sparse-coding technique and is trained with dictionaries of LR and HR patches. The network is trained in an end-to-end manner to learn the correspondence between input LR and output HR patches. The authors showed that the method outperforms classical SR techniques such as Bicubic SC [80]. F/SRCNN is currently commonly used in SR works as the reference method to compare the performance of new SR models.

Enhanced Deep Residual Network for Single Image Super-Resolution (EDSR) is based on SRResNet architecture [163]. EDSR uses a residual connection to learn the difference between input LR and output HR images, as it is more effective than learning their direct mapping. The authors also proposed to remove batch normalisation modules from their architecture. They justify it by claiming that the SR task does not require feature normalisation. This normalisation lowers the performance of the SR model.

Generative adversarial network for image super-resolution (SRGAN) was first work utilising the power of generative approach in SR task [148]. In contrast to previous networks, SRGAN is trained to generate the most probable SR reconstruction. A model is trained with a selective loss function. The authors designed an adversarial loss to classify HR images into super-resolved images and ground-truth HR images. Based on a MOS study, the authors showed that the participants perceived the quality of the restored HR images as higher compared to the image quality measured only by a PSNR score. They show that the SR task is hard to validate due to the ill-posed nature of assessing the quality of the image by the use of metrics. Although their SR was at that time outperforming other methods by producing highly pleasing images, authors were concerned with the application of their generative architecture to medical images.

3.3.2 Training details

The generated synthetic database was split into three subsets: training set (70%), validation set (15%), and test set (15%). Each subset was created ensuring that colon and oesophagus tissue were equally represented. The SR models are specific to the type of the probe but generic to the exact probe being used. Thus, the models do not need to be retrained for probes of the same type. Another type of probe, such as nCLE, would require a specifically trained model. nCLE and pCLE differ by the number of optical fibres and the design of the distal optics.

The data-sets were pre-processed in three steps. First, intensity values were normalised: $LR = LR - mean_{lr}/std_{lr}$ and $HR = HR - mean_{lr}/std_{lr}$. Second, pixels values were scaled of every frame individually in the range [0-1]. Third, non-overlapping patches of 64×64 pixels were extracted for the training phase, considering only pixels in the pCLE FoV. A stochastic patch-based training was used for training the networks, with a minibatch of size 54 patches to fit into the GPU memory (12GB).

The FSRCNN architecture was implemented based on the details in the original work [140] and official website[1], and EDSR architecture was implemented based on the details in the original work [163] and the official code published by authors at github[2]. The DL framework used to implement those networks was TensorFlow[3]. The SRGAN was trained based on publicly available code[4] implemented in PyTorch. Networks were trained with Adam as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, for 1000 epochs, learning rate 0.001 with weights initialisation drawing randomly from a Gaussian distribution with zero mean and standard deviation 0.001. The models were evaluated on the test set periodically every 100 steps.

Models were trained with patches from the training set. The patches from the validation set were used to monitor the loss during training, with the purpose to avoid overfitting. Since all the considered networks are fully convolutional, the test images were processed full size and no patch processing is required during the inference phase.

The behaviour of CNNs, especially in the context of SR, is strongly driven by the choice of a loss function, and the most popular one is MSE [175]. Zhao et al. [175] showed that MSE has two limitations: it does not converge to the global minimum and produces blocky artefacts. In addition to demonstrating that L1 loss outperforms MSE, the authors also introduced a new loss function SSIM+L1 by incorporating SSIM [31]. Accounting for these findings, we train FSRCNN and EDSR considering independently both L1 and SSIM+L1 to investigate their applicability for our data based on a quantitative comparison. Although this loss function steers the SR models towards the reconstitution of HR images with high

[1] http://mmlab.ie.cuhk.edu.hk/projects/FSRCNN.html
[2] https://github.com/LimBee/NTIRE2017
[3] https://www.tensorflow.org
[4] https://github.com/leehomyc/Photo-Realistic-Super-Resoluton

peak signal-to-noise ratios, this does not necessarily mean that the final images will provide a good quality perception, thus we also train a Generative network SRGAN with adversarial loss.

3.4 Results

Acknowledging the lack of proper ground truth for SR of pCLE and the ambiguous nature of established IQA metrics, a three-stage approach was designed for the evaluation of the proposed method using the three SR architectures considered in Section 3.3 “Implementation details”.

The first stage, presented in Section 3.4.2 “Experiments on synthetic data” and relying on the quantitative assessment, demonstrates the applicability of EBSR for pCLE in the ideal synthetic case where ground truth is available. In this quantitative stage, the inadequacy of the existing video registration-based high-resolution images as ground truth for EBSR training purpose is demonstrated.

The second stage, presented in Section 3.4.3 “Experiments on original data”, focuses on the quantitative assessment of our methods in the context of real input images and on the evaluation of our best model against other state-of-the-art SISR methods.

In the third stage, presented in Section 3.4.4 “Semi-quantitative analysis (MOS)” and performed to overcome the limitations of the quantitative assessment, a MOS study was carried out by recruiting nine independent experts, having 1-10 years of experience working with pCLE images.

3.4.1 Quantitative image quality assessment

For the quantitative analysis, the SR images were examined exploiting two complementary metrics: (i) SSIM to evaluate the similarity between the SR image and the $\widehat{HR}$ described in Section 2.1 “Image quality assessment”, and (ii) Global Contrast Factor (GCF) [34] as a reference-free metric for measuring image contrast which is one of the key characteristic of image quality in our context. The key idea behind GCF is to encapsulate overall image contrast as a weighted composition of local contrast measured in several different resolutions of the original image. The GCF

Table 3.1: Quantitative results obtained on full-size images from the test set for different training and testing strategies. The best results for each section are highlighted in bold.

metric method	SSIM with $\widehat{HR}$	GCF with $\widehat{HR}$	GCF with LR	Tot_{cs}
train on $pCLE_{syn}$ test with LR_{syn}				
LR	0.81±0.06	-0.42±0.31	0±0	0.46
EDSR L1	**0.87±0.06**	-0.22±0.13	0.21±0.31	0.67
FSRCNN L1	0.86±0.06	-0.26±0.16	0.17±0.19	0.63
EDSR SSIM+L1	**0.87±0.06**	-0.10±0.13	0.32±0.32	**0.71**
FSRCNN SSIM+L1	0.86±0.06	**-0.06±0.16**	**0.36±0.23**	**0.71**
SRGAN	0.76±0.06	-0.09±0.34	0.34±0.22	0.47
train on $pCLE_{org}$ test with LR_{org}				
LR	0.81±0.06	-0.24±0.37	0±0	0.44
EDSR L1	**0.83±0.06**	-0.24±0.29	0.01±0.13	0.50
FSRCNN L1	0.82±0.06	-0.15±0.30	0.09±0.11	0.52
EDSR SSIM+L1	0.82±0.06	-0.11±0.30	0.13±0.13	0.54
FSRCNN SSIM+L1	0.82±0.06	**-0.01±0.32**	**0.24±0.12**	**0.57**
SRGAN	0.75±0.05	-0.10±0.37	0.14±0.15	0.37
train on $pCLE_{syn}$ test with LR_{org}				
LR	0.81±0.06	-0.24±0.37	0±0	0.44
EDSR L1	0.81±0.06	0.18±0.29	0.42±0.35	0.61
FSRCNN L1	**0.82±0.06**	0.05±0.27	0.29±0.22	0.58
EDSR SSIM+L1	0.80±0.06	**0.33±0.31**	**0.57±0.36**	**0.65**
FSRCNN SSIM+L1	0.81±0.06	0.23±0.28	0.47±0.26	0.64
SRGAN	0.75±0.06	-0.04±0.44	0.21±0.26	0.38

is calculated only in grey-scale after transformation to perceptual luminances with gamma correction. The average local contrast C_i for current resolution is calculated as:

$$C_i = \frac{1}{w*h} * \sum_{i=1}^{w*h} lc_i, \tag{3.2}$$

where for each pixel i in the image with w width and h height, the local contrast lci is an average of 4 neighbouring pixels. The C_i is calculated for a series of images, where each image is twice smaller than its original version. In that way, four pixels in the bigger image becomes one pixel in the smaller image. The GCF is calculated as a weighted average over C_i from different resolutions, and weights were estimated by authors:

$$GCF = \sum_{i=1}^{N} wg_i * C_i, \tag{3.3}$$

where wg_i weigh factor, C_i is average local contrast.

Analysing both SSIM and GCF in combination leads to a more robust evaluation. SSIM alone cannot be dependent on when the reference image is unreliable, while improvements in GCF alone can be achieved deceitfully, for example, by adding a large amount of noise. Using these metrics, six scores for each SR method were extracted: mean and standard deviation of i) SSIM between SR and $\widehat{HR}$, ii) GCF differences between SR and LR and iii) GCF differences between SR and the $\widehat{HR}$. Finally, to determine which approach performs better, a composite score Tot_{cs} was defined. The Tot_{cs} captures both metrics SSIM and GCF as their arithmetic average. The Tot_{cs} represents effective quality improvement as both structural similarity and contrast improvement is weighted equally. To achieve the equal impact of the metrics on Tot_{cs}, scores are normalised and re-scaled to the range [0,1]. In our quantitative assessment, the score obtained by the initial LR_{org} was considered as a baseline reference.

3.4.2 Experiments on synthetic data

In the first experiment, synthetic data are used to demonstrate that our models work in the ideal situation where ground truth is available. The first section of Table 3.1 shows the scores obtained when the SR models are trained on $pCLE_{syn}$ and tested on LR_{syn}. Here it is evident that the EDSR and FSRCNN trained with SSIM+L1 obtain a noticeable improvement on the different quality factors with respect to the LR image. More specifically, in comparison with the initial LR image, the SSIM was increased by +0.06 when EDSR is used and by +0.05 when FSRCNN is used. These approaches also yield a GCF value that is very close to the GCF in $\widehat{HR}$ and an improvement of +0.32 and +0.36 in the GCF with respect to LR images. Statistical

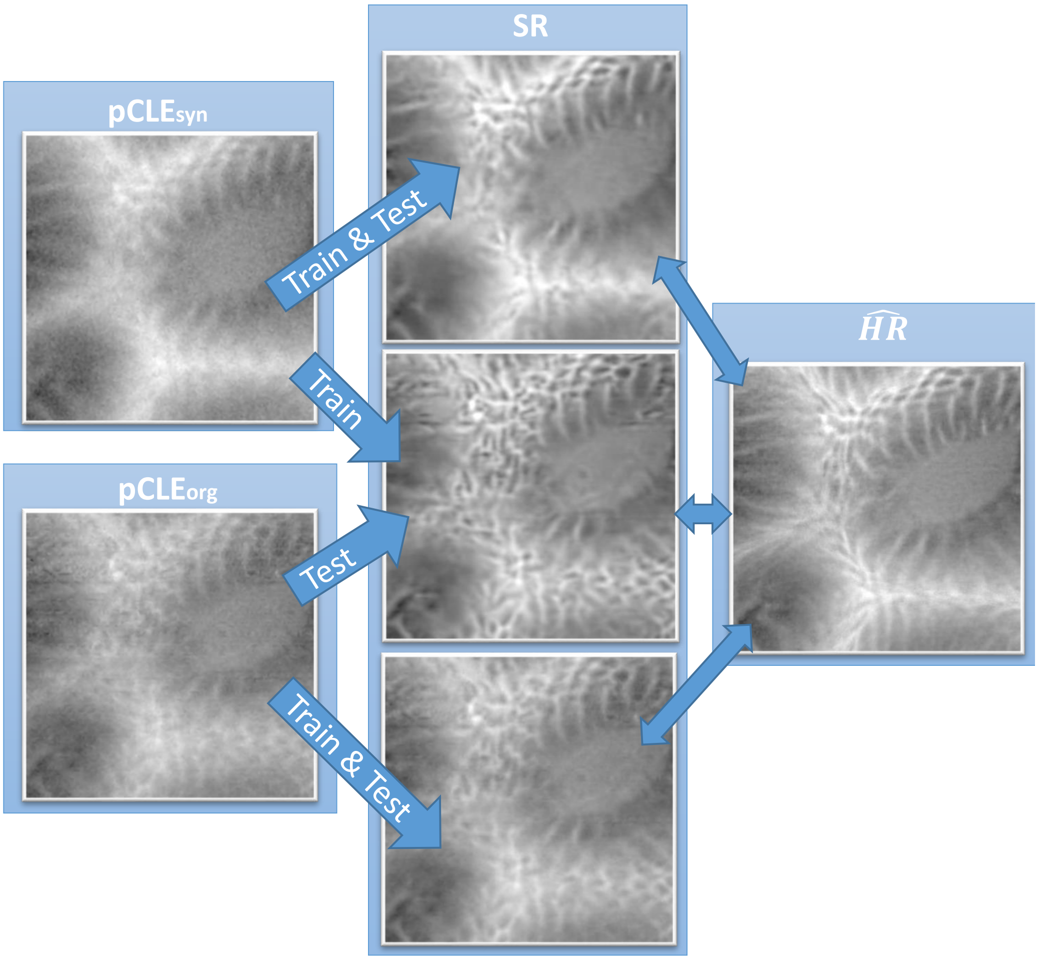


Figure 3.3: Example of SR images obtained when $pCLE_{syn}$ and $pCLE_{org}$ are used for train and test. From top to the bottom, the images in the middle represent the SR image obtained when: i) $pCLE_{syn}$ are used for train and test, ii) $pCLE_{syn}$ are used for train, and the $pCLE_{org}$ are used for test, and iii) $pCLE_{org}$ are used for train and test

significance of these improvements was assessed with a paired t-test (p-value less than 0.0001). From this experiment, it is possible to conclude that the proposed solution is capable of performing SR reconstruction when the models are trained on synthetic data with no domain gap at test time.

3.4.3 Experiments on original data

When real images are considered, the same conclusions is not reached. The results obtained by training on $pCLE_{org}$ and testing on LR_{org} are reported in the second section of Table 3.1 and here it is evident that all the different quality factors decrease. The best approach is the FSRCNN trained using SSIM+L1 as a loss function. With respect to the previous case, this approach loses 0.04 on the SSIM, and 0.12 on the Δ GCF with LR. This leads to a final reduction of 0.14 for the Tot_{cs} score. In this scenario, the deterioration of SSIM and GCF compared to the previous synthetic case can be due to the use of inadequate $\widehat{HR}$ images during the training (i.e. misalignment during the fusion, lack of compensation for motion deformations, etc.). Better results are instead obtained when the SR models performed on LR_{org} images are trained using the $pCLE_{syn}$ (last section of Table 3.1). Here, the quality factors increased when compared to the previous case, although they do not overcome the results obtained when the approach is trained and tested on synthetic data. EDSR, in particular, has a Tot_{cs} score of 0.65 that is 0.08 better than the best approach trained on $pCLE_{org}$ (the second section of Table 3.1) and 0.06 worse than the best approach trained and tested on $pCLE_{syn}$ (first section of Table 3.1). The GCF obtained here are in general much better when compared to the previous two cases.

An example of the visual results from the different training modalities is shown in Figure 3.3. In conclusion, our findings suggest that existing video-registration-based approaches are inadequate to serve as a ground truth for HR images in reference to original LR images. While EBSR approaches, such as the EDSR and FSRCNN when trained on synthetic data, can produce SR images that enhance the quality of the LR images.

To further validate our methodology, in Table 3.2 the results obtained by the best model of our approach (EDSR trained on synthetic data with SSIM + L1 as loss function) were compared against other state-of-the-art SISR methodologies. Specifically, in this experiment a Wiener deconvolution, a variational Bayesian inference approach with sparse and non-sparse priors [118], the SRGAN and EDSR networks pretrained on natural images were considered. The Wiener deconvolution was assumed to have a Gaussian point-spread function with the parameter σ=2 estimated experimentally from the training set. Finally, the last column of Table 3.2 includes the results of a contrast-enhancement approach obtained by sharpening the input with parameters similarly tuned on the trained set. Although our approach is not consistently outperforming the other on each quality score, when the combined

Table 3.2: Results of the proposed approach against state-of-the-art methods. The best results for each section are highlighted in bold.

metric method	SSIM with $\widehat{HR}$	GCF with $\widehat{HR}$	GCF with LR	Tot_{cs}
Proposed	0.80±0.06	0.33±0.31	0.57±0.36	**0.65**
Bayesian [118]	**0.81±0.06**	-0.26±0.37	-0.02±0.01	0.44
PreTrained SRGAN	0.79±0.06	-0.26±0.36	-0.01±0.01	0.40
PreTrained EDSR L1	**0.81±0.06**	-0.24±0.37	0.00±0.01	0.44
Wiener	0.77±0.07	-0.46±0.48	-0.22±0.24	0.28
Contrast-enhancement	0.65±0.09	**0.81±0.36**	**1.06±0.25**	0.50

score Tot_{cs} is considered, our method outperforms the others by a large margin.

3.4.4 Semi-quantitative analysis (MOS)

Due to our conclusions from previous sections, the MOS study was performed using images obtained from the models trained only with synthetic data. To perform the MOS, nine independent experts were asked to evaluate 46 randomly selected images each. Full-size LR_{org} were selected randomly from the test set of $pCLE_{org}$, and used to generate SR reconstructions. At each step, the SR images obtained by the three different methods (SRGAN, FSRCNN and EDSR) trained on synthetic data and a contrast-enhancement obtained by sharpening the input (used as a baseline) are shown to the user, in a randomly shuffled order. The input and the $\widehat{HR}$ are also displayed on the screen as references for the participants. For each of the four images, the user assigns a score between 1 (strongly disagree) to 5 (strongly agree) on three different questions:

- *Q1: Are there any artefacts/noise in the image?*
- *Q2: Can you see an improvement in contrast with respect to the input?*
- *Q3: Can you see an improvement in the details with respect to the input?*

To make sure that the questions were correctly interpreted, each participant received a short training before starting the study. The results on the MOS are shown in Figure 3.4. EDSR is the approach that achieves the best performance in Q2 and Q3. Instead, based on Q1, both FRSCNN and EDSR do not introduce a significant

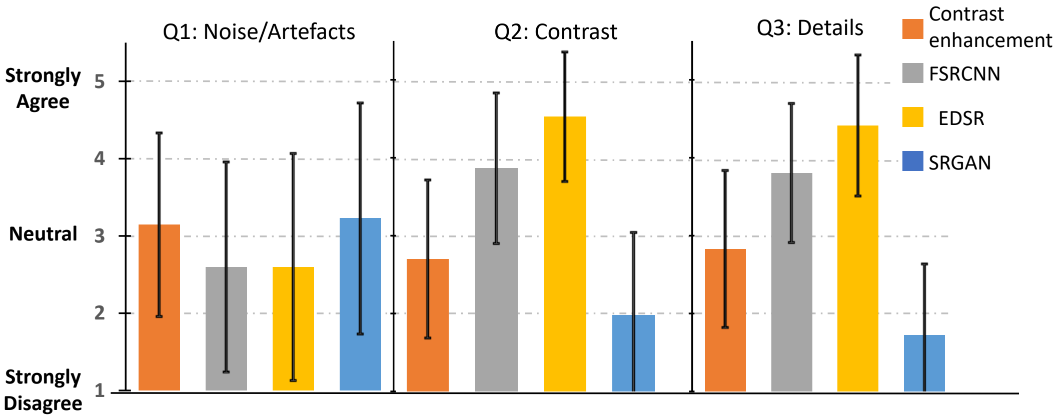


Figure 3.4: Results of the MOS using a contrast-enhancement approach, FSRCNN, EDSR and SRGAN. The plots report the results on the 3 different questions Q1-3.

amount of artefact or noise. The results of the MOS give us one more indication, which our training methodology allows improvements on the quality of the pCLE images. There is a high variance between individual responses of participants that leads to a high standard deviation of the scores. This is related to both the small number of participants and their scoring style. Also, images are very similar, which makes comparing them a hard task, and this impacts the variability of responses. MOS surveys tend to not be robust to those aspects [98].

Figure 3.5 shows a few examples of the obtained SR images using our proposed methodology. We also provide the movie [5] that shows generated pCLE sequence from five different configurations. All sequences were built from the validation set used for evaluating the performance of SR models we trained for pCLE super-resolution.

3.5 Discussion and conclusions

In this Chapter we addressed the challenge of SR for pCLE images. This is the first work to evaluate the potential of deep learning and EBSR in the pCLE context.

The main contribution of this work is to overcome the challenge of the lack of

[5] supplementary material available at `https://vimeo.com/426319061`

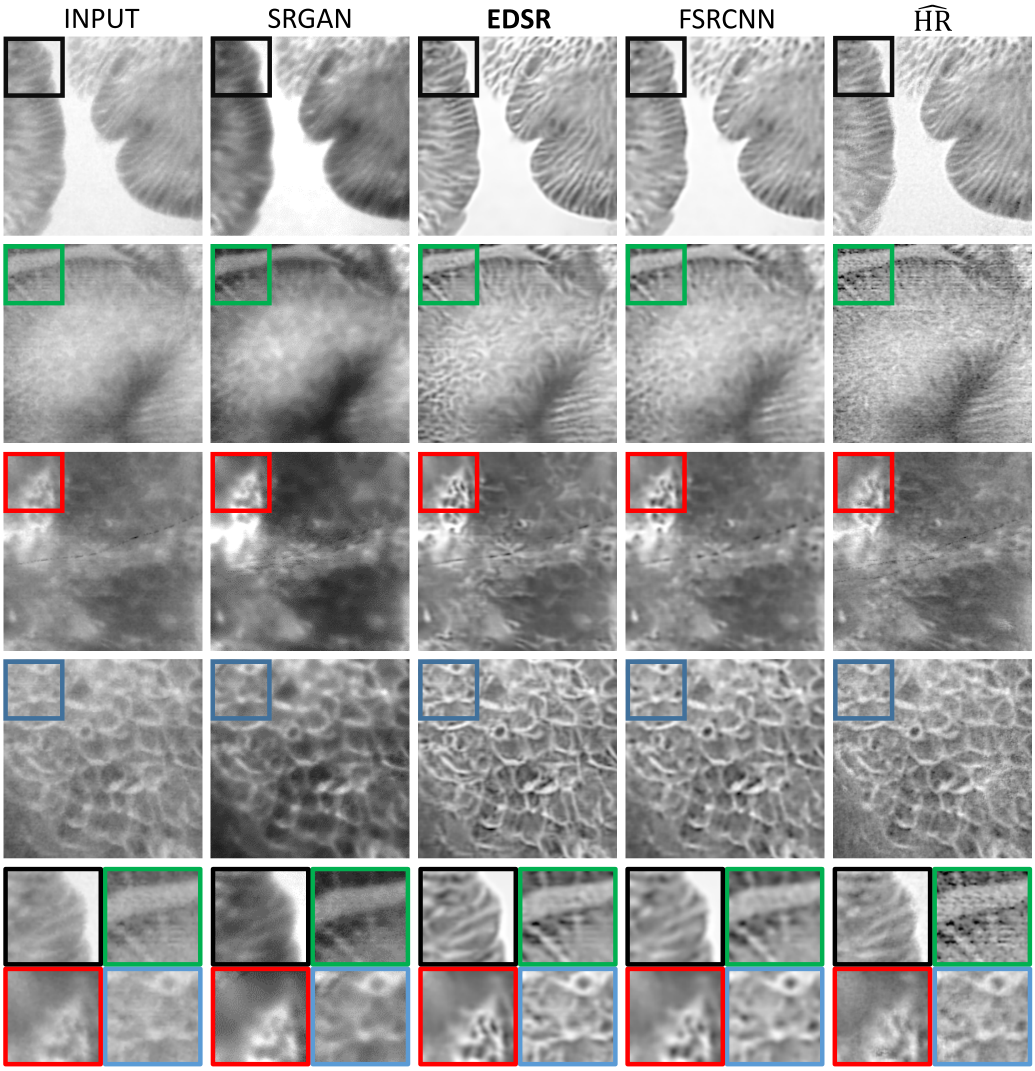


Figure 3.5: Example of visual results from the proposed approaches(from left to right): Input (left), SRGAN (middle left), EDSR (middle) and FSRCNN (middle right) $\widehat{HR}$ (right). Sequences are presented as follows: the top four rows show the full image, the two bottom rows show a zoomed crop from the top left corner of the image; each zoomed crop can be matched to the image by the colour of the border.

ground-truth data. A novel methodology to produce pseudo-ground-truth exploiting an existing video-registration method, and simulating realistic LR image based on a physical model of pCLE acquisition is proposed. The conclusions are that synthetic pCLE data can be used to train CNNs while applying them to real scenario data because of a physically inspired simulation process that reduces the domain gap between real and simulated images.

We examined two loss functions, L1 and SSIM+L1, for training purposes. As expected, the models trained with SSIM+L1 achieved better SSIM scores. However, images reconstructed for both loss functions are visually virtually indistinguishable from each other, which was confirmed by the MOS study and a Student T-test in the quantitative study. We speculate that there is not much difference between models optimised by these two loss functions. Additionally, SRGAN optimised with generative loss functions, performed poorly in comparison with models trained with L1 and SSIM+L1.

The robust IQA test based on the SSIM and GCF score confirmed the improvement of obtained results with respects to the input image. An analysis of perceptual quality of images with a MOS study recruiting nine independent pCLE experts showed that SR models give clinically interesting results. Experts perceived an improvement in the quality of the reconstructed images with respect to the input image, without noting a significant increase in the amount of noise and artefacts. We also learned that the way in which we setup the MOS survey is not the best choice to evaluate the image quality of very similar images when only a few participants do scoring. Thus, we need to reconsider the method for the next user study to achieve statistically significant results. The quantitative and semi-quantitative user perception analysis provided consistent conclusions.

Chapter 4

A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks

This Chapter is based on our research work **"Learning from Irregularly Sampled Data for Endomicroscopy Super-resolution: A Comparative Study of Sparse and Dense Approaches"** by Agnieszka Barbara Szczotka, Dzhoshkun Ismail Shakir, Daniele Ravì, Matthew J Clarkson, Stephen P Pereira, and Tom Vercauteren published in *International Journal of Computer Assisted Radiology and Surgery*, 15 (2020), pp. 1167-1175.

Contents

4.1 Motivation: reconstruction from irregularly sampled data with CNNs

In the previous Chapter, we have seen that end-to-end DL is pushing the boundaries of medical image computing in application to pCLE. We have shown that SR models can be effectively trained on a dataset of the synthetic pCLE reconstructions. We demonstrated that these synthetic data-driven models used to post-process original pCLE reconstruction improve their quality. Effectively, incorporating prior information about pCLE images into the DL model reduces the uncertainty introduced by the reconstruction process and enable higher quality reconstructions.

The models presented in the previous Chapter are the state-of-the-art DL SR techniques for natural image data, which rely on regularly sampled (Cartesian) images. To use these models directly on pCLE data, we reconstructed synthetic images as Cartesian images from irregular signals. This sampling irregularity comes from the inherent design of the pCLE probe, which relies on a coherent fibre bundle comprising many (>10k) cores that are irregularly distributed across the FoV. The nature of image acquisition through coherent fibre bundles constitutes a source of inherent limitations in pCLE, having a direct, negative impact on the image quality. The raw data that the pCLE devices produce, therefore remain challenging to use for both clinicians and computerised decision support systems.

Raw pCLE is distorted by a few artefacts, such as a honeycomb pattern, and so need to be corrected before reconstruction. During calibration and restoration, the raw image is transformed into a vector of corrected fibre signals and their locations in the space of the fibre FoV [28]. The irregular sampling domain of the signals can be accurately discretised as a set of locations in an over-sampled regular grid as a sparse image, and then interpolated.

Image reconstruction from sparse signals has been widely studied. We provided a detailed description of pCLE reconstruction from sparse signals currently implemented into clinical devices in Chapter 1 “Introduction”. As a brief reminder, existing pCLE image reconstruction approaches typically use Delaunay triangulation to linearly interpolate irregularly sampled signals onto a Cartesian grid [41].

These interpolation methods yield sharp images, but they are themselves prone to generating artefacts, such as triangle edge highlights or additional blur [41]. Another technique implemented specifically in the context of pCLE by Vercauteren et al. [41] is a reconstruction from scattered pCLE data with Nadaraya–Watson regression using handcrafted Gaussian kernels. The authors demonstrated that the method efficiently reconstructs pCLE images and mosaics, at the price of some additional blur in comparison with Delaunay reconstruction. Both reconstruction methods allow reconstruction of the Cartesian image, yet do not enhance image quality nor take into account any prior knowledge of the image space, except for regularisation related properties.

Although we have shown that DL has the potential to improve by post-processing the quality of the pCLE reconstructions, a potential limitation in the current CNN approaches is that the analysis starts from already reconstructed pCLE images, including reconstruction artefacts. While convolution layers are widely used, they have been identified as suboptimal for dealing with sparse data [173]. There are a few research works focusing on adapting the CNN framework to an unnatural fit for sparse inputs [183, 184, 173]. Generalising the conclusion from these studies to pCLE hints to the intuition that applying CNNs directly to irregularly sampled pCLE data may not be trivial. Hence, there is an unmet need for a unified, computationally-efficient, image reconstruction methodology that compensates for a range of limitations in pCLE, including the irregular sampling of raw pCLE signals.

4.2 An insight into an applications of CNNs to irregularly sampled data

Several approaches have been proposed to handle sparse data as input to CNN networks. Much of the available literature on exploring sparsity in the context of CNN input deals with the irregular data in an intuitive but ad hoc way: non-informative pixels are assigned zero, creating an artificial Cartesian image. For example, Li et al. [149] used that technique and assigned the missing points zeros on an LR image.

A similar workaround is to use an additional channel to encode the validity of each pixel like in Kohler et al. [125]; they passed a binary mask to the network. These solutions suffer from redundancy in image representation due to spurious data being fed to the convolutional layers.

Uhrig et al. [173] proposed a convolutional layer that jointly processes sparse images and sparse masks to achieve sparsity invariant CNNs. Their sparse layer is designed to account for missing data during the convolution operation by modelling the location of data points with the use of a mask. This is achieved by convolving the mask with a constant kernel of ones while optimising the solution through convolving the sparse image with trainable kernels.

Following the success of sparse CNNs, Hua, J. et al. [184] proposed to implement normalised convolution, as an extension of sparse convolution. They showed that using shared positive kernels for convolution with both an image and a mask is beneficial for upsampling depth maps. In both works, the information on sparsity is propagated to consecutive layers by the binary mask.

A demonstrated improvement in the proposed solutions is to use soft certainty maps, rather than propagating binary masks [183]. These maps are produced by updating the mask with convolution. This method worked well in a guided depth upsampling task and uses both RGB data and LiDAR to reconstruct depth maps.

These few research works [183, 184, 173] focusing on allowing sparse data as CNN input were introduce specifically for depth up-sampling task from irregularly sampled LiDAR data. The deep learning approaches have not yet been discussed for image reconstruction from irregular pCLE signals.

The main contribution presented in this Chapter is a suitability study of three DL approaches applied either directly to the irregularly sampled or reconstructed Cartesian images, respectively. The main focus of the work is to compare baseline pCLE image reconstructions obtained from classical interpolation methods to these dedicated DL approaches. To the best of our knowledge, it was at the time of publication the first work delivering a head-to-head comparison of sparse and Cartesian approaches for pCLE SISR.

The Second contribution of our study is a design of a novel trainable convolutional layer called an NW layer, which integrates Nadaraya–Watson (NW) kernel regression [2] into the DL framework, allowing principled handling of irregularly sampled data in the neural network. The proposed solution facilitates using sparse images as the input of the SR CNN directly, without the need for prior reconstruction, and also eliminating edge artefacts from input images. As far as we know, we are the first to propose, at the time of the publication, using NW kernel regression embedded in a CNN framework. We make use of it to design a network for medical image SR reconstruction from irregularly sampled pCLE signals.

We design this Chapter as an ablation study, that compares three ways of handling data sparsity for SR tasks with different methodological approaches. We study approaches as follows:

- *Cartesian*: processing reconstructed images with standard CNN. The model is not aware of the sparse nature of the signal and learns only how to super-resolve already reconstructed images. This approach is well-established in the DL community and justified in the previous Chapter on SISR for pCLE.

- *Sparse*: processing sparse images with standard CNN. The model has to learn both the SR task and the sparse representation of data needed to fill in the gaps in the implicit sampling pattern. This approach is a well-established solution in literature, yet has shown to be suboptimal.

- *Trainable Nadaraya–Watson regression*: processing sparse images with the sparsity mask, also explicitly used as input to a novel CNN layer. The model learns only SR since the explicit sparse representation is handled by the inherent design of the NW layer. This approach is a novel solution proposed as one of the contributions in this thesis.

The organisation of the Chapter is as follows: first in Section 4.3 “Data generation methodology for the comparative study”, we describe our data generation methodology of synthetic endomicroscopies for the evaluation of the comparative study. Then, in Section 4.4 “Irregularly sampled pCLE image reconstruction methods” we explain sparsity and reconstruction of these synthetic data. In Section 4.5 “Learning-based super-resolution approaches” we outline the methodology of the three compared approaches to handle data sparsity in the CNNs. We propose an experimental suite and results of our ablation study in Section 4.6 “Implementation details”. Finally, we conclude on the results of the ablation study in Section 4.8 “Discussion and conclusions”.

4.3 Data generation methodology for the comparative study

Since common IQA relies on ground truth images used as a reference in metrics such as PSNR, the lack of ground truth high-resolution pCLE images makes it difficult to evaluate and compare the quality of SR reconstructions.

To address the lack of the HR pCLE, in Chapter 3 “Single-image super-resolution for endomicroscopy”, we proposed to use a first-generation SR method—an offline mosaicking—to simulate HR endomicroscopy. We used mosaics as a source of HR content; unfortunately, these mosaics are not good enough estimate of HR images. The mosaicking resolves SR image from utilising overlap of video frames and therefore suffers from mis-registration artefacts, and not a uniform overlay of the frames, which cause nonuniform contribution of SR resolving capabilities of the mosaicking on the entire surface of the SR image. The mosaicking is also time-consuming, making it not applicable to the real-time workflow of pCLE.

In this work, similar to our previous solution, we used a triangulation-based reconstruction algorithm to simulate synthetic HR and LR endomicroscopy. However, in contrast, we took advantage of the availability of histopathological images as a source of HR signals instead of using imperfect mosaics. During the diagnostic process, histopathological images play a similar role to pCLE. Since histopathological images are acquired with a digital camera, histology does not suffer from the problems created by irregularly distributed fibre signals. Thus, histopathological images meet the criteria of HR signal source and serve the role of synthetic ground truth in our synthetic data set. The simulation pipeline is illustrated in Figure 4.1.

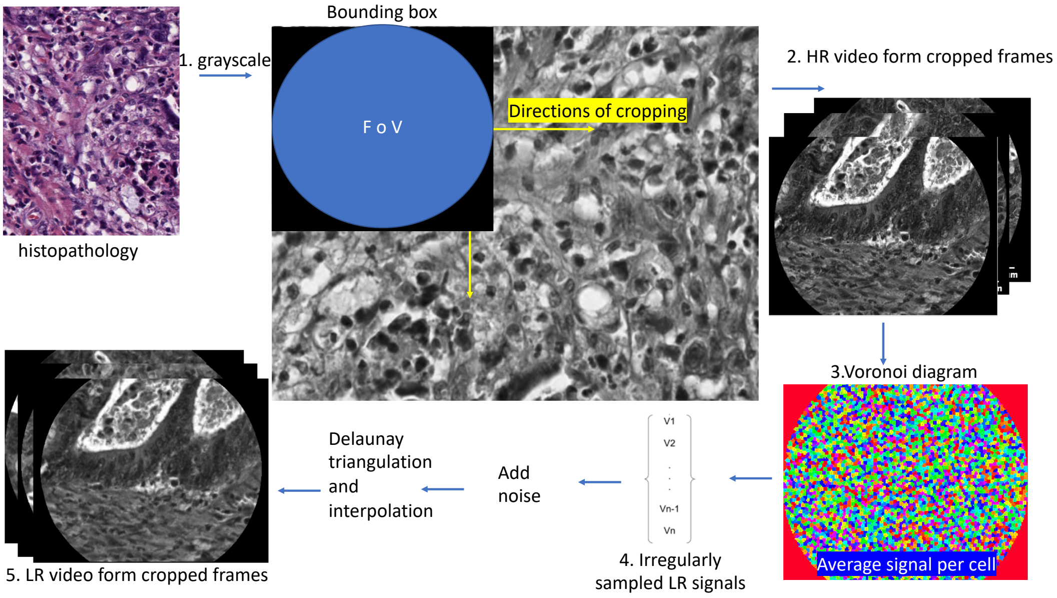

Bounding box
1. grayscale
F o V
Directions of cropping
2. HR video form cropped frames
histopathology
3.Voronoi diagram
Average signal per cell
V1
V2
Vn-1
Vn
4. Irregularly sampled LR signals
Add noise
Delaunay triangulation and interpolation
5. LR video form cropped frames

Figure 4.1: Illustration of the simulation for creation of synthetic data. Histological images are transformed to synthetic endomicroscopy.

Synthetic HR pCLE In the first part of the simulation, we transformed RGB HR histological images into greyscale HR pCLE-like videos. The simulation starts with transforming RGB images into greyscale images. Next, we randomly selected original pCLE videos from the Smart Atlas [82], retrieving information on a bundle FoV and fibres locations for each video, and we matched with the histological images. To crop pCLE-like frames from the greyscale image, we were moving a bounding box of the fibre's FoV from left to right, and from top to bottom in the image, with a step size equal to half of the bounding box size. These pCLE-like frames were stacked to create the synthetic HR pCLE video sequence.

Irregularly sampled synthetic LR pCLE To simulate LR pCLE videos, we used the physically inspired pCLE-specific downsampling presented in Chapter 3 "Single-image super-resolution for endomicroscopy". In our case, the sources of irregular signals for physically-inspired pCLE-specific downsampling are HR synthetic pCLE videos. They are rich in high frequencies and pixel-level details.

For every synthetic HR pCLE, we used the given associated fibres location to build Voronoi diagrams with each fibre in the centre of the Voronoi cell. Every cell corresponds to the one fibre signal, yet the cell space covers several pixels around the fibre on the HR image. Thus, to simulate signal loss, all HR pixels in that cell are averaged, and the average cell signal is used as a new LR fibre signal. The synthetic LR pCLE irregularly sampled images together with their associated given fibre position metadata may then be reconstructed as explained in Section 4.4 "Irregularly sampled pCLE image reconstruction methods" or used as input to a sparse super-resolution approach as detailed in Section 4.5 "Learning-based super-resolution approaches".

4.4 Irregularly sampled pCLE image reconstruction methods

Irregularly sampled data as sparse image Irregularly sampled data can be represented, with an arbitrary approximation quality, on a fine Cartesian grid as the sparse artificial Cartesian image S. Typically sparse image represents the sparse image space, where all non-informative pixels are set to 0. The pCLE sparse image is depicted in Figure 4.2; we can see fibre signals (informative pixels), surrounded by the non-informative space of zeros.

Input masks The information about the sampling grid is known from the nature of the acquisition. The position of the fibre signal in the Cartesian image corresponds to the position of the fibre within the bundle, which is given explicitly by a design of the fibre bundle and is provided with high accuracy by the manufacturer. We define pCLE mask M to be a representation of the position of the fibres in the Cartesian image space and represent it as a binary map. Intuitively, the image sparsity is encoded by ones and zeros for the informative and non-informative pixels, respectively, as a binary mask M of shape S. A sample M and its corresponding S is depicted in Figure 4.2.

Cartesian image reconstruction To reconstruct sparse images to Cartesian images, the missing information is typically interpolated. This also means that the reconstructed images are over-sampled, and only a subset of the pixels carry information [187].

Typically, pCLE noise is interpolated onto the reconstructed image from noisy signals. To achieve that, we simulate pCLE noise, by adding it to the new fibre signal, before the interpolation step. We add multiplicative and additive Gaussian noise to mimic a calibration imperfection and an acquisition noise, respectively.

Synthetic LR pixels with added noise, obtained in the last step of the data generation pipeline, are used as the irregularly sampled signals and reconstructed to the synthetic noisy Cartesian LR image using one of the baseline reconstruction approaches discussed hereafter. The synthetic LR pCLE has the size of the HR image, yet it is characterised by the lower image quality, noise, and reduced content of information due to simulated signal loss. Thanks to simulating signal distribution through the geometrical position of the fibres in the bundle, we simulate synthetic endomicroscopy as similar to real pCLE, characterised by typical triangulation artefact and noise patterns

The Cartesian images are reconstructed with our first baseline method, which is currently used in clinical practice; we refer to it as LINEAR BASELINE. This reference reconstruction method is based on linear interpolation and Delaunay triangulation [41]. We also provide a comparison to reconstructions obtained using NW kernel regression with a hand crafted single Gaussian kernel [53] as the second baseline method; we refer to it as GAUSS BASELINE. The results of the comparisons are presented in Section 4.7.1 “Evaluation on synthetic endomicroscopies”.

4.5 Learning-based super-resolution approaches

We exploit the irregularly sampled pCLE data for image reconstruction tasks with CNNs. We compare state-of-the-art dense and sparse approaches for handling image sparsity by CNNs in application to pCLE image reconstruction with classical pCLE reconstruction methods, including proposed by us novel trainable layer as a contender to established approaches.

Dense approach Reconstructed by Delaunay-based interpolation Cartesian image as input to the CNN, the reconstructed pCLE images can be treated as any natural image and input to the CNN directly. A convolutional layer $f_{u,v}$ is defined on the Cartesian grid as:

$$f_{u,v}(X) = \sum_{i,j=-k}^{k} x_{u+i,v+j} w_{i,j} \tag{4.1}$$

It considers all image X pixels as equally important regardless of their position u,v when convoluted with weights w.

Sparse approach A sparse image S is processed with a standard Cartesian CNN. S encapsulates both informative and non-informative pixels, and there is no explicit knowledge about the sampling pattern being fed to CNN. In that case, the network has to learn not only the mutual relations of informative pixels for the super-resolution task, but also handle the pixel sparsity. This dual role of the CNN kernels in the sparse approach was shown to lead to suboptimal results in the case of randomly sampled data [173].

NW layer Here the challenge is to adapt the CNN to work around the image sparsity by predicting the missing information. We propose the generalisation of NW kernel regression to pCLE image reconstruction from sparse data and their corresponding sparsity mask.

To incorporate NW kernel regression into the CNN framework, we propose a novel trainable CNN layer henceforth referred to as an "NW layer", which models the relation of the data points by the use of custom trainable kernels to perform local interpolation. This regression technique can be efficiently implemented using two convolutions and a pixel-wise division. We define the core NW operation as:

$$R_{u,v}(S,M) = \frac{\sum_{i,j=-k}^{k} S_{u+i,v+j} w_{i,j}}{\sum_{i,j=-k}^{k} M_{u+i,v+j} |w_{i,j}|} + b \tag{4.2}$$

$$M^{up} = \sum_{i,j=-k}^{k} M_{u+i,v+j} |w_{i,j}| \tag{4.3}$$

The NW layer takes as an input S and the corresponding known mask M. As it propagates through the network, the mask M can be seen as a probabilistic sparsity map. Initially $M \in \{0,1\}$, and the next M is updated as per Eq. (4.3) and becomes an approximation of the probability of obtaining reconstructions $R_{u,v}$ given $S_{u,v}$. The M^{up} holds the arbitrary probabilistic sparsity patterns, which are then propagated deeper to the consecutive NW layer. The outputs of the NW layer are reconstructed feature maps R estimated using an NW regression and updated probabilistic sparsity masks M. Finally, bias b is added to R. The graphical representation of the NW layer is presented in Figure 4.2.

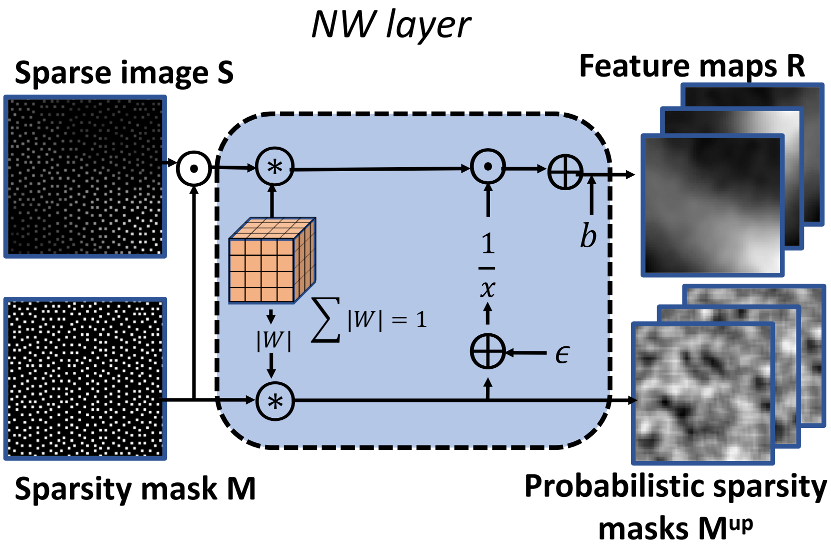


Figure 4.2: Graphical representation of NWNet layer. W is a kernel, b is a bias, $\odot$ is element-wise multiplication, $\oplus$ is addition, $\otimes$ is convolution, $\frac{1}{x}$ is element-wise division.

Classical NW kernel regression uses custom handcrafted, positive kernels. For generalisation of the kernel regression, however, our trainable NW layer allows for negative values. It is necessary for the convolution of the mask M to rely on the absolute value of W, as this operation is meant to capture the geometric influence of neighbouring pixels on the predicted values of $R_{u,v}$. For numerical stability, we also normalise the kernels such that $\sum_{i,j=-k}^{k} |w_{i,j}| = 1$, where w is a weight in position i, j in the W .

Thanks to shared kernels between the convolutions used in the NW operation, this layer has the same model complexity as the dense and sparse approaches. Model complexity is indeed measured by the number of parameters, or equivalently the number of kernels used. The only addition in terms of complexity is in terms of memory and computation: we keep track of the sparsity pattern and introduce a division of convolutions, which essentially is the key to enforcing Nadaraya–Watson regression in the CNN framework. The extra memory and computation requirement remain marginal in the context of our small 2D images, where IO often acts as a bottleneck in comparison to the GPU operation. The implementation used in this work was run on NVIDIA DGX-1 with 4 GPU cards Tesla V100 16GB; training of each network took around 31h, in total 175k steps with around 4 steps/sec. The test time per frame is on average 40 s on CPU, and 50 ms on GPU.

NWnet framework Multiple NW layers can be stacked to generalise and benefit NW kernel regression for irregularly sampled pCLE data. The trend in designing CNN architectures is that stacking many (deep) layers improves the performance of the network. This concept is related to the fact that deeper networks are capable of learning more complex mapping. This improvement of model performance was also demonstrated in application to SR, e.g. Dong et al. advanced their 3-layer architecture SRCNN [124] by increasing the number of layers in next-generation architecture called FSRCNN [140]. Given that the deeper CNN architecture has better performance, intuitively the same rule should apply to NWNet architectures: the deeper the NWNet, the better the generalisation. This assumption is based on the fact that the NW layer is built with CNN layers. The number of NW layers in the architecture is chosen experimentally, as it is common practice in the case of CNN layers. The choice of the depth of the architectures may be driven by factors, such as computational complexity, inference speed, amount of training data, and generalisation power. The details of the architecture designed in this work can be found in Section 4.6 “Implementation details” and Figure 4.5.

We assume that NWNet learns the sparsity of the input data, such that after a few NW layers, the output features map can be directed to classical CNN layers. This in turn facilitates end-to-end pipelines that can incorporate sparse inputs by combining NW layers with classical CNNs. As illustrated in Figure 4.3, we show how to combine NW layers into a deep(er) network. Each NW layer has t unique

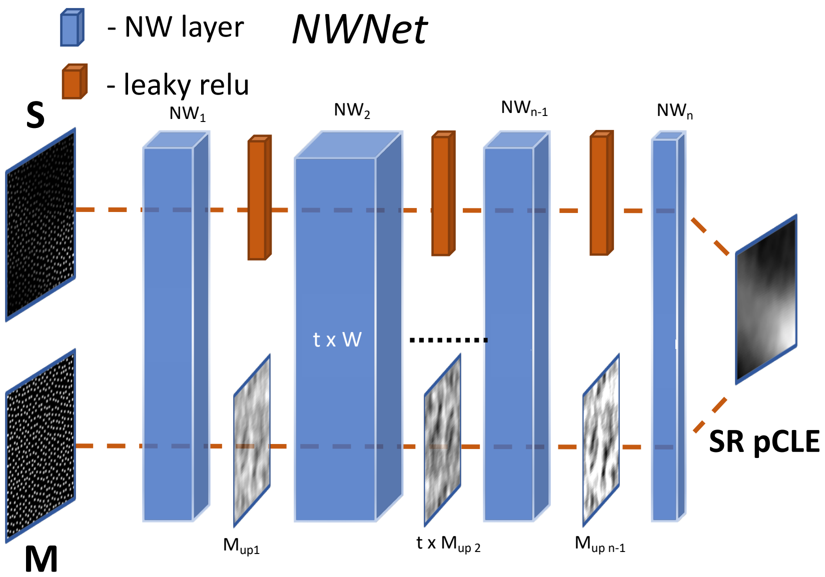


Figure 4.3: Graphical representation of NWNet framework. Each NW layer has t unique kernels W. The first layer NW_1 takes as input S and binary M and returns t feature maps R and t updated sparsity masks M_{up} which become the input for the next NW layer. The last NW layer of the NWNet framework returns only one feature map R, and masks M are discarded.

kernels W. The first layer NW_1 takes as input S and binary M and returns t feature maps R and t updated sparsity masks M_{up} which become the input for the next NW layer. The last NW layer of the NWNet framework returns only one feature map R, and masks M are discarded. NWNet in combination with complex deep learning models, such as EDSR [163], may allow for reconstruction of higher quality pCLE than the more typically used interpolation.

4.6 Implementation details

Data collection and pre-processing Synthetic videos were created using the methodology presented in Section 4.3 “Data generation methodology for the comparative study” from high-quality, large histopathological images. We created sets: a training set built with 540 files and a validation set built with 227 files from publicly available histological image data sets [143, 156]. The datasets contain samples showing tissue with abnormality such as colon cancer and breast cancer prepared with haematoxylin & eosin (H&E) stain.

The synthetic test set was created with ten histopathologies from publicly available data called Kather published in [146]. We used 10 large view H&E stained histological images of human colorectal cancer (CRC) and normal tissue, which were built with non-overlapping image patches, each 224x224 pixels (px) at 0.5 microns per pixel in RGB. The collections contain image samples with tissue classes such as adipose, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, colorectal adenocarcinoma epithelium. The samples with tissue slides were taken from NCT Biobank (National Center for Tumor Diseases, Heidelberg, Germany) and the UMM pathology archive (University Medical Center Mannheim, Mannheim, Germany). The synthetic test set facilitates making a comparative study between baseline solution, our proposed methodology, and classical CNN solutions.

To facilitate training, video sequences were normalised for each frame individually by subtracting mean and standard deviation of the LR frame as follows: $LR = (LR - mean_{LR})/std_{LR}$ and $HR = (HR - mean_{LR})/std_{LR}$, and then scaled to the range [0,1]. Synthetic LR images were transformed to the synthetic sparse LR images by setting zeros to all pixels which do not correspond to any fibre signal. The masks were generated as a binary image, where fibre positions from the LR image are set to ones. Lastly, to perform batch-based training, we extracted non-overlapping 64×64 sparse and Cartesian patches for the train and validation sets. The test sets were built with sparse and Cartesian full-size synthetic pCLE images.

Implementation details It was shown in Chapter 3 “Single-image super-resolution for endomicroscopy” that SISR improves the quality of over-sampled Cartesian pCLE images. Inspired by the EDSR network [163], the best performing architecture for our study on SISR, we design two architectures CNNnetSR and NWnetSR presented in Figure 4.5. The architectures were implemented in Python 3.6 using TensorFlow and Python Scientific Package. CNNnetSR was designed to perform the SR task from Cartesian or sparse LR pCLE images. The network is very similar to the EDSR architecture [163], but with two small improvements. First, we do not use upsampling layer, because synthetic LR and HR have the same size. In place of upsampling layer, we put a convolution layer with 32 filters. Second, the last convolutional layer aims at fusing the output feature maps from the penultimate layer into the Cartesian image. We find it more beneficial to use a kernel of size one, which is commonly used to reduce the number of features maps, than three. All convolutional layers have a kernel size of 3. The last layer uses a linear activation function.

We want to take advantage of SISR for NW kernel regression, so we design NWnetSR based on CNNnetSR by partly replacing standard convolution with NW layers. For the first NW layer, the kernel size is 9 across each image dimension. The size was chosen based on the known distribution of fibres across a Cartesian image to ensure that each convolution would capture more than 10 informative pixels. For deeper NW layers kernel size is 3. The NW weights were initialised with a truncated normal distribution with mean and standard deviation equal to 0.2 and 0.05, respectively.

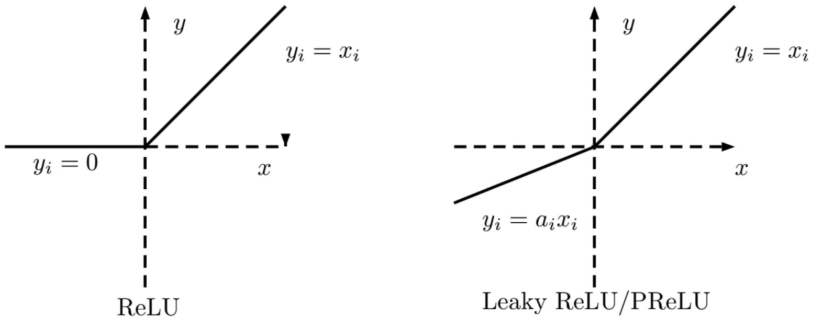


Figure 4.4: Relu (left) vs leaky-relu (right).

We also use Leaky Rectified Linear Unit aka leaky relu [115] as activation for NW layers. This choice is motivated by the fact that leaky relu replaces the zero part of the negative domain by a low slope allowing constant zero gradients to contribute to learning, as depicted in Figure 4.4.

Training strategy To achieve the best training results for each model individually, networks were trained with Population-Based Training (PBT) [162]. PBT is an optimisation technique design to find the best training parameters for the network. During PBT training, a population of models with different parameters is trained; these models are periodically validated and the weights from the best performing model in the population are copied to other members of the population. The PBT

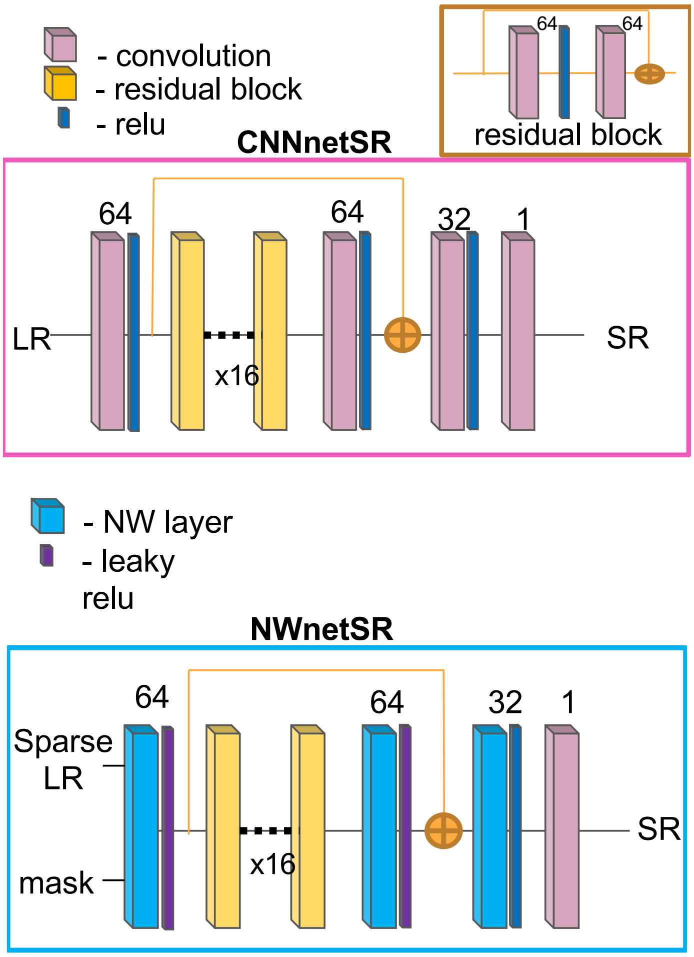


Figure 4.5: Reconstruction architectures: NWnetSR and CNNnetSR

was implemented with Ray[1]. We set the population size to 6 workers, where each member of the population uses a single GPU optimising a network for 10 epochs in one PBT iteration for a total of 100 iterations. The perturbation interval was set for every 20 iterations. The hyperparameter search applied to the 6 learning rates, which were initially set to 10^i where $i \in \{-2, -3, ..., -7\}$. We used the Adam optimiser and set β_1=0.9, β_2=0.999, $\varepsilon = 10^{-8}$. Based on results presented Chapter 3 "Single-image super-resolution for endomicroscopy", the models were trained with SSIM+L1 loss [175]. Finally, the best performing model from the population is used to generate results on the test sets described in Section 4.3 "Data generation methodology for the comparative study".

[1] `https://docs.ray.io/en/latest/index.html`

Experiments Our baseline methods are: linear interpolation based on Delaunay triangulation referred to LINEAR BASELINE, and NW kernel regression with a crafted single Gaussian kernel referred to GAUSS BASELINE.

We compared the performance of the DL models: CARTESIAN for CNNnetSR trained with Cartesian input, SPARSE for CNNnetSR trained with sparse input, and NWNET for NWnetSR trained with sparse input and corresponding mask. These models were handling irregular signals as input differently.

To quantify how the standard convolution performs on sparse pCLE images in comparison to using Cartesian reconstructions as the input, we trained two unique models based on the CNNnetSR network: CARTESIAN trained using reconstructed Cartesian images and SPARSE one trained with sparse images.

To test NW kernel regression benefits from generalisation via learning multiple kernels, we trained the NWnetSR as an EBSR network for the task of pCLE SR reconstruction with sparse input images and masks.

4.7 Results

4.7.1 Evaluation on synthetic endomicroscopies

The final performance of the models is evaluated by comparing the quality of the reconstructed SR pCLE with the HR synthetic images from the test set. To measure the image quality of the SR pCLE reconstructions, we design an IQA procedure which consists of two complementary metrics typically used for this task: PSNR and SSIM [31].

The summary results computed for the images from the test set, described in Section 4.3 “Data generation methodology for the comparative study”, are shown in Table 4.1. We also present detailed error analysis for PSNR in Figure 4.6 and Figure 4.7 and SSIM Figure 4.8 and Figure 4.9. Our error analysis includes histograms showing a distribution of the SSIM and PSNR scores for all frames in our test dataset, and a univariate distribution of observations to facilitate a comparison of the examined models.

The main observation is that each DL model outperforms the baseline interpolation technique. Training of the NWnetSR as SISR network generalises NW kernel regression and outperforms traditional NW regression with handcrafted Gaussian kernel, proving that the use of many kernels which are estimated on the data set is more beneficial than custom single Gaussian kernel. There is no significant improvement for PSNR and SSIM scores when comparing DL models between each other. They perform almost indistinguishably.

Table 4.1: IQA: comparison on average performance of 5 methods aimed at pCLE image reconstructions. Table shows average PSNR and SSIM score with standard deviation for all frames in 10 pseudo pCLE videos and all frames in total.

method	GAUSS BASLINE		LINEAR BASELINE		CARTESIAN		SPASRE		NWNET	
metric / video	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	26.6 ± 1.1	0.815 ± 0.024	25.1 ± 0.9	0.84 ± 0.015	32.3 ± 1.1	0.926 ± 0.010	32.4 ± 1.1	0.928 ± 0.010	32.3 ± 1.1	0.927 ± 0.010
2	25.1 ± 1.5	0.732 ± 0.026	24.2 ± 1.2	0.79 ± 0.018	30.2 ± 1.2	0.895 ± 0.016	30.2 ± 1.2	0.894 ± 0.016	30.1 ± 1.2	0.893 ± 0.016
3	24.4 ± 0.7	0.709 ± 0.019	23.6 ± 0.7	0.77 ± 0.017	29.2 ± 0.6	0.882 ± 0.015	29.2 ± 0.6	0.880 ± 0.014	29.2 ± 0.6	0.879 ± 0.014
4	28.6 ± 2.0	0.801 ± 0.035	25.7 ± 1.8	0.82 ± 0.029	33.6 ± 1.8	0.907 ± 0.023	33.4 ± 1.5	0.905 ± 0.023	33.5 ± 1.6	0.905 ± 0.023
5	24.5 ± 1.0	0.705 ± 0.024	24.7 ± 1.0	0.77 ± 0.018	29.4 ± 0.8	0.884 ± 0.016	29.5 ± 0.8	0.882 ± 0.016	29.5 ± 0.8	0.882 ± 0.016
6	26.9 ± 1.5	0.775 ± 0.028	26.2 ± 1.8	0.82 ± 0.020	31.9 ± 1.3	0.902 ± 0.016	31.8 ± 1.2	0.900 ± 0.016	31.8 ± 1.2	0.900 ± 0.016
7	25.7 ± 1.1	0.764 ± 0.027	24.2 ± 1.1	0.79 ± 0.017	31.1 ± 1.1	0.899 ± 0.013	31.1 ± 1.0	0.898 ± 0.013	31.1 ± 1.1	0.897 ± 0.013
8	25.7 ± 1.4	0.774 ± 0.034	24.1 ± 1.1	0.81 ± 0.020	31.1 ± 1.5	0.910 ± 0.015	31.1 ± 1.4	0.909 ± 0.015	31.0 ± 1.4	0.908 ± 0.015
9	24.9 ± 0.7	0.743 ± 0.015	24.0 ± 1.0	0.80 ± 0.011	30.0 ± 0.7	0.902 ± 0.010	30.1 ± 0.6	0.901 ± 0.010	30.0 ± 0.6	0.900 ± 0.010
10	25.9 ± 0.8	0.761 ± 0.020	25.4 ± 0.8	0.81 ± 0.014	31.2 ± 0.8	0.910 ± 0.012	31.3 ± 0.7	0.909 ± 0.012	31.2 ± 0.7	0.908 ± 0.012
all	25.9 ± 1.0	0.758 ± 0.025	24.7 ± 1.2	0.80 ± 0.018	31.0 ± 1.1	0.902 ± 0.015	31.0 ± 1.0	0.907 ± 0.016	31.0 ± 1.1	0.907 ± 0.015

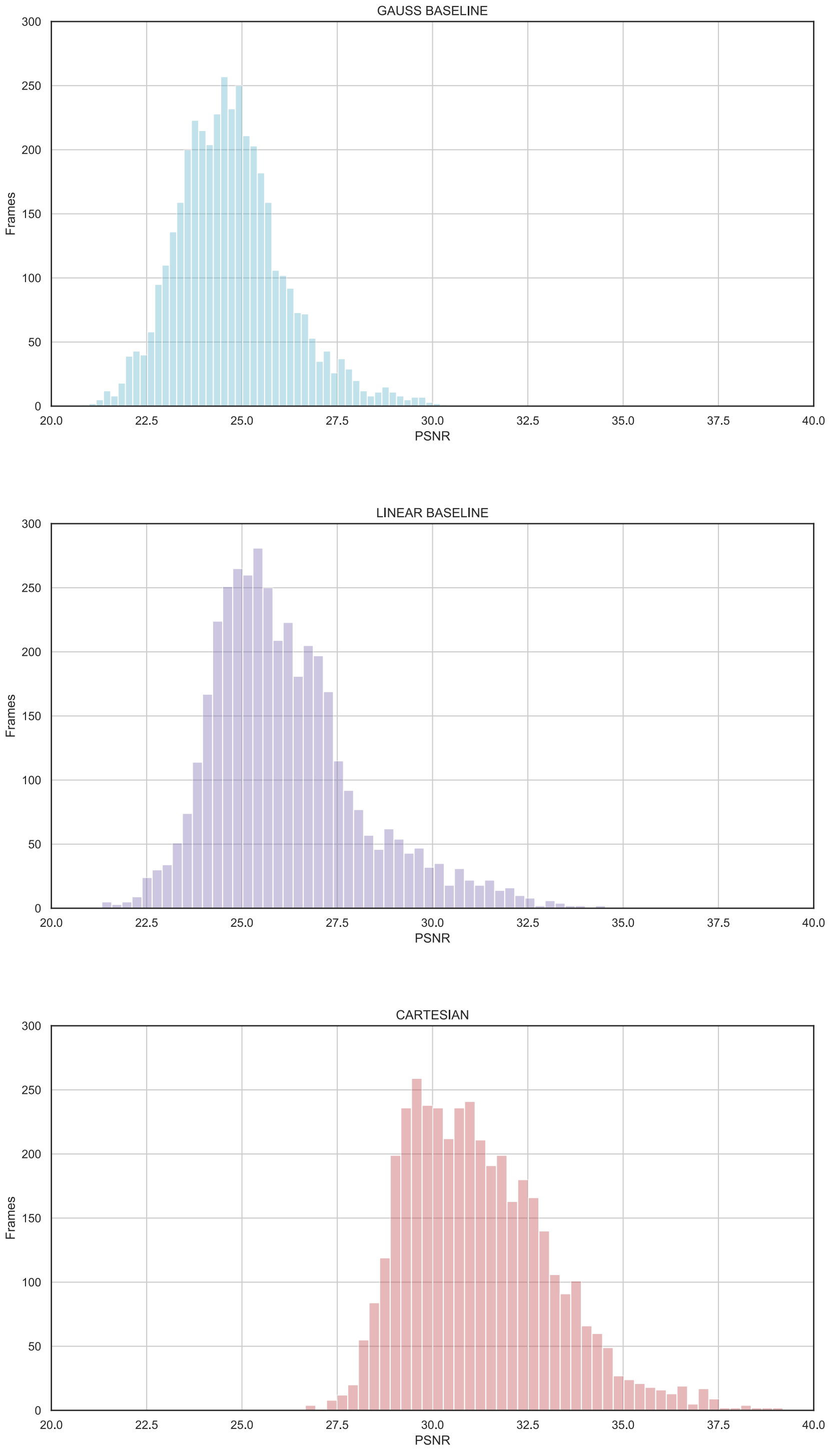


Figure 4.6: PSNR error analysis for the test set. Histograms of PSNR scores for baselines and CARTESIAN model.

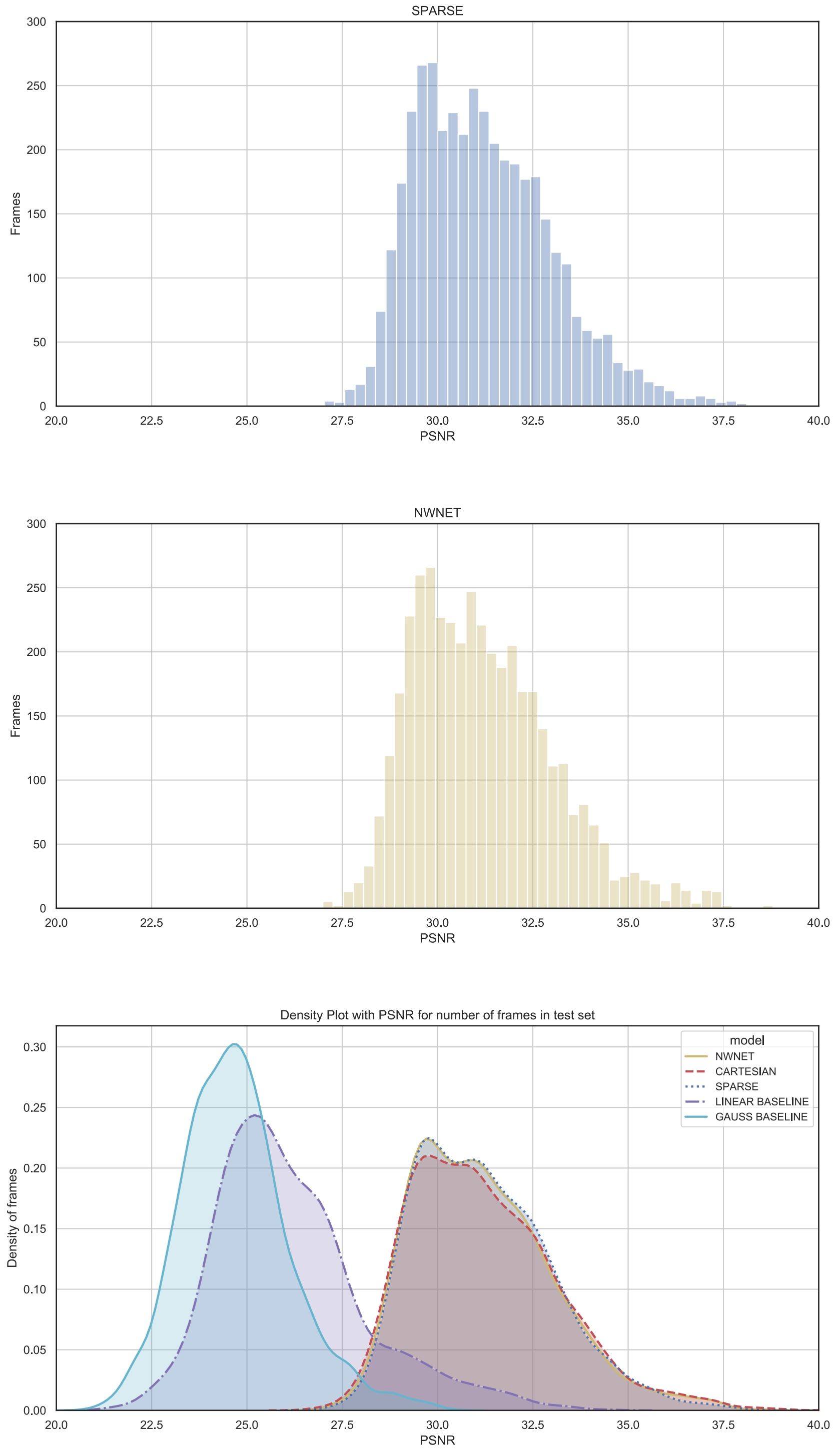


Figure 4.7: PSNR error analysis for the test set. Histograms of PSNR scores for SPARSE and NWNET model; and the normalised density plot comparing all models together.

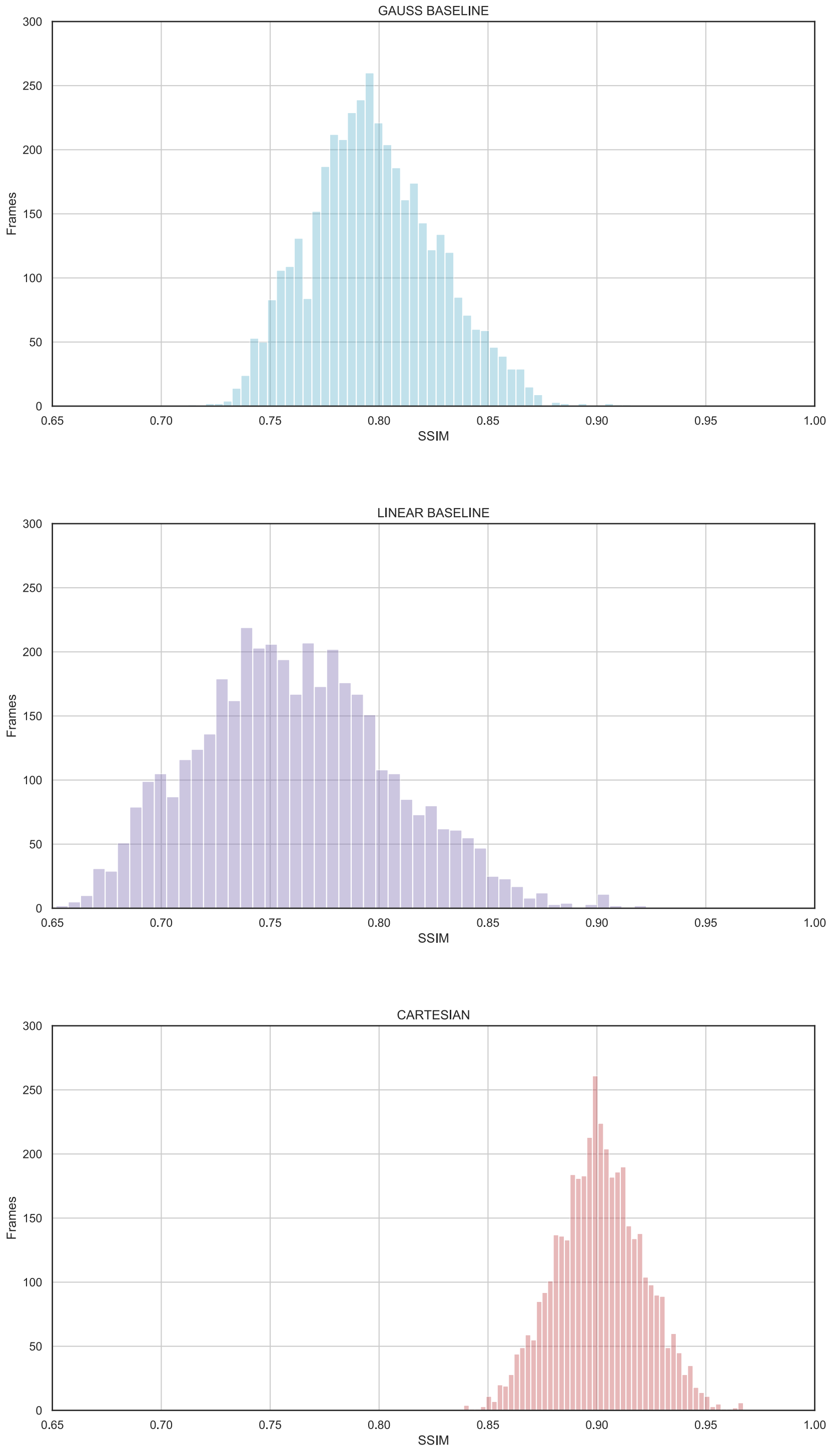


Figure 4.8: SSIM error analysis for the test set. Histograms of SSIM scores for baselines and CARTESIAN model.

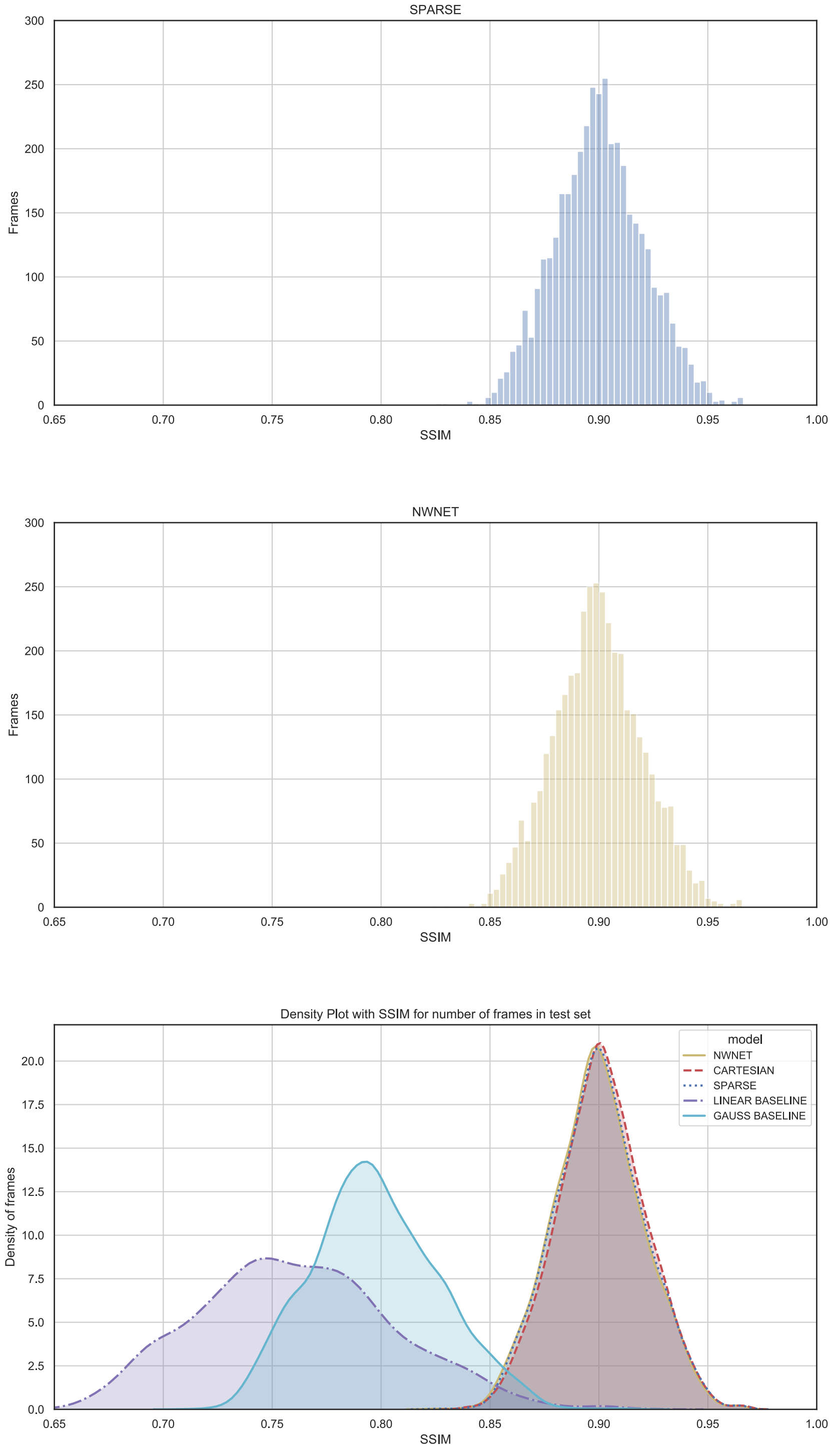
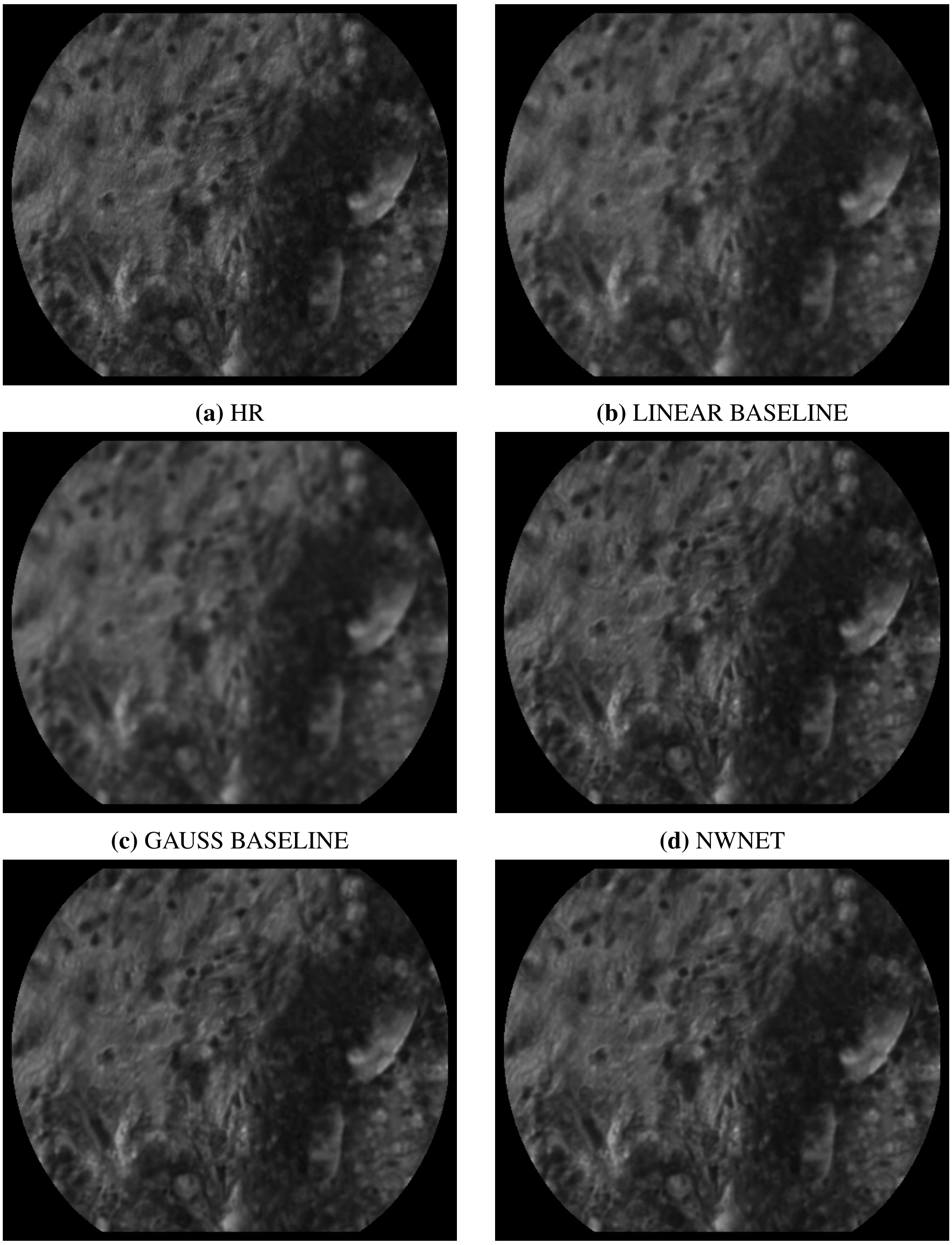


Figure 4.9: SSIM error analysis for the test set. Histograms of SSIM scores for SPARSE and NWNET model; and the normalised density plot comparing all models together.

We also provide example reconstructions which are the best and worst cases, and the one which has shown the biggest and smallest improvement based on these metric scores in Figure 4.10, Figure 4.11, Figure 4.12, and Figure 4.13. Our team of four researchers, specialised in medical imaging, analysed qualitatively SR reconstructions. The analysis was performed based on team experience with pCLE images by performing an informal visual inspection of the test set. We have chosen to not run any formal qualitative test on synthetic images, as we observed that SR reconstructions differ slightly on a pixel level, but that differences do not affect how the image is perceived as a whole, and maybe a reason for slightly different metrics score during IQA. In the opinion of our clinical collaborators, SR reconstructions are clearly more aesthetically pleasing than baseline reconstructions LINEAR BASELINE and GAUSS BASELINE. Neither of SR has triangulation artefacts, and every SR reconstruction has significantly reduced noise, additionally benefiting from improved contrast and visibility of details.

The results confirm that the NW layer among other layers handling sparse data is a suitable choice for image reconstruction and yields increasingly good image-quality results from sparse pCLE data to Cartesian SR image.



(a) HR

(b) LINEAR BASELINE

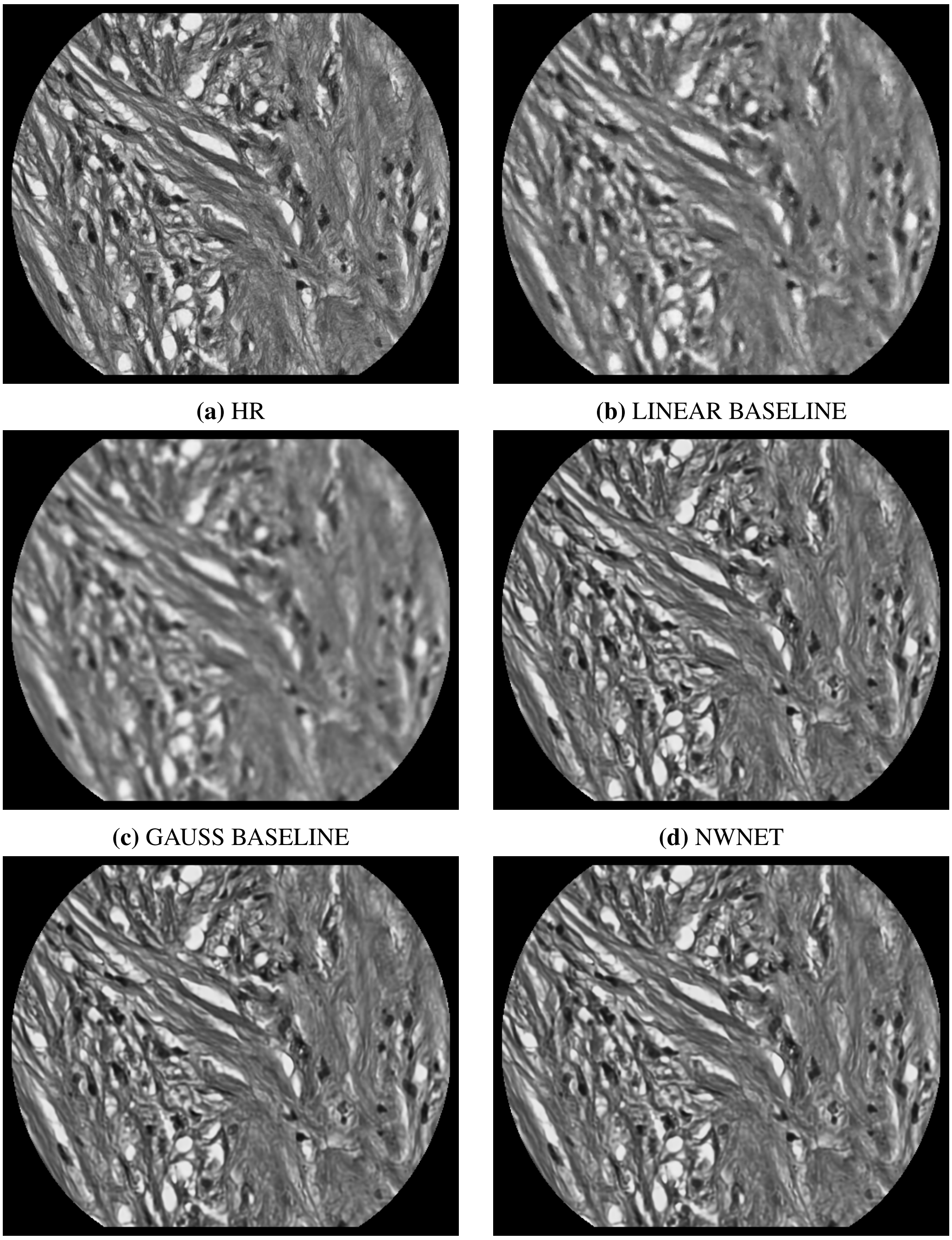
(c) GAUSS BASELINE

(d) NWNET

(e) CARTESIAN

(f) SPARSE

Figure 4.10: Frames with the highest PSNR and SSIM score in the test set for each model. These frames also represent the biggest improvement in PSNR score in reference to the LINEAR BASELINE.



(a) HR

(b) LINEAR BASELINE

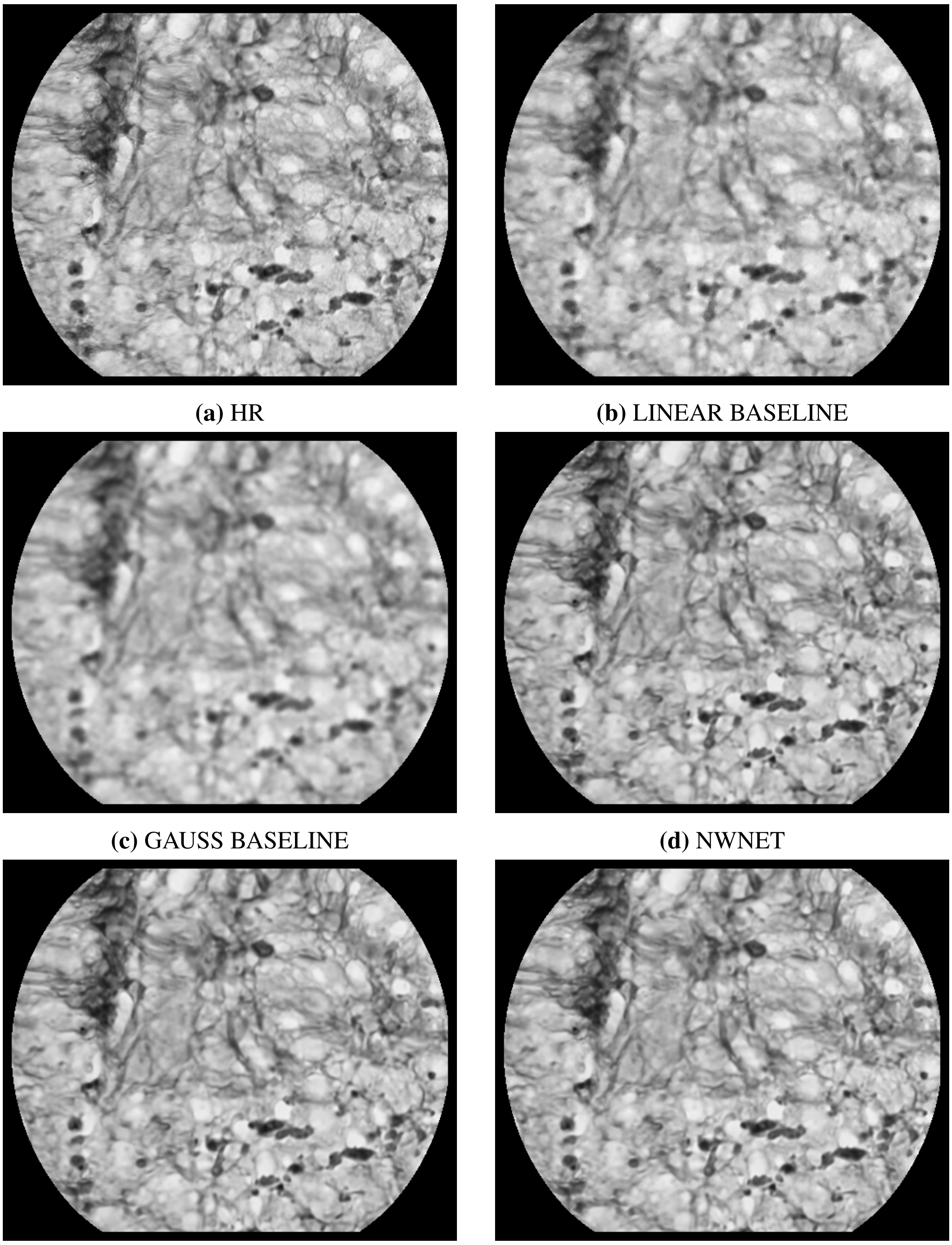
(c) GAUSS BASELINE

(d) NWNET

(e) CARTESIAN

(f) SPARSE

Figure 4.11: Frames with the smallest improvement in PSNR score in reference to the LINEAR BASELINE in the test set for each model.



(a) HR

(b) LINEAR BASELINE

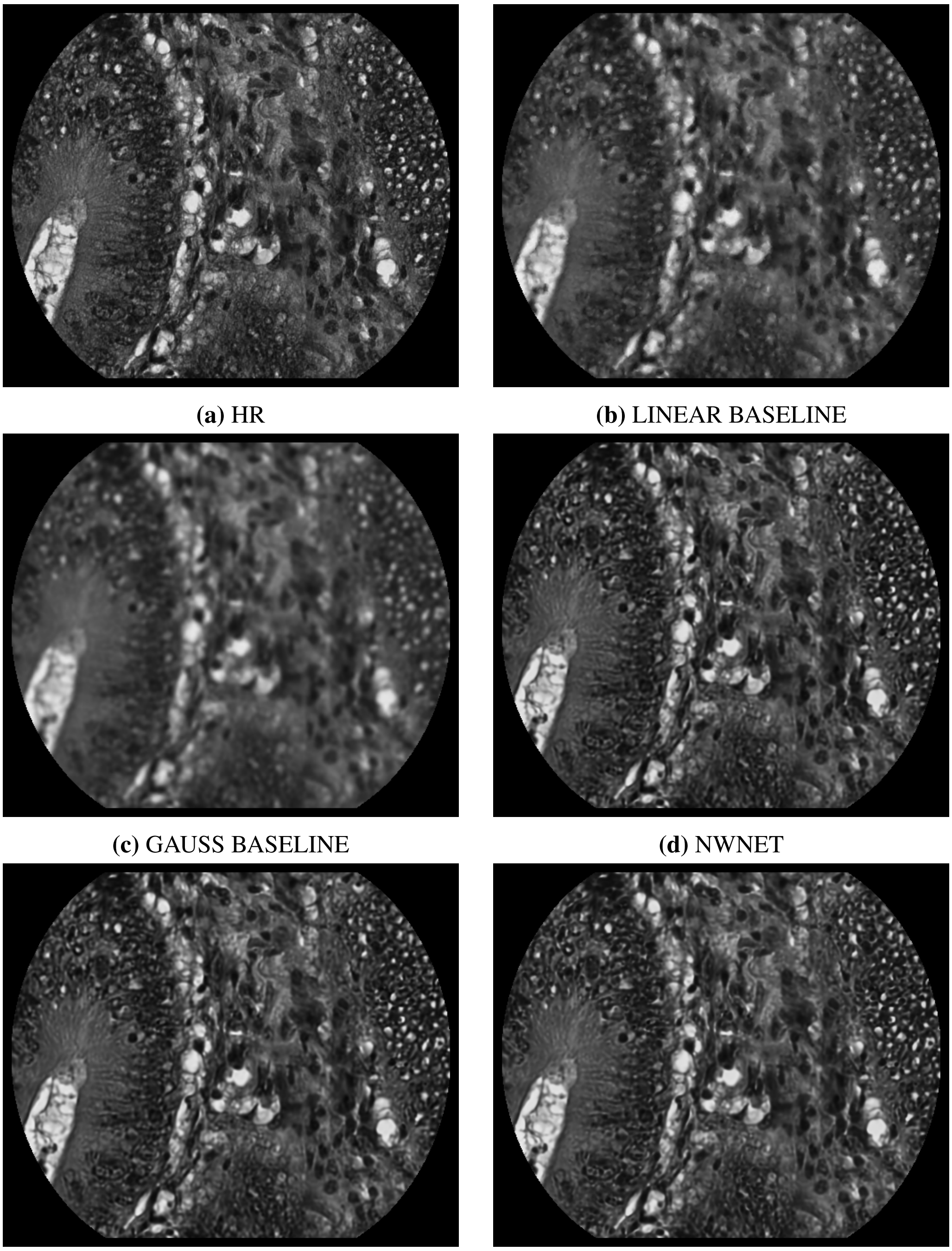
(c) GAUSS BASELINE

(d) NWNET

(e) CARTESIAN

(f) SPARSE

Figure 4.12: Frames with the biggest improvement in SSIM score in reference to the LINEAR BASELINE in the test set for each model.



(a) HR

(b) LINEAR BASELINE

(c) GAUSS BASELINE

(d) NWNET

(e) CARTESIAN

(f) SPARSE

Figure 4.13: Frames with the smallest improvement in SSIM score in reference to LINEAR BASELINE in test set for each model.

4.7.2 Application to original endomicroscopies

Despite the absence of ground truth when working with real data, visualising results on real data is very informative. We generated five videos[2] created from the original images. Videos display deep learning-based SR and the original reconstruction. Noting that ground truth HR images are not available in the context of real pCLE data, for this qualitative comparison, our deep learning models were trained on the synthetic data. We then used these models to generate the results for the real pCLE data.

These exemplar results illustrate that our research, even when trained on synthetic data, generalises well to real pCLE data. Similarly to the synthetic case, the three deep learning solutions all perform well on real pCLE data, providing the expected improvement in resolution and image quality. Also, as in the synthetic case, all deep learning solutions visually outperform the state-of-the-art baselines linear interpolation reconstruction method (aka LINEAR BASELINE) and single layer NW kernel regression with handcrafted Gaussian kernel (aka GAUSS BASELINE).

In Chapter 3 “Single-image super-resolution for endomicroscopy”, we have shown that Cartesian DL SR outperforms state-of-the-art baseline referred to as LINEAR BASELINE. We demonstrated it with both quantitative analysis and through a human study run as a Mean Opinion Score survey. Thus, we have already shown that the Cartesian DL SR performs better than the baseline reconstruction. In this comparative study, as we showed that Cartesian and Sparse learning perform equally well in terms of quantitative analysis, and building on experience from previous work on SISR, we believe that a user study essentially will give the same conclusions as our quantitative results, which clearly indicates the benefit of deep-learning approaches over the state-of-the-art baselines.

[2] available in Chapter 7 “Appendix - video reconstructions”, Figure 7.2

4.8 Discussion and conclusions

In this research work, we focused on providing a comparative study of learning-based approaches for pCLE SR. In the context of pCLE, this is the first work that proposes end-to-end deep learning image reconstructions from irregularly sampled fibre data and provides a head-to-head comparison of sparse and Cartesian approaches for this task.

We proposed the NW layer which enables the use of sparse images as input and performs deep generalised Nadaraya– Watson kernel regression. NWnet uses multiple stacked learnt kernels in contrast to the classical single hand-crafted Gaussian kernels in Nadaraya-Watson regression (aka GAUSS BASELINE). NWnet significantly outperforms its classical version. Up to our knowledge, we are the first to implement generalised NW regression into a DL framework and apply it to image reconstruction of pCLE. We have shown the successful implementation of the reconstruction pipeline, which combines NW and CNN layers and is trained in a supervised manner as SISR, reconstructing super-resolved images from sparse input images. Our results indicate that the idea of trained kernels by NWnet framework is effective for the purpose of learning from irregularly sampled data.

We demonstrated that learning-based methods reconstruct pCLE images with improved image quality. Our results show that these methods outperform baseline solutions. Yet, no significant differences between deep methods were found. Although somewhat negative, we believe this result is important because of its counter-intuitive nature.

The observation that all DL approaches perform equally well with no statistically significant difference may mean that standard deep CNNs are powerful enough to solve the reconstruction of sparse pCLE images regardless of their sparsity. A potential reason for it may lie in the fact that fibres in the pCLE bundle are distributed in a pseudo-regular pattern (quasi-hexagonal). This pseudo-regularity is probably key to steer the Cartesian and sparse model towards a suitable solution.

Learning methodologies for handling sparse input data in a similar manner to NWnet have been proposed for the domain of depth upsampling from LiDAR data which provides sparse clouds of points. Previous related studies on randomly distributed LiDAR data, which do not display the same pseudo-regularity pattern, showed that a sparse approach that explicitly accounts for data sparsity is helpful. Although accounting for data sparsity benefits depth upsampling and improves the results in comparison with standard Cartesian CNN, in our case, that conclusion did not hold true. In contrast to published works on depth upsampling we showed that the importance of explicitly using this mask is less influential than for depth upsampling applications. In the case of LiDAR data, the cloud of points is distributed randomly and does not exhibit any pseudo-regularity, which may be a more important difficulty for a Cartesian CNN. The sparse learning problem may indeed be more complex in such cases than in the case of pCLE data, where the pCLE fibre signals are distributed in the pseudo-regular pattern due to the geometrical constraint of packing fibres into a bundle.

Chapter 5

Zero-shot super-resolution for endomicroscopy

This Chapter is based on our research work **"Zero-shot super-resolution with a physically-motivated downsampling kernel for endomicroscopy."** by Agnieszka Barbara Szczotka, Dzhoshkun Ismail Shakir, Matthew J. Clarkson, Stephen P. Pereira, and Tom Vercauteren published in *IEEE Transactions on Medical Imaging 2021.*

Contents

5.1 Motivation: blind SR with a physically-motivated downsampling kernel

In Chapter 3 “Single-image super-resolution for endomicroscopy”, it was shown that DL-based SR models give promising results in improving the pCLE image quality. Several supervised models were trained on simulated pCLE data. HR pCLE images were estimated with an offline registration and mosaicking method; synthetic LR images were created from the HR images with the use of a physically inspired simulation algorithm. This work has shown that DL models trained on such synthetic dataset translate well and can be used for improving the image quality of real pCLE images. However, the proposed data synthesis suffers from several weak points of the registration, such as limited availability of videos with undisturbed optical flow, lack of uniform SR powers on the entire field of view, and misregistration artefacts.

In Chapter 4 “A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks”, a solution for dealing with the irregularity of pCLE signals [227] allowing inputting sparse pCLE data to the CNN network was proposed. To make a systematic comparison between different CNN layers, the histopathological images were used as the equivalent of HR images to simulate LR pCLE-like images. The major drawback for pCLE is that this work is limited to the use of synthetic images based on histopathologies, without consideration of the real pCLE images.

Both solutions proposed in this thesis have focused on SISR algorithms that are trained on pairs of LR-HR images. Unfortunately, in pCLE, acquisition of HR videos is impossible, since high definition pCLE probes do not exist. The lack of HR pCLE thus severely limits the application of SISR to improve its image quality. Although attempts have been made to design DL-based SR techniques for pCLE [187, 227, 208, 209], there is an apparent need for a non-reference SR pipeline that does not need any ground truth images.

Our team used the potential of adversarial unsupervised training to get SR models without relying on HR pCLE images [206]; the work is coauthored but not reported as a contribution in this thesis. Instead of pCLE data, natural images were used as a source of HR data for the training. To ensure that the GAN is trained towards generating meaningful pCLE-like reconstructions, this work added a physically-inspired cycle consistency loss based on the reconstruction method using fibre pattern presented in [187]. Even though many successful works on GAN have been published, the downside of that method is that there is still uncertainty about whether GAN-based models generate results that deviate from the target domain. Moreover, there are still very few published works on blind SR for natural images [189, 215] and none for pCLE that discuss how to tackle SR on realistic images, as seen in clinical practice, without reliance on HR images.

In this Chapter, we expand on the originally proposed ZSSR methodology [189] to maximise its benefits specifically for pCLE data. The influential work of Shocher et al. [189] gave rise to a renewed interest in using deep non-reference SR for natural images [215]. The core of the Zero-Shot (ZS) framework for SR relies on two things: an image to train a model in a self-supervised ZS manner, and a downscaling kernel to reduce the resolution of this image. We propose a real-time ZSSR pipeline that uses a downscaling kernel, adapted explicitly for pCLE, and also a new ZS training strategy exploiting several primary frames from one pCLE video.

The essential part of the ZSSR framework is the downscaling kernel. Given the strong association between the downscaling kernel and the quality of the super-resolved image, it is essential to simulate LR pCLE as realistically as possible. To do so, we developed the necessary improvements by exploiting the geometrical quasi-hexagonal fibre pattern and the nature of pCLE noise. We designed a downscaling procedure that encapsulates a more realistic Voronoi-based downsampling kernel and mimics realistic pCLE noise.

Not only does ZS learning allow us to train a network without relying on HR ground truth images, but it also requires much fewer data to train the network. In the context of pCLE, we can nonetheless take advantage of the fact that the same bundle acquires all frames in any given pCLE video. Under the assumption that the temporal non-linearity of the bundle transfer function is constant, the frames in the video are correlated with each other by the same downscaling kernel. Thus, we can expand the training of our ZSSR to use a plurality of frames from the video, instead of using only one frame from that video.

The main objective of this work is to investigate non-reference methods for improving the image quality of pCLE videos. This topic constitutes a new domain with largely unstudied potential, and to be best of our knowledge, and at the time of publications, this work was the first to explore it for pCLE. The main contributions of this work are listed below:

- The design of a Voronoi-based downscaling procedure, which simulates realistic kernels that take under consideration the geometrical distribution of the pCLE signal sourced from irregularly distributed fibres.

- The extension of ZSSR to use several consecutive video frames during training to improve ZSSR in pCLE. Instead of using only a single image, we propose to take advantage of several consecutive frames from pCLE videos to extend the training data set, allowing for more generalised ZS training.

- Demonstrating that the SR pipeline in conjunction with the training on data that realistically models the nature of the pCLE noise reduces the prominence of the imaging noise occurring in pCLE.

The result is the first real-time ZSSR pipeline using a downscaling kernel tailored to the physics of pCLE, and with enhanced denoising capabilities.

5.2 Application of ZSSR to pCLE

5.2.1 Materials

In our experiments, we used two data sets: the first composed of clinical pCLE video sequences, and the second being a simulated video endomicroscopy dataset. The clinical data is used to test our methodology qualitatively, and the simulated data are used to test our methodology quantitatively.

Although we can test the IQ improvement qualitatively, this remains a complex and somewhat subjective task for human raters. We thus also use additional synthetic data to address the apparent lack of HR images for quantitative analysis. The pseudo endomicroscopies are a synthetic equivalent of real HR pCLE, and their availability facilitates quantitative validation of the proposed methodology. To obtain the pseudo endomicroscopy dataset we simulate it with the method originally introduced in [227]. This method mimics both signal distribution through a geometrical position of the fibres in the bundle and generates realistic pCLE noise. Thanks to that, the resulting images are characterised by typical triangulation artefacts and noise patterns, making our synthetic test cases more similar to real pCLE than basic simulation and hence reduce the domain gap.

For simulating synthetic endomicroscopy, we utilise the histological dataset named Kather[1] published in [146]. We used 10 large view hematoxylin & eosin (H&E) stained histological images of human colorectal cancer (CRC) and normal tissue, which were built with non-overlapping image patches, each 224x224 pixels (px) at 0.5 microns per pixel in RGB. The collections contain image samples of tissue classes such as adipose, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, colorectal adenocarcinoma epithelium. The samples with tissue slides were taken from NCT Biobank (National Center for Tumor Diseases, Heidelberg, Germany) and the UMM pathology archive (University Medical Center Mannheim, Mannheim, Germany). This dataset is a good choice for two reasons. First, histological images show structures similar to those which can be seen on real pCLE. In particular, the Kather dataset contains multi-class textures found in colorectal cancer histology. Secondly, the histological data were digitised with an HD camera and saved using lossless compression. The images thus show tissues in very high-quality magnification, without any significantly visible noise and compression artefacts. These high-quality images serve well as the equivalent of ground truth images.

We selected seven original pCLE video sequences from the pCLE-oriented

[1] `https://doi.org/10.5281/zenodo.1214456`

education platform Cellvizio.net[2]. Cellvizio.net contains high-quality video sequences presenting typical tissue found in conditions, such as Barrett's esophagus, gastric diseases, pancreatic cysts, irritable bowel syndrome, inflammatory bowel diseases. The video sequences are a good representation of pCLE clinical applications, as Cellvizio.net dataset provides video content allowing for comprehensive training of pCLE experts. The videos for the experiments were selected ensuring that our methodology is tested for diverse types of tissue and cancer stages. Examples of the selected video exports with their SR versions can be found in the supplementary material.

5.2.2 Zero-shot SR

Zero-shot framework allows training models capable of generalising from minimal dataset a distribution that was not seen during training time. In SR, the research work [189] explores information redundancy within one frame only to train an image-specific CNN network capable of predicting a super-resolved image from itself. It was shown that natural images contain recurring patches that capture the image representation well, so they may contribute to improving image content through the use of SR approaches [58].

Based on that finding, Shocher et al. used the LR frame as a source of LR-HR patches for training the image specific network to improve image resolution [189]. These patches are downscaled with either a bicubic kernel or an estimated kernel to simulate signal loss. Extracted patches are further augmented with eight rotation/flips to generate a bigger dataset. Pairs of LR-HR patches are used in a supervised manner to train a network. It is important to give attention to the fact that the input image has a dual role in ZSSR. It serves as the source HR image, giving rise to the paired LR-HR image patches for training, and also serves as the LR image at inference time. Predicted SR estimations are added to the training data set for further gradual training.

[2] Available at `http://www.cellvizio.net/`

5.2.3 Architecture and training

We designed a ZSSR pipeline (cf. Figure 5.1) tailored to video pCLE taking inspiration from the original ZSSR paper [189]. Formally, HR image I_{HR} is related to LR image I_{LR} by degradation process D:

$$I_{LR} = D(I_{HR}, \rho), \tag{5.1}$$

where ρ are parameters of the degradation process, such as scaling s, blurring h, motion g factors and noise n inherent from the imaging system. The blind ZSSR aims in recovering HR approximation of the ground truth $\widehat{I}_{HR}$ from I_{LR} under the assumption of known D:

$$\widehat{I}_{HR} = F(I_{LR}; \theta, D(I_{LR}, \rho)), \tag{5.2}$$

where F is the SR process and θ denotes the parameters of F. θ dependents on the SR process, and in the case of the DL model they are: network parameters, training hyper-parameters etc. In blind ZSSR I_{HR} is not available. The optimisation of F is based on the assumption that $F(\theta, D)$ is preserved with different scales and does not change depending on the input image scale. Thus, instead of I_{HR}, we used I_{LR} to get its downsampled version based on D. As the results, the objective of ZSSR is to find F by minimising the loss function $loss$ between the generated $\widehat{I}_{HR}$ and the test image I_{LR}.

$$\widehat{\theta} = \arg\min_{\theta} loss(\widehat{I}_{HR}, I_{LR}) + \lambda \omega(\theta), \tag{5.3}$$

where $\omega(\theta)$ is the regularisation term, and λ is the trade-off parameter.

Network architecture Identically to [189], we build an eight-layer CNN network with 64 kernels and ReLU as the activation function in each layer. We also use one skip-connection to learn the residual image instead of mapping it directly. We have shown that averaging of penultimate CNN layer [227] improves SR, thus differently to [189], in the proposed network the last CNN layer uses a kernel of size one and the linear activation function to reconstruct an SR image instead of

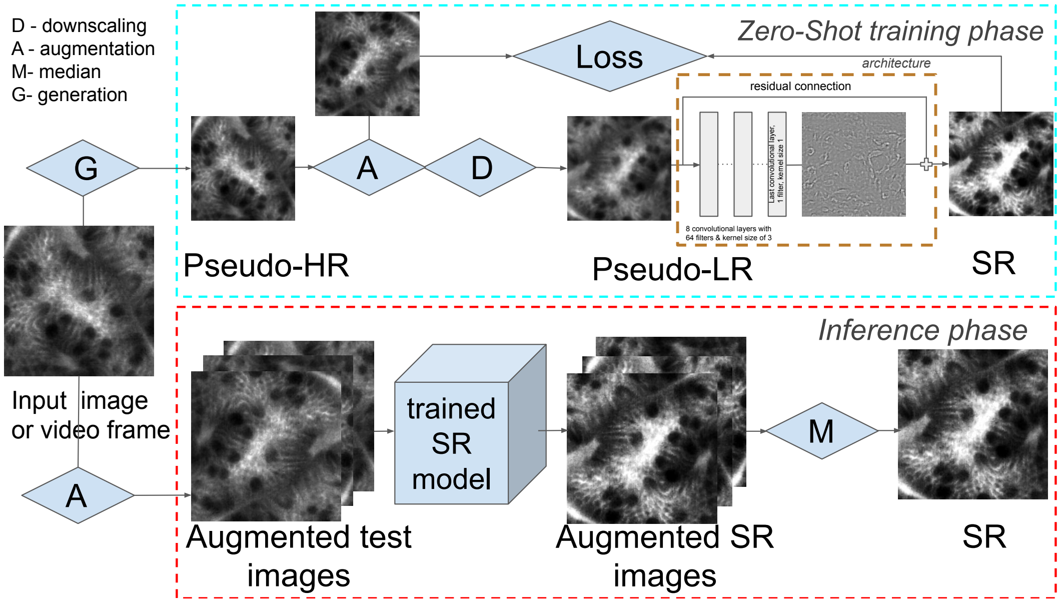


Figure 5.1: Graphical illustration of the proposed pCLE video ZSSR pipeline. The cyan (upper) block represents the training phase, and the red (lower) box represents the inference phase. The training starts with the input image or video. Note that, by design, the same input is used for both (zero-shot) training and inference. To extend the pipeline for handling multi-frame video, we iterate through a video stream, and each frame one by one becomes the input image. The input is used to generate (**G**) the pseudo-HR, then the downscaling (**D**) creates the pseudo-LR with a Voronoi (or bicubic for comparison) kernel with augmentation (**A**) is also used for the training of the SR network. For the inference phase, the input video in its original size and augmented to create eight images per frame, enhanced by a trained SR network. The final SR is a median (**M**) of these images after reverting the augmentation to the original frame. Detailed description is in Section 5.2.3.

learning another kernel.

Training procedure Our training approach is illustrated in Figure 5.1. In brief, training of the network starts from the input image. First, the input image is transformed to a pseudo-HR image, and next downscaled to a pseudo-LR image, which is further augmented to generate a set of pseudo-LR images. The training of ZSSR, marked by the cyan dashed box in Figure 5.1, is supervised on the LR-HR image patches. The patches are extracted from the pseudo-HR pCLE. A detailed description of the bespoke training data generation is provided in Section 5.3 “pCLE-specific zero-shot training data generation”. In this paragraph, we focus on our adaptations of the training procedures presented in [189].

pCLE has a circular FoV, resulting in an all-black region outside the circular region of interest. To avoid performing convolutions with those non-informative black pixels, we cropped rectangular patches from the pseudo-HR and LR images. A patch size of 340 pixels was chosen, as this is the largest crop region that can be applied across all videos. The central rectangular crop also facilitates augmentation, in particular rotation. We augmented the test images by using a combination of four rotations of {0, 90, 180, 270} degrees and two flips {`left, upside down`}, yielding eight unique augmentations per image.

Experiments on pCLE videos exploit multiple frames, instead of using the only first frame from the sequence. Starting from the 1st frame, we iterate through each video to extract consecutive frames into the training set. We expand our training dataset from 1 single frame to several frames from the video sequence, which is around 6 frames for each video that contains on average 60 frames (10% total number of frames in the video). More details are presented Section 5.4 "Experiments on synthetic images". These input train data are used during the training of the ZSSR video. Each image in the train set is cropped into the patch and augmented, as we describe it for the input image in the previous paragraph.

In the original ZSSR work, the pipeline benefits from tests performed during training. Shocher at al. [189] evaluate their model several times during training and create intermediate predictions. They add them to the training data as "HR fathers" to enhance the training dataset. In a similar vein, we evaluate our model 10 times during training, and every time we add the newly predicted image to the input data set, as the HR image. After every 100 epochs, we use the test image to evaluate the current model. The super-resolved image produced on each evaluation becomes the new "HR father", and it is added to the training set. Naturally, this grows the training set by one image each time, eventually leading to a ten-image training set in the end. To sum up, after performing 100 epochs, the model is evaluated. This is repeated until a total of 1000 epochs is reached. In the original work, authors sample "HR fathers" from the training dataset with a lower confidence level than the input image augmentations to avoid learning the wrong representation. In contrast, we unified the sampling to be equal for all images in our training set to avoid tuning additional hyperparameter (sampling rate) and simplify the implementation of our pipeline.

We trained the network with Adam as the optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We update the learning rate lr periodically with an exponential decay defined as: $lr \leftarrow lr \times dr^{\frac{\text{global step}}{\text{decay steps}}}$ where empirically we set the decay rate dr to 0.95 and decay step to 1000. Training starts with the learning rate of 0.001, gradually decreasing until it reaches 10^{-7} at the end of the training. Additionally, we use mini-batch training with a batch size of 8 to update weights more reliably than when only one training image is used.

Choice of loss Zhang et al. have shown that features extracted by any deep network can be adapted as a perceptual similarity metric [195]. They designed a framework called LIPIPS to calibrate three deep networks with scores of the images judged by human raters. Their approach improves the performance of the network in terms of similarity to human perception and outperforms the widely used PSNR and SSIM metrics. Another observation of theirs is that any DL network, regardless of training style, architecture or data used for training its weights, is sufficient to play the role of the general feature extractor in this context. Based on those observations, we decided to utilise the power of their framework in our pCLE application.

Following initial experiments with using the reference-based IQA metric LIPIPS [195] directly as a loss, we built a custom loss function by introducing an additional L1-based term. We used the pre-trained LIPIPS network[3], setting the linear calibration parameter to `net-lin` and selecting the VGG architecture as the network variant parameter. The network is pre-trained for RGB images (three-colour channels). pCLE videos have only one channel, so we converted our videos to RGB by replicating the available scalar channel. Our initial experiments confirmed that LIPIPS works well, as expected, for generating sharp and detailed images. Yet it enhances all small details in the pCLE images, including the characteristic pCLE noise. We observe that LIPIPS is not robust enough to handle pCLE noise. To help our model distinguish noise from the pCLE signal, we regularise the loss by adding

[3] Available at `https://github.com/richzhang/PerceptualSimilarity`

an L1 norm term. It steers the model in the direction of denoising. In summary, our new loss is defined as:

$$loss = \underbrace{\sum_{l} \frac{1}{H_l, W_l} \sum_{w,l} \parallel w_l \odot \hat{\vartheta}_r^l - \hat{\vartheta}_p^l \parallel_2^2}_{\text{LIPIPS}} + \underbrace{\lambda \frac{1}{n} \sum_{i=1}^{n} |I_r - I_p|}_{\text{L1}}, \tag{5.4}$$

where LIPIPS term [195] is computed with features given for a layer l as $\hat{\vartheta}_r^l$, $\hat{\vartheta}_p^l$ $\in \mathbb{R}^{H_l \times W_l \times C_l}$ that are extracted from a reference I_r and a predicted I_p image, respectively. The LIPIPS term is a channel-wise C sum of an average $L2$ distance in spatial dimension $\{W, H\}$ of the feature stack scaled by vector $w^l \in \mathbb{R}_l^C$. $L1$ term is an average distance between reference and prediction scaled by λ, which was set empirically to 5.

Using the trained model for predictions The inference is marked by the red box in Figure 5.1. Predictions are generated from the trained model by inference from the video stream, in which the first frame or several frames were used for training depending on the experiment setup. For prediction, we use original images, as seen in the clinic. The generation of a pseudo-HR is only used for training purposes. The near real-time inference per frame of our shallow network allows the predictor to process the video stream efficiently. It was shown that augmentation of the input image by rotations and flips improves SR [157]. Following that idea, we created eight augmented input frames. The augmentation is implemented in the same way as during training, and use arbitrary flips and rotation. We used augmented frames to obtain eight predictions. These predictions were transformed back to the test image reference frame, and the final SR image is a median image of the eight predicted SR images.

5.3 pCLE-specific zero-shot training data generation

5.3.1 Downscaling kernels in SR

Previous studies on DL-based SR have concentrated primarily on using a default bicubic kernel with anti-aliasing as counterparts of real downscaling kernels. The

NTIRE challenge [170, 219] aims at developing SR methods able to not only perform well on simulated images with a known bicubic kernel but also on “real” cases with unknown downscaling kernels. The results of the challenge demonstrated that the misestimation of the downscaling kernel affects the quality of super-resolved images. Furthermore, the authors of the original ZSSR work acknowledge these limitations, and they confirmed in their analysis that the performance of their SR pipeline is affected strongly by the choice of downscaling kernel.

Real images are influenced by limitations of the imaging system, such as sensor noise, non-ideal point spread function, and reconstruction artefacts. In reality, the downscaling kernel is typically not known, and when choosing one, there are several important technical considerations to be taken, such as the optical model, the noise model, etc. Despite considerable efforts in the past, precise kernel estimation has proved a difficult goal. DL models will underperform when trained on pCLE images downscaled with the default bicubic kernel, as it departs considerably from the acquisition physics of pCLE. All these are downgrading factors and influence the acquisition model, so should be considered when estimating a realistic downscaling kernel.

The problem of estimation of the downscaling kernels is not a new one. In the pre-DL SR literature, Irani et al., for example, estimated kernels blindly and showed that it improves their SR [116]. In order to improve super-resolved images, researchers are currently looking at better ways to include downscaling kernels into non-reference SR pipelines. For instance, in [189], the authors have fused both previously developed classical kernel estimation techniques [116] with the novel DL architecture. Although there are many classical solutions, the field of DL methods is evolving rapidly, with much space to develop novel deep learning methods with realistic kernels.

5.3.2 Noise modelling in pCLE

The works exploiting DL for SR are extensive, but they are primarily concerned with artificially generated LR images lacking any noise, and only recently the interest in real-world image SR increased [222]. Practically, noisy images are typically considered for denoising tasks [201]. Yet, there is a demonstrated improvement coming from using SR and denoising jointly [174, 194, 215]. In [189], the authors have shown that adding a small amount of Gaussian noise to LR images appears to have a positive impact on their ZSSR pipeline. They noticed that the optional addition of noise helps the network to distinguish between the correlated mapping of LR-HR pairs and uncorrelated noise. They showed that when models are trained with noise, it directly impacts SR results towards higher PSNR [189].

The pCLE noise contributes importantly to lower the pCLE image quality, creating characteristic textured noise patterns affecting many neighbouring pixels around a noisy fibre. Until now, there has been no systematic study considering the impact of simulating noise in order to train an SR network for the additional task of denoising in pCLE. To the best of our knowledge, there is hardly any published evidence in the literature regarding the effects of using noisy LR pCLE in ZSSR.

Having in mind the specificity of pCLE, we adjust the ZSSR pipeline by: first, generating a pseudo-HR image from the test image to obtain an estimate of the higher information density than test image displays, which is described in detail in Section 5.3.3 "Pseudo-HR generation"; and second, downscaling procedure based on Voronoi diagram build for a fibre bundle that embeds simulation of pCLE noise that enables signal loss from pseudo-HR to pseudo-LR, which all steps are described in Section 5.3.4 "Voronoi-based downscaling", and presented on Figure 5.2.

5.3.3 Pseudo-HR generation

The standard pCLE reconstruction algorithm relies on a Delaunay triangulation to linearly interpolate noisy fibre signals onto an oversampled Cartesian grid. The oversampling ratio is chosen to minimise information loss stemming from the reconstruction, leading to an average of 7 pixels per fibre [28]. A bundle built with 25k fibres would thus capture 25k unique fibre signals, and those are used to reconstruct a Cartesian image with around 175k pixels, excluding any black borders.

Consequently, the pixel information rate is only 1/7 on the oversampled original pCLE reconstruction.

The generation of a pseudo-HR image from the input image is depicted in Figure 5.2. This stage is designed to reduce redundant pixels generated by the standard pCLE reconstruction, while allowing for some data loss, and reducing the distortion of the pCLE textured noise pattern. We reduce the redundancy of information in the oversampled input image by reconstructing the sparse pCLE input data on a grid n times smaller, in terms of a total number of pixels, than the original one. For that, we rely on the standard pCLE reconstruction algorithm. We thus obtain an average of $n/7$ pixel per fibre in the reconstruction. Such a newly created image is referred to as a pseudo-HR image, as it displays much less information redundancy at the reconstructed pixel level. Thanks to reconstructing the pseudo-HR on a twice smaller grid than the original pCLE image, we create more compact information that provides high-frequency information by making the pixel information rate 4/7. This highly packed information serves as an estimate of the HR with a higher density of information than the original test image. Another beneficial property coming from the reconstruction of the pseudo-HR image is that it does not include the typical pCLE noise pattern. Thanks to that, the noise in the image does not adversely impact its information content and potentially paves the way for better SR to be obtained.

5.3.4 Voronoi-based downscaling

A well-known limitation of the bicubic downscaling kernel when applied to natural images is its inaccurate approximation of the real LR creation process [170]. This discrepancy is even higher in the case of pCLE, given the irregular fibre-induced sampling and reconstruction characteristics of the device. Hence, a better downscaling approach is needed in our case. To generate LR pCLE images, we designed a novel downscaling procedure based on the pattern of fibre and reference pCLE reconstruction method. The downscaling procedure consists of several steps: kernel estimation, acquisition of LR signals with a Voronoi vectorisation, noise generation, and reconstruction of the LR image as illustrated in Figure 5.2.

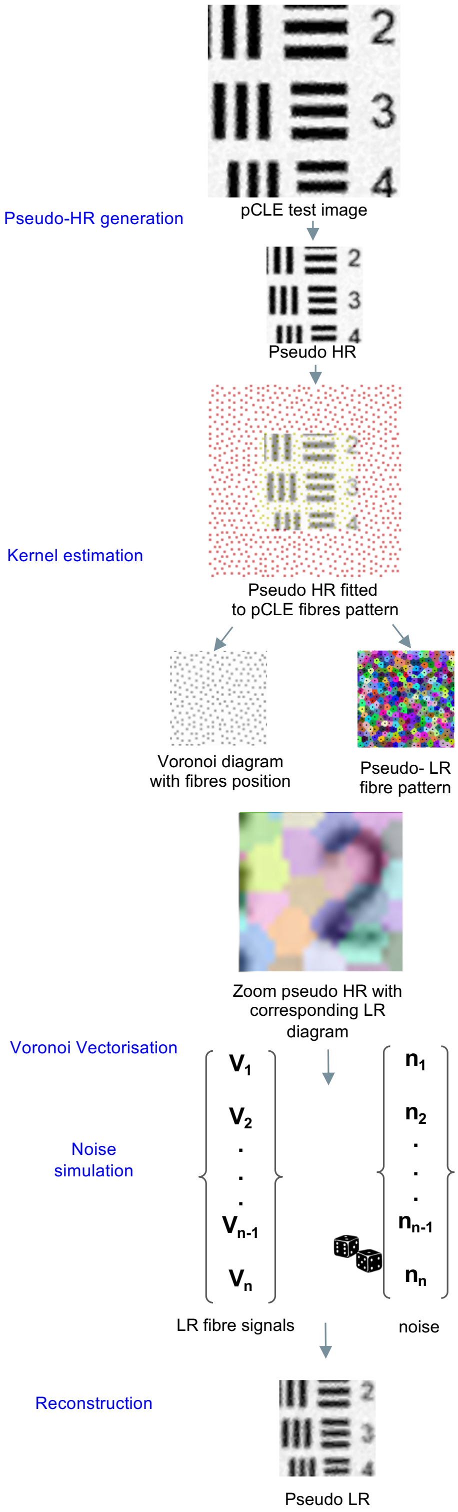


Figure 5.2: Voronoi downscaling: A test pCLE image is reconstructed as a pseudo-HR; A pCLE fibre pattern is fitted to a pseudo-HR space to estimate kernel as a pseudo-LR Voronoi diagram; The averaged signals retrieved from the diagram by a Voronoi vectorisation are reconstructed to form a pseudo-LR pCLE.

Kernel estimation In the first step, we use the pseudo-HR image to construct the pseudo-LR fibre pattern used to create a kernel in the process of Voronoi downscaling. To achieve that, we retrieve the original pCLE fibre pattern from the input image metadata. To construct the pseudo-LR fibre pattern, we associated the pseudo-HR image with the original fibre pattern. The fibres in the original pattern are distributed in a pseudo-regular pattern (quasi-hexagonal). Thus, pseudo-HR's exact position within a fibre pattern space does not play a significant role, as the characteristic distribution of fibres is preserved within a bundle. As the original pattern has a circular shape, the most straightforward implementation is to fit the pseudo-HR area to the centre of the pattern to avoid edge problems. As depicted in Figure 5.2, the pseudo-HR image spans a smaller area than the fibre bundle of the test image does. Fibres outside the pseudo-HR space are discarded (marked red in Figure 5.2). Consequently, the geometrical distribution of fibres remains unchanged, but the density of fibres per structure in the image decreases. Thanks to the fact that we reuse the fibre pattern from the input image, we ensure that we preserve the signal acquisition's nature through the fibre bundle, including typical fibre signals and their geometrical position.

Voronoi vectorisation We use the new fibre pattern to simulate signal loss for the acquisition of the LR signals. Every fibre signal contributes to the reconstruction of the neighbouring pixels. This contribution is defined by the Delaunay triangulation created using the position of fibres in the fibre pattern. To acquire the pseudo-LR signal, we use the pseudo-LR fibre pattern to find each fibre position. The region of influence of each fibre is given as the space covered by the corresponding cell in the Voronoi diagram, which is the dual of the Delaunay triangulation.

Formally, let $X = f_1 \dots f_n$ be a set of n fibres' position within fibre pattern in $\mathbf{R}^d$. The Voronoi diagram generated by X is the partition of the fibre pattern into n convex cells, the Voronoi cells, V_i, where each V_i contains all points of $\mathbf{R}^d$ closer to f_i than to any other point:

$$V_i = \{x : \forall j \neq i, d(x, f_i) \leqslant d(x, f_j)\}, \tag{5.5}$$

where $d(x,y)$ is the Euclidean distance between x and y. As presented in Figure 5.2, we construct the diagram in such a way that each cell contains one fibre in its centre. Each cell in the Voronoi diagram represents the fibre's FoV and its contribution to the image pixels. Effectively, the signal in the cell V_i is discrete and includes m pixels p. We estimate the LR signal in each cell LR_i by averaging the pseudo-HR $\widehat{HR}_s$ signal covered by each fibre region of influence:

$$LR_i = \frac{1}{m}\sum_{p=1}^{m} \widehat{HR}_s(p),\ p \in V_i, \tag{5.6}$$

We create a vector of the pseudo-LR signal $\overrightarrow{LR} = \{LR_i, ..., LR_n\}$ of the length n which equals the number of fibres in the pseudo-LR pattern.

Noise generation In the seminal ZSSR work [189], the authors studied the application of their pipeline to real, noisy images. They noticed that SR training may benefit from adding extra small Gaussian noise to pseudo-LR images. However, while Gaussian noise can be considered appropriate for natural images, this does not translate directly to our case. pCLE images display a characteristic type of noise. This is produced due to the optical limits of the system, and a specific image reconstruction algorithm, which is very different to a standard digital camera.

The pCLE noise has a distinct texture that stems from the fact that the current reconstruction algorithm interpolates noisy fibre signals along with the tissue signal onto the Cartesian grid affecting several pixels. Depending on the experimental setup, we can also add noise to the LR signal prior to pseudo-LR image reconstruction, as described below.

Based on our previous experiences with simulating synthetic pCLE images [187, 227, 206], in order to generate realistic LR, we used the pCLE noise model suggested in [28]. The authors defined two types of noise: acquisition noise that can be modelled as additive noise; and calibration imperfection that is modelled by multiplicative noise. Similarly, to our previous works, we simulated both noise types by sampling them from a Gaussian distribution with a mean of zero. In contrast to our previous approach, we simulate varying levels of noise in each augmented pseudo-LR frame. We achieved that by drawing noise from Gaussian distributions with a different standard deviation value per frame. We set sigmas σ_i as $\sigma_1 + c_1$ for additive noise and $\sigma_2 + c_2$ for multiplicative noise, where σ_i is chosen empirically and set σ_1 to 0.03 and σ_2 to 0.05, c_i is a random number between [-0.025, 0.025] drawn from a uniform distribution for each frame independently. The range was chosen empirically by visually inspecting noise at the images.

The noise is added directly to the pseudo-LR fibre signals, before interpolation-based reconstruction.

Reconstruction Finally, we reconstruct pseudo-LR with the "gold standard" reconstruction method based on the Delaney triangulation and interpolation. We use pseudo-LR fibre pattern to construct triangulation and acquired pseudo-LR signals with or without noise to interpolate them onto a Cartesian image of the same size as the pseudo-HR image.

5.4 Experiments on synthetic images

We study the application of zero-shot learning to the pCLE SR task using synthetic data. Special care was taken to simulate realistic endomicroscopy data to ensure that the domain gap with real pCLE data is reduced as much as possible. We compare super-resolved images obtained on the synthetic test set with the corresponding synthetic HR pCLE. As the baseline method, we used the reference reconstruction method currently used in the clinic, which amounts to the identity transformation in case no synthetic noise is added to the simulated LR input image. For comparison, we use IQA reference-based metrics: PSNR, SSIM, VGG loss, L1 loss, LIPIPS and Gradient Magnitude Similarity Deviation (GMSD) [119].

Based on the comparison of the super-resolved images with synthetic HR data, which serves as ground truth, we test whether there is a difference between compared video sequences on a frame-by-frame basis.

Our aim was to test the following research hypotheses (**H***i*):

H1 : Voronoi downscaling leads to better pCLE ZSSR than bicubic downscaling

H2 : pCLE ZSSR benefits from joint training of both the SR and the denoising tasks.

H3 : ZSSR benefits from expanding the training dataset from a single frame to several video frames.

H4 : pCLE ZSSR performs better in comparison to SISR and state-of-the-art DCNN

For each research hypothesis above, we constructed a corresponding statistical null hypothesis and analysed the statistical significance of the IQA improvement with the two-sided student t-test using a significance level of $p = 0.05$. The proposed research hypotheses are evaluated with predictions generated by seven DL models and baseline:

- model 1 - baseline linear interpolation,
- model 2 - ZSSR trained with noise free frames and Voronoi kernel,
- model 3 - ZSSR trained with noise free frames and Cartesian kernel,
- model 4 - ZSSR trained with noisy frames and Voronoi kernel,
- model 5 - ZSSR trained with noisy frames and Cartesian kernel,
- model 6 - ZSSR trained with noisy video frames and Voronoi kernel,
- model 7 - SISR trained with noisy frames and Voronoi kernel,
- model 8 - pre-trained DCNN [174].

Sample synthetic videos are in Chapter 7 “Appendix - video reconstructions”, Figure 7.5. The quantitative results from our IQA on synthetic pCLE is presented in Table 5.1, with models referred by number in round brackets. Here, we provide an experimental evaluation of the results per hypothesis for our ablation study.

Table 5.1: Image Quality Assessment for synthetic experiments. We run ablation study to investigate the performance of models employing either Voronoi (2, 4) or Bicubic kernel (3, 5) for simulated noise or noise-free data. We compare Video ZSSR with Voronoi kernel (6) against SISR trained on synthetic data (7), state-of-the-art DCNN (8) [174], baseline interpolation reference method (2).

model	baseline	ZSSR					SISR	DCNN [174]
training	interpolation	noise-free frame		noisy frame		noisy video	simulation	pre-trained
kernel	LR (1)	Voronoi (2)	Cartesian (3)	Voronoi (4)	Cartesian (5)	Voronoi (6)	Voronoi (7)	Cartesian (8)
PSNR	27.99 ± 1.08	28.86 ± 1.08	28.27 ± 0.99	30.16 ± 1.22	28.23 ± 0.93	30.67 ± 1.27	30.99 ± 1.29	28.04 ± 1.08
SSIM	0.851 ± 0.020	0.880 ± 0.017	0.862 ± 0.014	0.878 ± 0.015	0.849 ± 0.021	0.890 ± 0.014	0.902 ± 0.012	0.852 ± 0.020
LIPIPS	0.781 ± 0.017	0.806 ± 0.015	0.785 ± 0.012	0.806 ± 0.014	0.777 ± 0.018	0.817 ± 0.012	0.816 ± 0.014	0.782 ± 0.017
GMSD	0.940 ± 0.006	0.954 ± 0.007	0.953 ± 0.004	0.955 ± 0.004	0.943 ± 0.006	0.960 ± 0.004	0.961 ± 0.004	0.940 ± 0.006
L1 loss	0.973 ± 0.003	0.975 ± 0.003	0.974 ± 0.003	0.980 ± 0.003	0.973 ± 0.003	0.981 ± 0.003	0.982 ± 0.003	0.973 ± 0.003
VGG	2.62 ± 0.19	2.39 ± 0.17	2.70 ± 0.14	2.37 ± 0.16	2.69 ± 0.20	2.24 ± 0.14	2.25 ± 0.16	2.60 ± 0.19

Study of the downscaling kernel choice (H1)

In our ablation study, we compare super-resolved images generated by models trained with our proposed Voronoi kernel against models trained with a baseline bicubic kernel. The models marked in Table 5.1 by numbers 2-5 were trained for both downscaling kernels independently. The same models have also been used in the study of the effect of noise on the pipeline.

As described in more detail in Section 5.3 "pCLE-specific zero-shot training data generation", the pseudo-HR image is obtained by reconstructing the input image on a grid twice times smaller than the original grid. Based on visual inspections of the images, we found that this downscaling rate is optimal for creating a smaller image without lose of information. This new image is used in place of the input image during a training phase serving as an HR image.

The pseudo-HR is also a part of the pCLE downscaling procedure described in Section 5.3 "pCLE-specific zero-shot training data generation" and marked by **D** in Figure 5.1. The protocol for downscaling with the bicubic kernel is designed as follows. First, we downscale the pseudo-HR image using the native TensorFlow `tf.image.resize` function with the bicubic kernel, anti-aliasing and an empirically chosen scale value of 3. Our ZSSR architecture is designed to use the LR and HR images of the same size and does therefore not embed any upsampling layer. To comply with our pipeline, we upsample the downscaled intermediate image with the bicubic upsampling to go back to the size of pseudo-HR. Since the Cartesian downscaling does not use fibre patterns, it is impossible to simulate typical pCLE noise with it. Nonetheless, we add multiplicative and additive noise pixel-wise on the Cartesian LR images for training model 5 only.

The results in Table 5.1 demonstrate the suitability of the Voronoi kernel in ZSSR. Utilising Voronoi downscaling yields consistently better results than bicubic downscaling for all metrics. We run a two-sided paired t-test for all frames inferred by models 2, 3 and 4, 5 which confirm that there were a statistically significant improvement in image quality of the reconstructed SR images from models exploiting Voronoi downscaling. From the observer perspective, videos enhanced by models taking advantage of Voronoi downscaling were visibly sharper than those that use bicubic downscaling; sample videos are in Chapter 7 “Appendix - video reconstructions”, Figure 7.5 . Therefore, we confirm that models trained with images downscaled with Voronoi kernel (models 2 and 4) yield better reconstructions than ones trained with Cartesian downscaling (models 3 and 5).

Study of the impact of simulated noise on ZSSR (H2)

We trained the ZSSR pipeline with two types of data: noise-free and noisy pseudo-LR. To test the influence of noise on the image quality of pCLE, during every training iteration, the LR frame is augmented by simulating noise. Because ZSSR limits training to one data sample, it is essential to augment the training dataset, ensuring variability of transformation. Thus, a new random noise signal is drawn each time weights are updated. This makes the network capable of capturing not only one noise pattern characteristic for the test image, but also the random nature of the noise. The experiments were repeated for each downscaling kernel. The quantitative scores are shown in Table 5.1 for models 2-5.

It is unsurprising to find that the models trained on noise-free images (models 2 and 3) produce models unable to distinguish between true signal and noise on the test pCLE images, while models which account for noise (models 4 and 5) generalise towards denoising. One of the key findings is that for pCLE data image quality improvement can be implemented for both super-resolving and denoising capabilities in one model. This holds for both types of downscaling kernels, yet models trained with Voronoi kernel (models 2 and 4) tailored to the specific pCLE acquisition noise yield results with significantly higher PSNR of 2.0 dB. It can be attributed to the fact that Cartesian noise does not represent typical pCLE patterns well. Therefore, this is supporting evidence that our physically inspired downscaling Voronoi kernel accounting for pCLE noise fits the pCLE ZSSR task better. Voronoi downscaling models noise patterns by interpolating them during the reconstruction of the image from the noisy fibre signals, which is demonstrated to be a crucial element of pCLE denoising.

Lastly, it is important to note that the models trained only for SR task not taking noise into consideration (models 2 and 3) generate higher SSIM, than models which perform SR and denoising together (models 4 and 5). Although SSIM scores are slightly lower for the SSIM on a combined SR and denoising task, that difference is attributed to the fact that SSIM is a robust metric in the presence of the noisy signal. SSIM does not measure the difference in noise between images precisely enough, since it is more suited as a structural metric, and PSNR is a better metric in this case. There is a minimal difference for GMSD, LIPIPS and VGG in favour of models trained with noisy data. This slight difference, like in the case of SSIM, may be attributed to the fact that these metrics are not designed to measure the impact of noise only, but rather complex structural similarity. The L1 loss behaves similarly to PSNR giving a better result for models aware of the noise. Informal visual inspection of the images aligns with quantitative results. It is noticeable that images with reduced noise are characterised by higher quality.

Study of the training set size extended to multiple-frames training in video ZSSR (H3)

The previous experiments were limited to use of the only one (first) frame from the video sequence. We observe that several consecutive frames show similar content. Frames are correlated and characterised by small information entropy. These frames altogether are nonetheless more informative than one frame and can be used to build a more robust training dataset. Importantly, the video is acquired with the same fibre bundle, and from the same tissue and patient, thus one downscaling kernel can be used for all frames in the video. Based on this observation, we expand our training dataset from 1 single frame to the first 10% of the frames from the video sequence, which is around 6 frames for each video.

The results in Table 5.1 obtained from the model trained with several frames (model 6) only slightly outperform with statistical significance the models trained with one frame (models 2-5). The improvement for PSNR is small; only 0.5dB when compared with the model trained on one image and Voronoi kernel with noise simulation. It may suggest that it is sufficient to train the denoising model with one frame, which is augmented every iteration with newly drawn noise. Although the increase in the number of frames gives only a small increase in PSNR and L1, it significantly benefits SSIM, LIPIPS, VGG and GMSD. Since SSIM, LIPIPS, VGG and GMSD measures the structural similarity between images, we can conclude that the extended dataset benefits SR more than the denoising task. On the other hand, there is only a modest improvement visible the naked eye during informal visual inspections of videos. These findings reinforce the general belief that bigger datasets help in the training of neural networks, which is also valid for zero-shot learning.

Study of performance of training type – Supervised SISR and pre-trained state-of-the-art DCNN vs unsupervised ZSSR (H4)

In order to provide a comparison of unsupervised zero-shot learning to the supervised single image SR, we also trained the same network architecture presented in Figure 5.1 using a SISR approach (model 7). Similarly to our previous study [227], we use the synthetic test and train dataset, and trained the residual mapping between synthetic pairs of HR-LR images. We also compared ZSSR to the state-of-the-art network for joint denoising, and SR called DCNN with a pre-trained model 'dncnn3'[4] (model 8). Those (models 7 and 8) are used to generate a prediction for a synthetic test set.

We report in Table 5.1 that ZSSR (model 6) statistically does not outperform SISR (model 7). As a matter of fact, supervised SISR performs significantly better. However, our findings are not surprising, since it is expected that models trained in a supervised manner perform better than the ones trained as unsupervised as in the case of ZSSR. Additionally, the informal visual inspection does not capture any significant differences between videos, which suggest that both models are trained towards the same/similar solution. This, in turn, provides more confidence to unsupervised ZSSR in the reconstruction of pCLE images. It also highlights that kernel choice may be a powerful driver towards converging models to the optimal solution, and data size, while still important, may play a smaller role. On the other hand, ZSSR (model 6) outperforms the DCNN (model 8) by a large margin. It is

[4] Available at `https://github.com/cszn/KAIR`

related to both the fact that DCNN is pre-trained with natural images for denoising the Gaussian noise and uses the bicubic kernel as a downscaling process. We have shown that both elements are not suitable for pCLE, in **H1** and **H2**, as it has a very distinct noise type and Voronoi kernel mimics downscaling better than the bicubic kernel.

5.5 Experiments on original images

In our experiments on synthetic data, we demonstrated that the most significant improvement in the image quality of pCLE is achieved by the supervised SISR and the unsupervised video ZSSR models. In this section, we describe how we translated these two models to our original data, and evaluate the quality of obtained reconstructions with reference to each other, and the baseline interpolation method by subjective assessment of the image quality performed as a user study.

Thanks to the reference-free ZS training scheme, we can train SR models on original pCLE videos, not relying on (unavailable) real HR images. We trained seven independent ZSSR models, one per test video selected from our original dataset described in Section 5.2.1 “Materials”. We designed Voronoi kernels specifically to account for the acquisition of each original video with its unique fibre pattern. We also simulated pCLE-like noise with sigmas σ_1 and σ_2 set as 0.1 for additive noise and 0.5 for multiplicative noise. The sigmas were chosen based on informal visual inspection of images. For the training, we used the first 10% of all frames in each video, which have around 60 frames in total. This number of frames allows for the creation of a representative sample of the video and allows for fast training. Once models converge, we interfere results for all the frames in the test video on which the model was trained to create SR video reconstruction. While ZSSR has a blind training protocol that allows creating video-specific models directly from original data, SISR cannot be retrained easily without ground truth HR pCLE images. Nonetheless, it was shown that original pCLE reconstruction benefits from models trained on synthetic data [187]. Thus, we re-use the SISR model trained on synthetic data presented in Section 5.4 “Experiments on synthetic

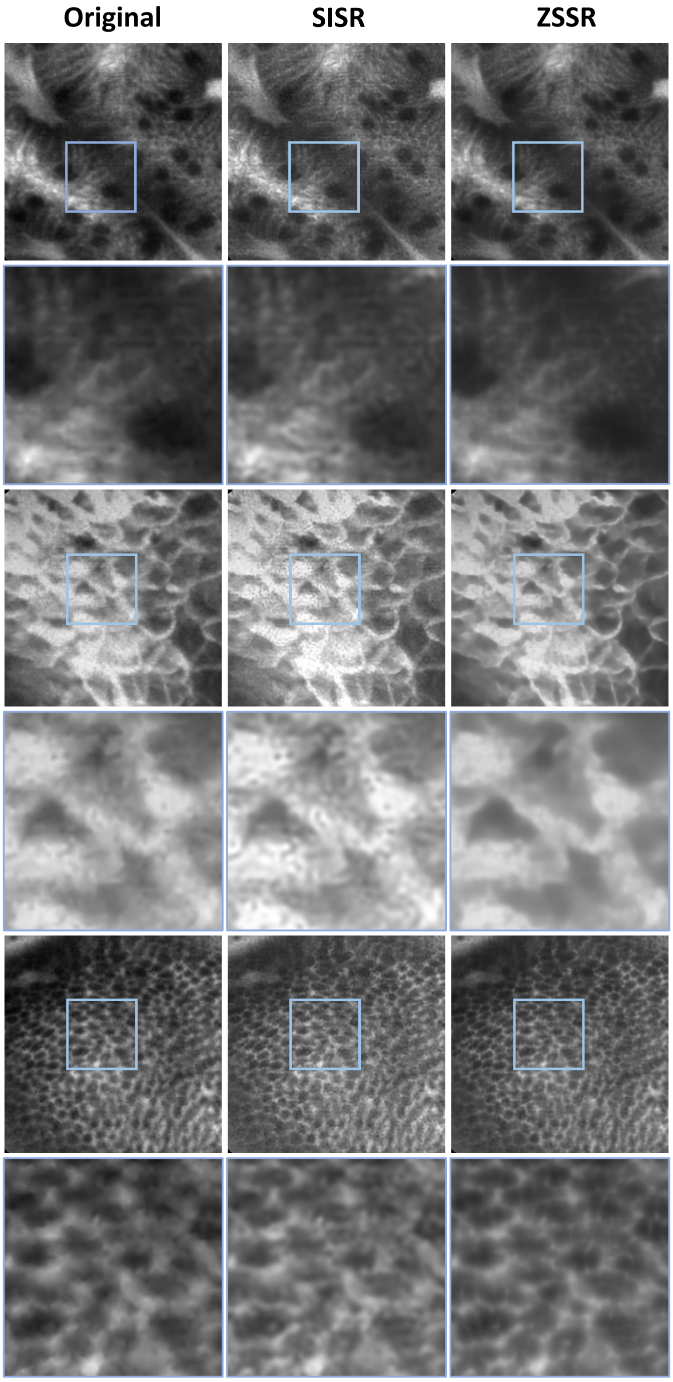


Figure 5.3: The results for two different tissues for the baseline, SISR and video ZSSR (starting from the left). Each odd row presents a patch from an original pCLE image; each even row presents the zoomed square marked with blue on the image in the row above it.

images" and use it to infer results for the original test videos. Thanks to the realistic simulation of the synthetic training dataset, we expect that this SISR model performance translates well for the task of improving the quality of original data. The example results are depicted in Figure 5.3.

In the absence of original ground truth HR images as reference, well-established reference-based metrics are ill-suited to assess the quality of the super-resolved images obtained from this data. To evaluate our proposed methodology in application to real pCLE videos, we asked observers to compare images reconstructed from three methods: ZSSR, SISR and baseline. The preferences of the observers indicate an improvement in the pCLE image quality.

Among the most common user studies for classifying image quality such as a single-stimulus, a double-stimulus, and similarity judgements, it was shown that a forced-choice pair-wise comparison is the most accurate and time-efficient [98]. The most consistent results are generated by a cognitively easy approach. In the forced-choice approach, observers need to compare only two images at a time and make a quick binary decision without any rating scales. Based on the suitability of a forced-choice approach, we designed an image quality assessment survey based on a two-alternative forced-choice (2AFC) paradigm to test the following research hypotheses (**H***i*):

H5 : video ZSSR with simulated noise yields super-resolved pCLE images, which are preferable to the baseline reconstruction by both experts and non-experts.

H6 : SISR yields a pCLE super-resolved image, which is preferable to the baseline reconstruction by both experts and non-experts.

H7 : video ZSSR with simulated noise outperforms SISR in improving the image quality of the reconstructed pCLE by the judgement of both groups of raters.

Survey structure We compare three methods by directly inferred pair comparisons: baseline vs SISR, baseline vs ZSSR, and SISR vs ZSSR. It is essential to consider that pCLE reconstructions are similar, and the cyclic relations may occur.

To avoid any error coming from such cyclic relationships, we decided to test all possible combinations giving three direct comparisons.

The images used in the survey were extracted from all test videos for each of methods: SISR, ZSSR and baseline. For each comparison, we extracted 12 random frames from each video. Frames are used only once. The same frame is never re-used in another comparison. Thanks to extracting multiple frames from each video, we can assess results on both frame basis and cumulative video basis. We ensure that each video comes from a different tissue, patient and fibre, assuring high variability of data dynamic range among real clinical cases.

Each question tests one of the three comparisons. There are 252 questions in the entire survey, 84 questions for each comparison pair. All questions were randomised to avoid correlations between questions, such as the same video, frame or model tested in adjacent questions. For each question, two images from two different methods are displayed in a random order, and with different randomisation.

Observers The image quality of the pCLE reconstructions is assessed by the subjective perception of image quality of expert and non-expert observers. All recruited in our study experts have over two years of experience working with pCLE images. Non-expert users did not have exposure to pCLE images prior to the survey.

Experimental procedure The observers received the survey as a website. On the first page, the display conditions of the survey were given to the observers. The observers were asked to run a survey on a computer screen size bigger than 13" in full window mode, to ensure visualisation conditions mimicking a clinical set-up. Second, they were instructed how to use the survey tool for a more detailed comparison with the ability to zoom in on the images. Lastly, users were shown a sample question. Since the task of comparing two images is intuitive and relatively easy for human observers, it does not require extensive training before starting the survey.

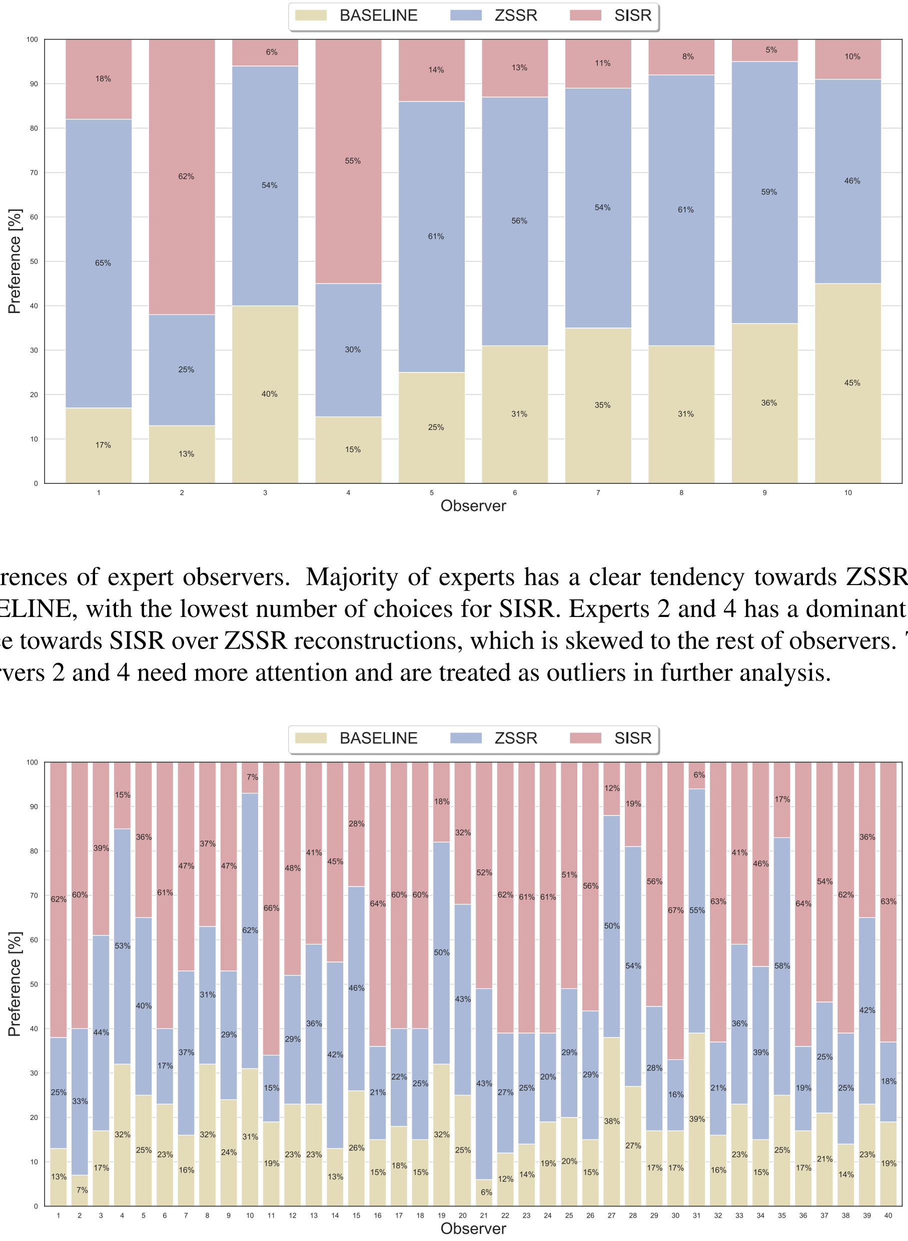
The observers were shown two pCLE reconstructions arising from two of the tested methods next to each other. They were asked to select a higher quality image. Users were instructed to assess images in less than 15 seconds per question, but no time restriction was imposed by the survey tool. The observers spent 26 minutes on average to fill in the survey. They had to answer every question in the survey and were not allowed to come back to a previous question.

5.5.1 Results

There were 10 experts and 40 non-experts taking part in the survey. Since expertise plays a key role in understanding images, we stratify the analysis by the group. The responses to the survey are summarised in Table 5.2, and detailed answers per observer are visualised in Figure 5.4.

Table 5.2: Summary of preferences for Image Quality Assessment survey for both experts and nonexperts. There are three tests: Baseline vs. ZSSR (1); baseline vs. SISR (2); and SISR vs. ZSSR (3). The preferences are given as the percentage (%) of frames chosen by all observers in each comparison.

	Test 1		Test 2		Test 3	
model	BASELINE	ZSSR	BASELINE	SISR	SISR	ZSSR
experts	23	77	64	36	25	75
nonexperts	36	64	25	75	62	38



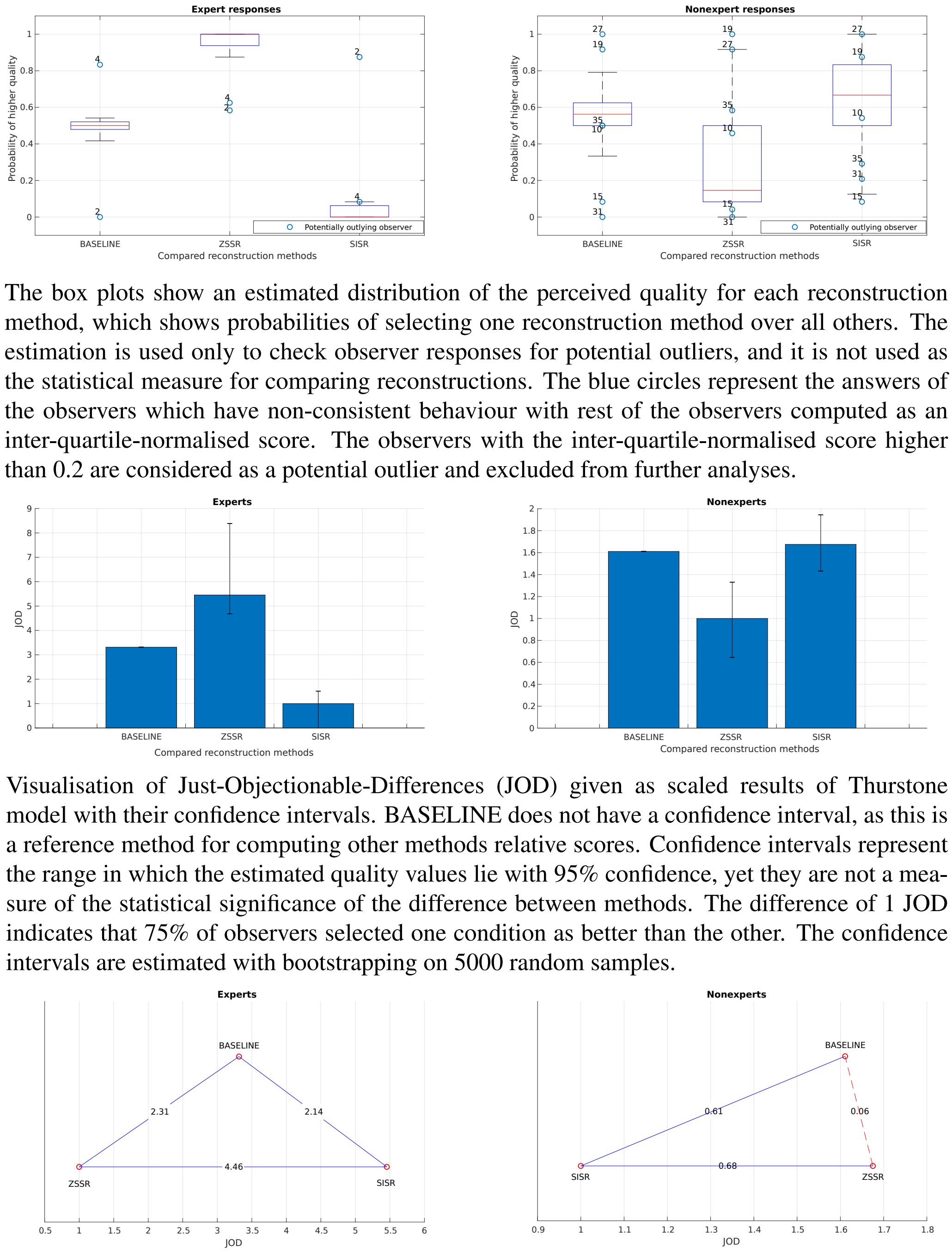
(a) Preferences of expert observers. Majority of experts has a clear tendency towards ZSSR over BASELINE, with the lowest number of choices for SISR. Experts 2 and 4 has a dominant preference towards SISR over ZSSR reconstructions, which is skewed to the rest of observers. Thus, observers 2 and 4 need more attention and are treated as outliers in further analysis.

(b) Preferences of non-expert observers. There is a clear preference for SISR reconstruction for the majority of observers. In contrast, no dominant preference occurs for neither ZSSR and BASELINE. In comparison to others, observers, e.g. 4, 10, 15, 19, 27, 28, 31, 35 do not show a preference for SISR; they should be investigated as statistically valid outliers.

Figure 5.4: The preferences of the observers for each of the reconstruction methods: BASELINE, ZSSR and SISR. The preference is shown as a percentage of the choices calculated as a number of choices for each method from total possible choices in the survey for an expert (5.4a) and a non-expert (5.4b).

Although categorical data can be ranked via direct vote count, additional care is needed to measure the magnitude of the differences between conditions by statistical analysis. Hypothesis testing of 2AFC requires modelling a quality scale based solely on the comparison ranks. Among the most popular methods for pair-wise comparisons in image quality is the Thurstone-Mosteller model (Thurstone's Law of Comparative Judgment, Case V) [15]. The model represents linear paired comparisons with probabilities of preference mapped to scales. The main assumption of the model is that users opinions about the quality of the stimulus can be estimated with a Gaussian distribution where each stimulus in the test is independent and identically distributed, meaning with equal (or zero) correlations and equal variance. The results of comparisons refer to the image quality difference that is scaled as Just-Objectionable-Differences (JODs). JOD measures which stimulus is closer to an ideal quality reference. Particularly, we used the Thurstone Case V scaling method with outlier analysis and statistical testing[5] to analyse the observer preferences. The results of the statistical analysis are in Figure 5.5.

[5] Available at `https://github.com/mantiuk/pwcmp`



(a) The box plots show an estimated distribution of the perceived quality for each reconstruction method, which shows probabilities of selecting one reconstruction method over all others. The estimation is used only to check observer responses for potential outliers, and it is not used as the statistical measure for comparing reconstructions. The blue circles represent the answers of the observers which have non-consistent behaviour with rest of the observers computed as an inter-quartile-normalised score. The observers with the inter-quartile-normalised score higher than 0.2 are considered as a potential outlier and excluded from further analyses.

(b) Visualisation of Just-Objectionable-Differences (JOD) given as scaled results of Thurstone model with their confidence intervals. BASELINE does not have a confidence interval, as this is a reference method for computing other methods relative scores. Confidence intervals represent the range in which the estimated quality values lie with 95% confidence, yet they are not a measure of the statistical significance of the difference between methods. The difference of 1 JOD indicates that 75% of observers selected one condition as better than the other. The confidence intervals are estimated with bootstrapping on 5000 random samples.

(c) The representation of an analysis of a statistical significance for the comparison of the reconstruction methods. Red points indicate reconstruction method, and they are connected to the compared methods of the interest. The statistically significant differences are shown as solid blue lines, as opposed to red dashed lines. The x-axis represents the scaling scores from plot (5.5b), and the y-axis supports schematic visualisation of the difference between methods in form of triangles.

Figure 5.5: Statistical analysis of the pairwise comparison experiments with Thurstone model with open source implementation available at `https://github.com/mantiuk/pwcmp`.

H5: ZSSR vs BASELINE

The results for both observer groups show broadly similar traits. The experts show clear preference by giving 77% of votes towards ZSSR reconstructions over baseline solution with the statistical significance confirmed by fitting a Thurstone-Mosteller model. The preference towards ZSSR is also seen among non-expert users, yet results vary between videos, and statistical significance was not achieved for their cumulative choice.

H6: SISR vs BASELINE

An interesting pattern emerges when examining the preferences for SISR. It is very apparent that experts prefer the baseline method (64% of votes) over the learning-based SISR approach. On the other hand, non-expert observers showed polar behaviour and chose the SISR model as their favourite (75% of votes). Both voting are shown statistically significant in our analysis. This extreme difference may be attributed to the understanding of the pCLE domain and experts' ability to reject unnaturally enhanced images with possible non-informative domain-specific artefacts.

H7: ZSSR vs SISR

Experts prefer ZSSR (75% of votes) over SISR, and non-experts choose SISR instead (62% of votes). Consequently, with previous results, we observe the same pattern throughout the survey, which confirms the cyclic relationship for the performance of the reconstruction method. This confirms with statistical significance that both tested groups behaved consistently.

The findings indicate that both groups find the quality of learning-based super-resolved image better than baseline interpolation methods, although observers exposed varying opinion of which model is better. Tested videos with their SR versions are included in the appendix Figure 7.4.

Qualitative image assessment

In addition to the user study outcomes, we provide some qualitative assessment of the results. Both DL methods restore images with higher contrast and enhanced details. Also, the triangulation artefact, which is seen on original images reconstructed with the baseline method, is removed by both DL methods. The main distinction between ZSSR and SISR reconstructions is the denoising power of ZSSR. SISR reconstruction contains characteristic pCLE noise that is even enhanced compared to original images, while ZSSR reconstructions are significantly de-noised. Also, ZSSR reconstruction reveals small structures which are not distinguishable from noise. Overall, we find enhanced images easier to interpret, as the biological structures are more apparent when noise and artefacts are removed. Those pixel-level changes are an essential clue for the clinician allowing differentiating cellular structures. We may see the clinical impact of that change is as a more robust source of information for the diagnosis than previously available with standard reconstruction methods.

5.6 Discussion

Thanks to the blind schema of the ZSSR training, we explore several benefits for pCLE videos. Consequently, ZSSR is trained for one fibre, one patient, a patient-specific medical problem, allowing building a model, improving image quality without the interference of external information, which may misinterpret other patients, case, tissue, fibre quickly. We redefined ZSSR pipeline from [189] to adjust its application to pCLE videos which are significantly different to natural images acquired with standard cameras. We made improvements such as a pCLE inspired kernel design, noise simulation and benefiting from the multi-frame training dataset, and also drop some of the proposed initial enhancements not suitable for our application.

We observed that when noise is not taken under consideration, ZSSR enhances all structures on the images, including noise. Following the original work on ZSSR [189], authors have shown that Gaussian noise may help training, yet it does not simulate pCLE noise well. The pCLE noise is characteristic of the local fibre distribution/pattern. We found that accounting for this specific pCLE noise improve ZSSR denoising capabilities. We also found that the amount of noise affects results strongly and need careful consideration when estimated for the training. Differently from the original work, we did not observe improvement from training images for different scales. Thus, we train models for one scale only.

The primary assumption for applying the ZS learning for achieving SR is that inter-occurring patches within image carry enough low-entropy information to learn its representation and consequently improve image quality by fusion of that information. We observed that frames in the pCLE video are correlated, and due to the nature of the procedure, they tend to show only a limited number of tissue types and medical cases, and there are less dynamic than natural image videos often used for entertainment. This in turn facilitates the implementation of ZSSR to the entire video by intuitively expanding the training dataset to several frames from the video. A larger training set created from the video is beneficial for ZS. All the frames in the data-set are associated with the same kernel, yet several frames have together richer content than just the single frame, still maintaining low entropy of information (similar frames). This benefits our pipeline and helps to train and generalise models performing SR reconstructions. Additionally, we also democratise sampling predicted frames known as "HR fathers" from the training set.

We did not use back-projection, an iterative process of projecting details potentially lost in SR image space from LR image space, which was introduced in [57, 6], and also implemented into ZSSR [208]. When the SR image is downscaled, super-resolved details are lost; thus, under an assumption that the downscaling is a perfect estimate of the real downscaling kernel, downscaled SR and the test image should be identical. In back-projection, the difference is used to correct SR reconstruction by propagating back lost details from LR space. Unfortunately, it is not the case for imperfect pCLE images. The difference between SR and LR image, among with informative content, contains the contribution of noise mixed with triangulation pattern, and if propagated, add artefacts to SR image.

The original work on ZSSR [189] is shown to not only have real-time inference time. In the presented adaptation of the ZSSR to pCLE, model complexity and thus inference time are unchanged compared to the original implementation. Additionally, ZSSR can be trained online. I found that extending the amount of training data and allowing the network to be trained for denoising and SR jointly increases the time taken for the model to converge. I also observed that image quality is increasing rapidly during the very first iterations of training and then the change in image quality improvement slows down until loss saturates. That leads to the conclusion that shortening training time may reduce the time and resources of delivering the model without a significant loss in the expected image quality.

5.7 Conclusions

We proposed a novel method for improving the quality of endomicroscopy. Our pipeline combines Zero-Shot learning, which encapsulates a downscaling kernel tailored to the acquisition physics by incorporating fibre bundle geometry and noise simulation. ZSSR is a patient-specific training approach that considers all available information and parameters, such as fibre bundle type and its unique fibre pattern, tissue type, and clinical case. It also makes ZSSR immune to external data, such as another fibre configuration, tissue type, or clinical case, which could potentially be misleading. The specificity of our method is not the only advantage. Its flexibility is deemed of high interest, since the method can be trained with very little data in comparison to state-of-the-art supervised approaches. Limited necessary computational resources make it easier to implement in a real-world clinical set-up, for example, as a reconstruction algorithm using the first few frames for calibration.

The quantitative and qualitative image assessment confirms the superiority of our ZSSR reconstruction method over the baseline interpolation-based reconstruction algorithm. The baseline method does not use any prior information, but just naïve interpolation to reconstruct images. In comparison, DL-based reconstructions use multiple kernels trained on prior information provided by the pCLE dataset. These kernels, which are trained to reconstruct HR from LR images, allow for more informed reconstructions of the images that enhance their quality. ZSSR serves as a more effective alternative to SISR, which is itself restricted in applicability to the real pCLE domain due to the lack of ground truth HR pCLE images.

Chapter 6

Conclusions, limitations and future research lines

Contents

In this dissertation, I have presented research on improving the image quality of endomicroscopy by proposing online, DL-based super-resolution algorithms. In this last Chapter, I highlight the main outcomes and provide a critical view on their benefits and limitations, as well as propose further research lines on the remaining challenges in applying DL-based super-resolution to pCLE.

6.1 Summary of contributions

In this dissertation, pCLE is introduced as a state-of-the-art imaging system used in clinical practice for *in situ* and real-time *in vivo* optical biopsy. In particular, recent works using Cellvizio, developed by Mauna Kea Technologies in France, have demonstrated the positive impact of introducing pCLE as a new imaging modality in gastrointestinal and pancreaticobiliary diseases. Due to a close collaboration with the manufacturer and the availability of the pCLE design details, the developments in this thesis are primarily established based on data acquired by Cellvizio.

The modality is based on the acquisition of the signal as video sequences through an optical fibre bundle and miniaturised optics, that can be integrated into endoscopic and needle-based devices for a wide range of interventional workflows. However, the very nature of the acquisition, via a bundle constructed of thousands of optical fibres, means that there is a significant limitation in the pCLE image quality. This research is conducted in response to the need for improving the clinical readability of the images which is currently hampered by the relatively low image quality. The proposed solutions may help in improving pCLE specificity and sensitivity, and as a result, would help pCLE become a routine diagnostic tool.

The methodical advancements proposed in this work are focused on the geometry of pCLE signal acquisition. In pCLE, the signals produced by each fibre are non-uniformly distributed over the bundle's field of view in a characteristic irregular and a quasi-hexagonal geometrical pattern, where each fibre produces a single-pixel signal. The pixel signals contain both tissue signal and noise. The current pCLE reconstruction algorithm that interpolates noisy pCLE signals onto an oversampled Cartesian grid is known to be sub-optimal. While this resampling allows compensating for artefacts, such as the honeycomb pattern produced by non-Cartesian fibre arrangements, it does not have any denoising properties and introduces edge artefacts caused by linear interpolation across the edges of the underpinning Delaunay triangulation [53].

In the next sections, the outcomes from the testing hypotheses, given in Section 1.6 "Research objective, questions and hypotheses", are presented. The hypotheses were tested in the contribution chapters; here, the reader can find a summary of the results achieved for each hypothesis highlighted at the beginning of the section.

6.1.1 The image quality improvement of endomicroscopy

"The current probe-based confocal laser endomicroscopy reconstruction is suboptimal and does not use any prior information. Using prior information during the reconstruction process is possible and allows for a better representation of reconstructed endomicroscopies."

The main research hypothesis in this thesis

The principal aim of advancing the pCLE experience is to improve its image quality by proposing online enhancement methods alleviating the limitations that arise from the hardware design and the current reconstruction. The focus of the research presented in this thesis is on the benefits of incorporating prior information about pCLE into the reconstruction algorithm in order to augment the oversampled image by increasing the number of the high-frequency, informative pixels. This is achieved by developing dedicated data-driven SR algorithms. Each contribution, presented in this dissertation, is a novel, dedicated to pCLE, SR pipeline that leverages the power of deep learning and incorporates specific details on image acquisition and the physics of the modality. As CNN-based algorithms learn the distribution of training data, providing prior information reduces the uncertainty introduced by the reconstruction process and enables higher quality reconstructions. The quantitative and qualitative assessment of the presented research shows that super-resolved pCLE images have better quality than the one interpolated using the baseline method. Thus, the proposed super-resolution pipelines outperform the currently used in clinical devices gold standard reconstruction.

Validation of the image quality improvement

The experiments are validated quantitatively with the most established metrics such as PSNR and SSIM. I explored several other metrics for IQA of pCLE, such as GCF [34] or LIPIPS [195]. The results coming from those metrics were aligned with SSIM and PSNR and did not provide more specific information on image quality improvement. In addition to the quantitative results confirming the usefulness of pCLE image enhancement, I also received encouraging feedback from our clinical partners, who found the enhanced images easier to interpret. In particular, they saw potentially valuable diagnostic information in enhanced structural details. Besides benchmarking the proposed SR algorithms with common metrics, I run user studies that allow getting the qualitative measure of observer perception and expert opinions. Thanks to the MOS study, presented in Chapter 3 "Single-image super-resolution for endomicroscopy", we get insights into understanding how SR reconstruction is perceived by pCLE experts. While 2AFC, presented in Chapter 5 "Zero-shot super-resolution for endomicroscopy", tested a big number of images in two groups of users, both surveys gave consistent results confirming that SR has a positive impact on perceived pCLE image quality.

6.1.2 Online inference

"Prior information is available and allows for real-time processing of pCLE."

1st specific research hypothesis in this thesis

The prior information is encapsulated in both: data used for training and acquisition physics embedded into either data simulation or the SR architecture and training strategy. Although the training of DL algorithms is offline, the prior information captured from either data or specific for pCLE processing embedded into the training pipeline is available during the inference phase via learnt data representation. Each contribution presented in this thesis allows for an online inference. There is a trade-off between the resolution improvement and the real-time performance of the algorithms. Pushing the resolution forward in real-time by using computational approaches is the main technical requirement for contributed novelties. The latency of the proposed SR reconstruction will in large part depend on the type of GPU used. All proposed networks capable of online performance during inference having up to 16M parameters. Specific inference time depends highly on the IO pipeline - time needed to load images from hard drives into GPU memory. Although all algorithms presented in this thesis are capable of real-time inference, the specific implementation aiming at optimising their speed goes beyond the scope of this thesis. It is worth mentioning that there are highly specialised engineering solutions on how to load data into GPU[1], and how to implement trained models into AI accelerators, which are highly optimised such as FPGA[2] or AI chips [204].

6.1.3 pCLE pseudo ground truth data for EBSR

"A temporal redundancy in the pCLE video sequence allows for the fusion of the temporally aliased high-frequencies to reconstruct detailed frames."

2nd specific research hypothesis in this thesis

CNN-based EBSR has been achieving state-of-the-art image enhancement and has surpassed, by a large margin, the first-generation SR methods. However, the limitation of applying EBSR to pCLE is the lack of high-quality images serving as prior information for the algorithm. There is a lack of HR ground-truth data that are necessary to train supervised models. The methodological contribution, presented in Chapter 3 "Single-image super-resolution for endomicroscopy", is to overcome the problem of missing data by generating the pCLE pseudo-ground-truth data for the training. The novel, realistic data simulation ensures the reduction of the domain gap between synthetic and original images. The frames in the original video sequences are fused with an offline registration algorithm [53]; next, the resulting mosaics are cropped based on the inverse transformation of the mosaicking to obtain HR data. The estimated frames are then downsampled with the proposed method that uses fibre arrangements to build a graph-based LR fibre sig-

[1] https://developer.nvidia.com/blog/gpudirect-storage/

[2] https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/fpga-gpu.html

nal estimator. Those LR signals are, after adding noise, used to reconstruct LR image counterparts. The experiments show that the temporal redundancy in the pCLE video sequence allows for the retrieval of the sub-pixel information, which is a source of resolution and content enrichment for the images. The simulated signal loss that accounts for acquisition physics allows mimicking original pCLE images more realistically than with standard bicubic downsampling. Those pseudo-ground-truth images serve as training data allowing creating models that provide convincing super-resolution reconstruction. The realistic simulation of endomicroscopies that use original data reduces the domain gap and enables using supervised training in application to pCLE. Realistic data often suffers from a lack of continuous information flow between frames because of noise corruption and unstable movements of the probe. The limitation of this solution is the availability of videos that can be registered without suffering from significant misalignment artefacts. In this case, suitable videos are selected manually, which is time-consuming. Additionally, the registration does not super-resolve images uniformly, as it depends on the overlapping content between frames. HR data are obtained by transformation of original images containing triangulation artefacts and interpolated noise, and those are often not reduced by mosaicking. The mentioned artefacts hamper the performance of the model, making it difficult to distinguish signal from uncorrelated noise and artefacts, and they are often visible in the simulated HR images. Strong artefacts that can be mistaken by the algorithm as tissue and enhanced as well, which might be apparent in the SR images that rarely are perfectly denoised. Even with these limitations, these results show that EBSR can be effectively trained with synthetic data and improve the image quality of the original pCLE. The proposed realistic data simulation exploits a temporal redundancy in the pCLE video sequence by using registration. It provides a high-quality estimation of ground truth images that can be used for the training of DL models.

6.1.4 Irregular pCLE signals as CNN input

"Accounting for the fibre pattern of the pCLE bundle in the reconstruction algorithm benefits the quality of the reconstructed endomicroscopies."

3rd specific research hypothesis in this thesis

Since the vast majority of DL techniques, including SISR used for the pCLE reconstruction, rely on Cartesian images, I propose a solution that facilitates using sparse images as the input of the SR CNNs directly, with no prior reconstruction. In Chapter 4 "A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks", a unified, computationally efficient, methodology that generalises Nadaraya–Watson (NW) kernel regression as a part of the DL framework is presented. The proposed NW layer is used as a building block in a dedicated architecture, which allows for kernel optimisation that models a signal dependency in sparse images precisely. The layer captures data sparsity by learning the sparse image representation and propagates it through a network in end-to-end training, allowing ConvNet-like approaches on irregular sampling grids. It eliminates the need for using interpolated pCLE input images with artefacts, such as triangulation edge or interpolated noise. The intrinsic value of NW regression implemented in the DL reconstruction algorithm is in incorporating knowledge of the physical constraints of the pCLE acquisition process and fibre bundle design. To make a systematic comparison between different CNN layers, I used histopathological images as the equivalent of HR images and simulate pCLE-like LR images. This simulation is a variant of the earlier one used in Chapter 3 "Single-image super-resolution for endomicroscopy", as it shares the same core idea but a different source of HR signals. Using the histological data allows avoiding artefacts occurring at the HR images which were an issue with registered original images, while still maintaining similar biological information. Having reliable reference data helps to benchmark the different approaches to pCLE super-resolution. The proposed deep NW outperforms classical NW regression with a hand-crafted Gaussian kernel, by providing sharper and more detailed images. Fortunately, the NW layer performs equally to standard CNN and does not provide any additional advantage in image quality improvement. When models trained with synthetic data are used to restore original images, they provide some improvement, yet it seems less significant to one seen on synthetic data. As the same observation was derived in Chapter 3 “Single-image super-resolution for endomicroscopy”, I conclude that the domain gap, although small, still affects the performance of the models. Adequately, capturing realistic image statistics is essential to drive pCLE image improvement. The proposed layer designed to generalise Nadaraya–Watson (NW) kernel regression, which accounts for the fibre pattern of the pCLE bundle, was shown to bring improvement in application to pCLE reconstruction compared to its classical version but did not yield further improvement as a part of DL architecture.

6.1.5 Blind super-resolution with the realistic pCLE kernel

“A physical model of pCLE inserted into the reconstruction algorithm drives quality improvement of the reconstructed endomicroscopies.”

4rd specific research hypothesis in this thesis

Building on the positive results from the previous section, the reliance on ground-truth images for the training is still the limiting factor. In Chapter 5 “Zero-shot super-resolution for endomicroscopy”, I investigated the adaptation of the SR approach using unsupervised training that does not need HR images for the training. I used the power of zero-shot learning that uses internal statistics from only one image alongside a downsampling kernel to achieve SR. Based on previous learnings that a prior on fibre arrangements helps to simulate pseudo-pCLE images, the design of downsampling kernel uses Voronoi diagrams generated from original fibre positions mapped onto a lower resolution grid. Besides simulating the signal loss based on bundle geometry, the typical pCLE noise is also added to LR fibre signals. Consecutive frames contain similar content, thus video statistics should be more robust and informative for DL modelling than a single image. To exploit the internal statistics in pCLE sequences, the training data are extended from using just one single image to several frames from the video. The conducted ablation study shows that ZSSR benefits from the Voronoi kernel that accounts for noise when compared to one without noise and bicubic kernels. It also achieves competitive results compared to SISR in the quantitative tests, and the user study shows that pCLE experts prefer ZSSR to SISR. This contribution allows overcoming the lack of HR images and relies on a very small training set. It is an advantage in medical imaging applications where limited data is often an issue. Moreover, reliance on the video sequence to improve itself could be seen as safer, since no external patients, tissues or abnormalities are introduced in improving image representation. The downside of the ZS approach is that a new network needs to be created each time during calibration of the probe, just before the examination. Although real-time training without delaying the procedure is possible, the models would be trained and used directly without much control over the automated training process driven by objective metrics and without extensive prior validation by the experts. This could bring uncertainty over the calibration outcomes, and may not be easy to translate into clinical devices. The proposed adaptation of the ZSSR framework, which using a physical model of pCLE that implements Voronoi downscaling and noise simulation, was shown to improve the image quality of super-resolved reconstructions.

6.2 Future perspectives on pCLE SR

As DL-based SR is becoming a more mature research domain, there are plenty of emerging research avenues that could bring further benefits to improving the image quality of pCLE. DL modelling is a data-driven domain. The success of training a highly performing model increases with the amount of data. The major drawback in pCLE is the lack of availability of ground truth data. Although in this thesis it was shown that simulated data might be a viable substitute for real data, the drawback is the difficulty in covering all possible clinical scenarios with their natural occurrence in the population, as the emerging new applications of endomicroscopy are growing constantly and access to clinical data is limited. The supervised SR, in my opinion, maybe interesting to explore, yet the lack of benchmark data will remain a challenge. The most suitable data are only acquired during the clinical procedure, and they might not be easily accessible and extendable for DL.

In this thesis, I showed that making use of temporal information that is available in temporal aliasing of pCLE videos in the proposed simulation of HR images was beneficial for effective SR training. CNNs might be able to directly fuse the same information instead of being captured in simulated data. In my contributions, I have considered each frame in pCLE video independently, yet there is a possibility to switch from spatial convolution to perform spatial-temporal convolution. 3D convolution is more expensive computationally than the 2D one, yet with computational performance growing exponentially, they might be within reach for online processing in application to SR soon.

Future research on improving pCLE image quality has the potential to apply unsupervised learning. Very recent work on blind SR primarily has focused either on GANs or estimation of the downscaling kernels. In this thesis, I showed that the Voronoi-based kernel is more suitable and provides a significantly better result than a bicubic kernel. Although the Voronoi-based kernel encapsulates fibre geometry and noise, there could be more benefits coming from exploiting the non-linearity of fibres' response and consider each fibre with their independent point-spread function that captures fibres cross-coupling and transfer function. Those additional considerations may be added analytically or driven by the DL approach. There is a clear trend among recent works to estimate kernels with DL modelling [196], add iterative corrections of the kernel [199], and even consider multi-spatial kernels for one image [197]. Those kernels can be part of the SR pipeline or used as a plug-in into blind SR algorithms.

I have proposed how to input irregular signals by embedding Nadaray-Watson kernel regression as a novel NW layer into the CNN framework. Although my research has shown that standard CNN is as powerful as the dedicated NW layer in performing information expansion in the network to obtain HR grid reconstruction, there might be the advantage of using this layer in networks that aim to reduce downstream parameters e.g. for the classification task or kernel estimation. The main benefit of that approach is that the network would not require formal reconstruction of the input image, and allow inputting sparse data directly. The SR pipelines presented in this dissertation start from the reconstruction of pre-processed images after the calibration stage, yet there are few works primarily adopting raw images as the input of DL algorithms. It is challenging to achieve mapping between the raw input image and output reference frame, and this alignment is a key to training the SR model. The NW layer defines the geometry of sparse signals and might be well suited and bring an advantage over standard convolution used on raw data. Beyond pCLE, I believe that this research might benefit other applications in which data is defined on a graph structure giving the possibility of transfer any DL methods for regularly sampled data to sparse data. The NW layer is computationally efficient and can readily be adapted to either graph-in/scalar-out or graph-in/image-out.

Improving clinical diagnosis is the long-term goal of the presented work but demonstrating such impact is beyond the scope of this dissertation. The primary focus was to present novel methods to improve pCLE images without reliance on infeasible ground truth and compare them with baseline reconstruction and state-of-the-art methods. The overall specificity and sensitivity of pCLE were not always above the threshold acceptable for clinical demand, and the fundamental assumption is that improved image quality would help pCLE become a widespread tool. Beyond technological advancements, it is crucial to assess the impact of enhanced pCLE images on clinical diagnosis. The clinical study would check the overall experience of the clinicians with new super-resolved pCLE and show whether enhanced images allow for a better understanding and evaluation of pCLE compared to standard images. Such clinical studies are the next stage in bringing SR pCLE into hospitals and regular clinical use.

Chapter 7

Appendix - video reconstructions

Supplementary material contains videos with example reconstructions of real pCLE images created with the proposed SR methods with accompanying reconstruction obtained from the gold standard reference method. The videos are accessible through the URL detailed under each miniature.

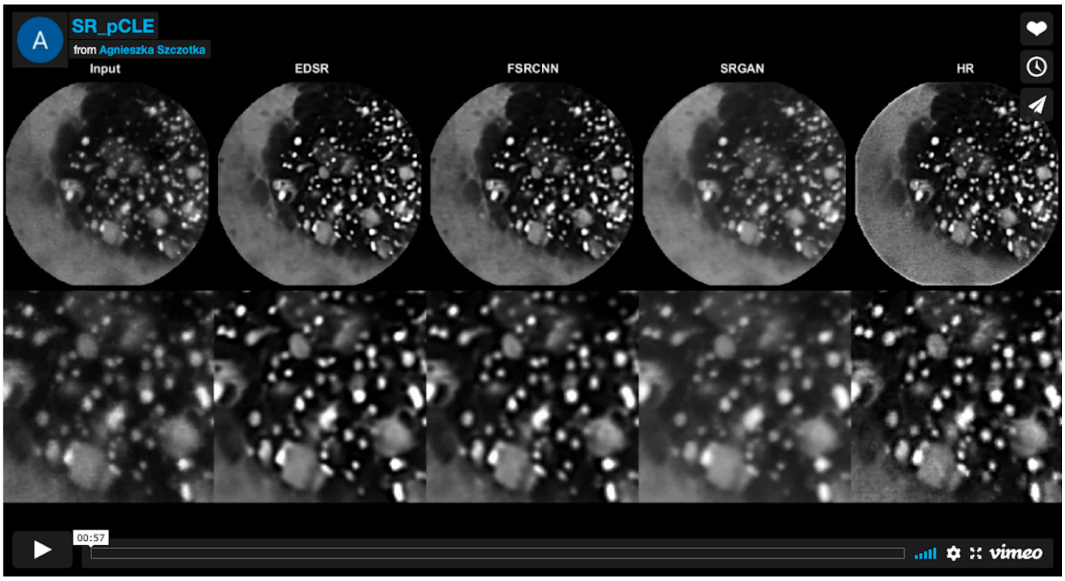
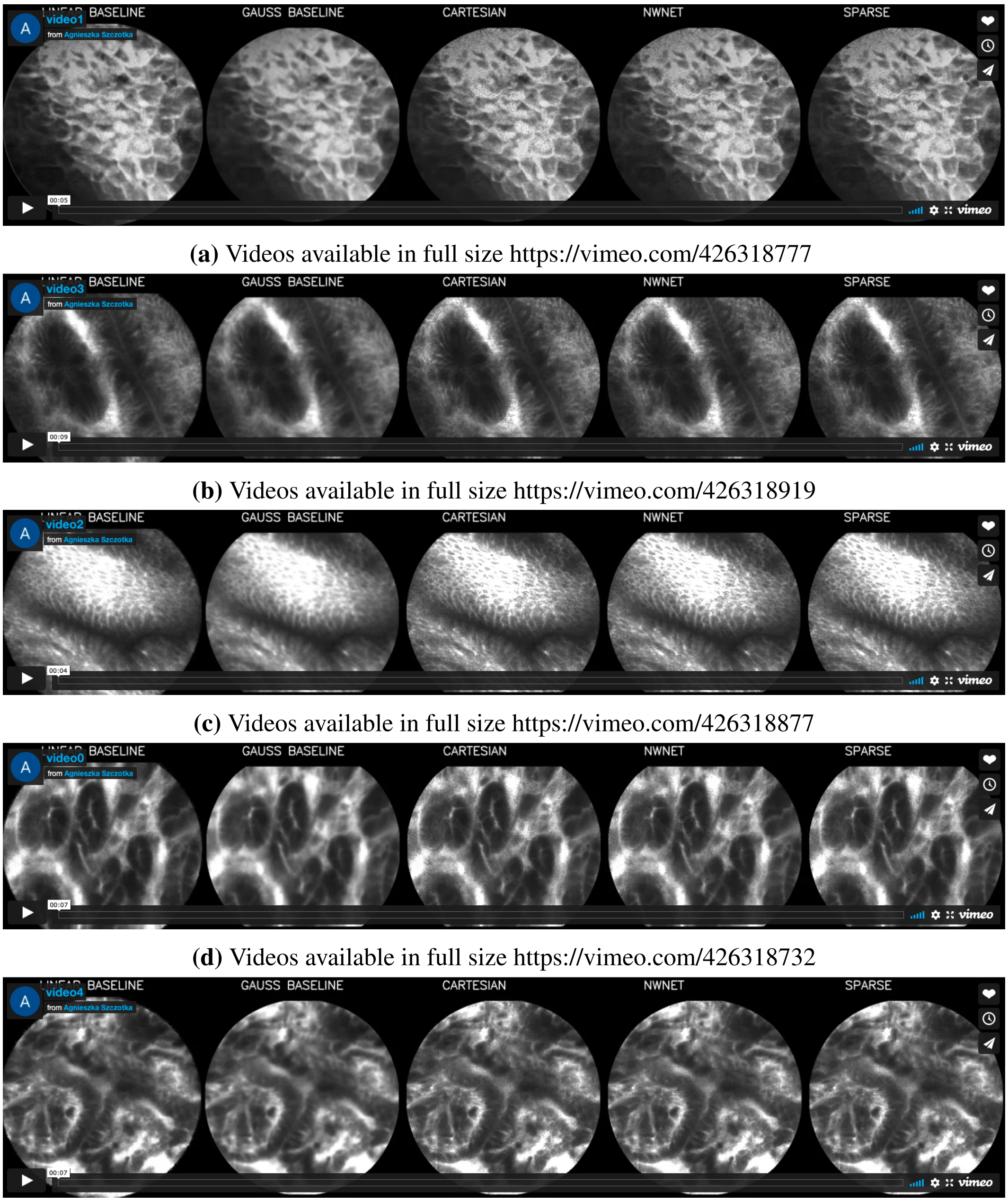


Figure 7.1: Example reconstructions from contribution presented in Chapter 3 ”Single-image super-resolution for endomicroscopy”. Videos available in full size https://vimeo.com/426319061



(a) Videos available in full size https://vimeo.com/426318777

(b) Videos available in full size https://vimeo.com/426318919

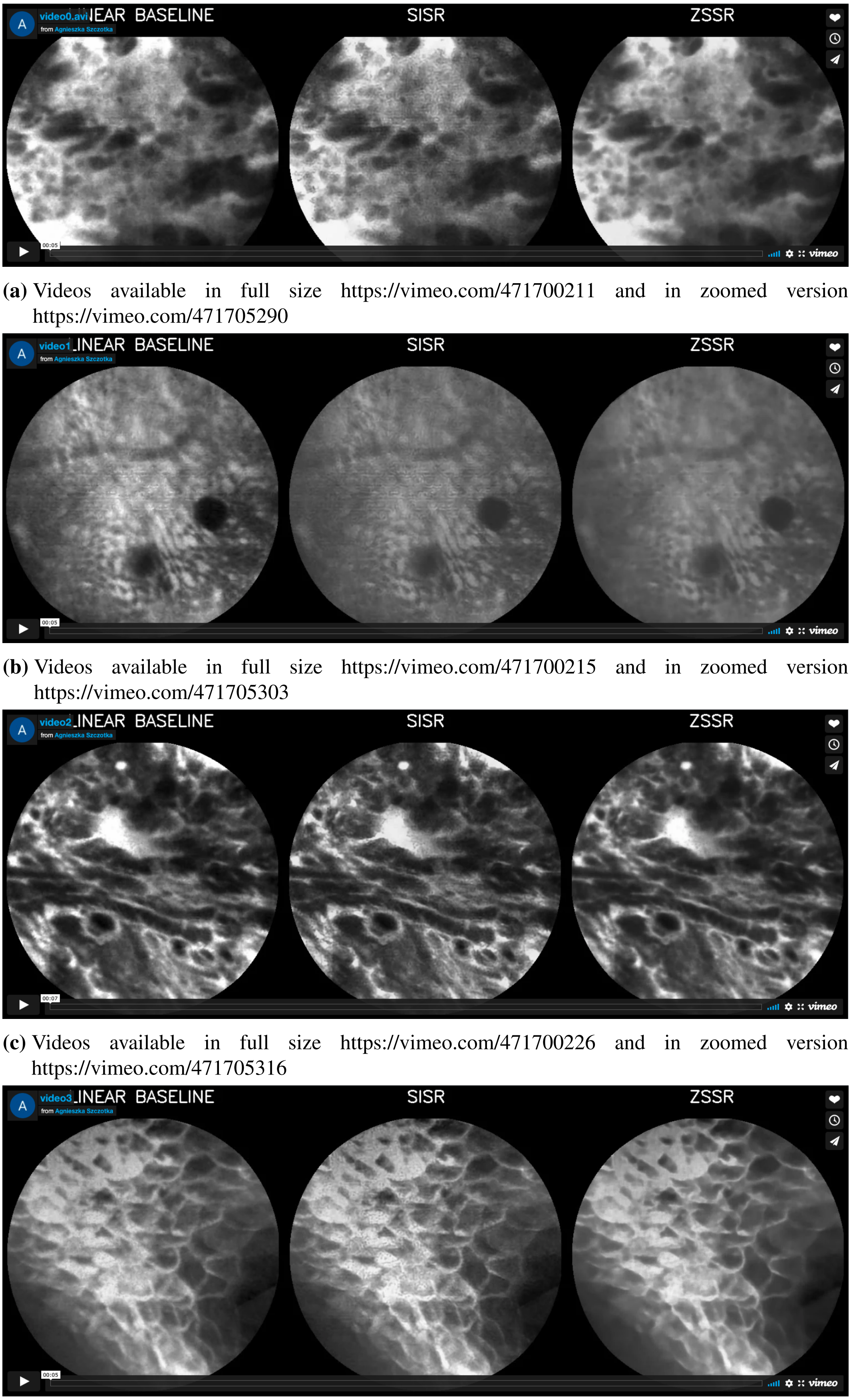
(c) Videos available in full size https://vimeo.com/426318877

(d) Videos available in full size https://vimeo.com/426318732

(e) Videos available in full size https://vimeo.com/426318998

Figure 7.2: Example reconstructions from contribution presented Chapter 4 ”A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks”.

Chapter 4 “A Comparative study of Sparse and Dense approaches for endomicroscpy in convolutional neural networks”



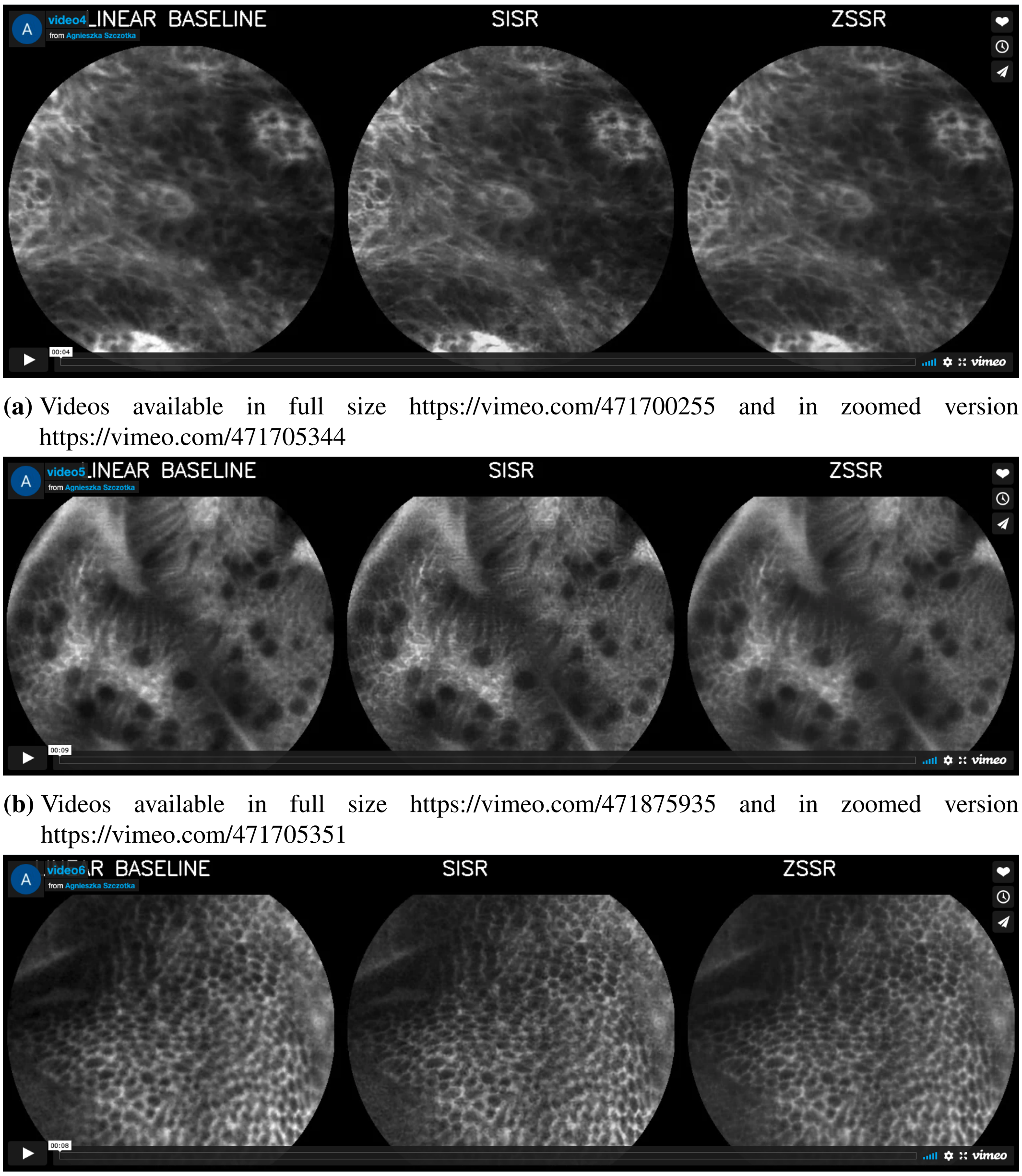
(a) Videos available in full size https://vimeo.com/471700211 and in zoomed version https://vimeo.com/471705290

(b) Videos available in full size https://vimeo.com/471700215 and in zoomed version https://vimeo.com/471705303

(c) Videos available in full size https://vimeo.com/471700226 and in zoomed version https://vimeo.com/471705316

(d) Videos available in full size https://vimeo.com/471875612 and in zoomed version https://vimeo.com/471705322

Figure 7.3: Example reconstructions from contribution presented in Chapter 5 ”Zero-shot super-resolution for endomicroscopy”.

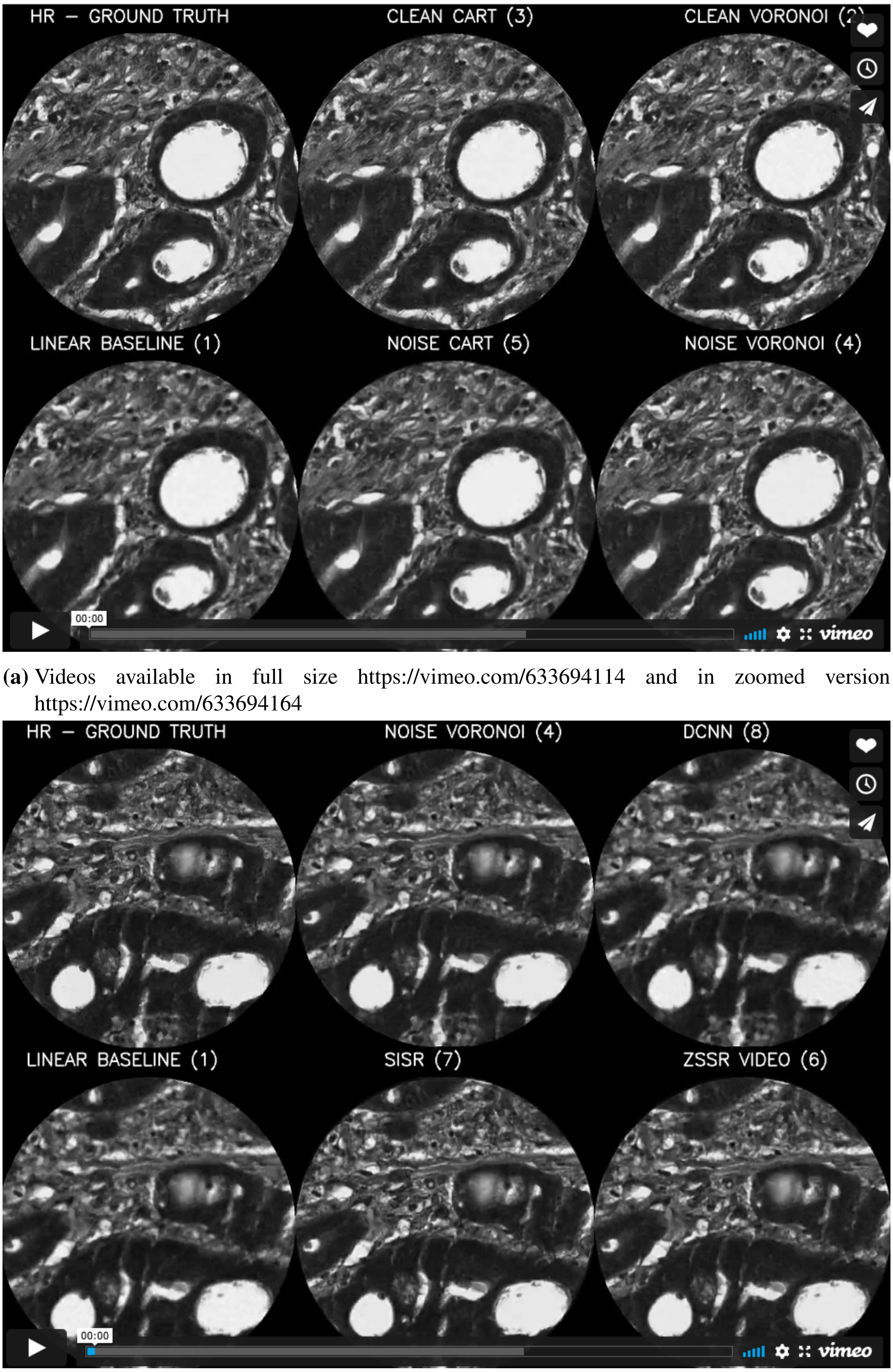


(a) Videos available in full size https://vimeo.com/471700255 and in zoomed version https://vimeo.com/471705344

(b) Videos available in full size https://vimeo.com/471875935 and in zoomed version https://vimeo.com/471705351

(c) Videos available in full size https://vimeo.com/471878136 and in zoomed version https://vimeo.com/471878141

Figure 7.4: Example reconstructions from contribution presented in Chapter 5 ”Zero-shot super-resolution for endomicroscopy”.



(a) Videos available in full size https://vimeo.com/633694114 and in zoomed version https://vimeo.com/633694164

(b) Videos available in full size https://vimeo.com/633694185 and in zoomed version https://vimeo.com/633694385

Figure 7.5: Example synthetic reconstructions from contribution presented Chapter 5 ”Zero-shot super-resolution for endomicroscopy”.

Bibliography

[1] ET Whitaker. “On the functions which are represented by the expansion of interpolating theory”. In: *Proc. Roy. Soc. Edinburgh*. Vol. 35. 1915, pp. 181–194.

[2] Elizbar A Nadaraya. “On estimating regression”. In: *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142.

[3] RY Tsai. “Multiframe image restoration and registration”. In: *Adv. Comput. Vis. Image Process.* 1.2 (1984), pp. 317–339.

[4] Marvin Minsky. “Memoir on inventing the confocal scanning microscope”. In: *Scanning* 10.4 (1988), pp. 128–138.

[5] Michal Irani and Shmuel Peleg. “Improving resolution by image registration”. In: *CVGIP: Graphical Models and Image Processing* 53.3 (1991), pp. 231–239.

[6] Michal Irani and Shmuel Peleg. “Improving resolution by image registration”. In: *CVGIP: Graphical Models and Image Processing* 53.3 (1991), pp. 231–239.

[7] Arthur F Gmitro and David Aziz. “Confocal microscopy through a fiber-optic imaging bundle”. In: *Optics Letters* 18.8 (1993), pp. 565–567.

[8] Ahmet M Eskicioglu and Paul S Fisher. “Image quality measures and their performance”. In: *IEEE Transactions on Communications* 43.12 (1995), pp. 2959–2965.

[9] S. Borman and R. L. Stevenson. “Super-resolution from image sequences-a review”. In: *1998 Midwest Symposium on Circuits and Systems (Cat. No. 98CB36268)*. Sept. 1998, pp. 374–378.

[10] Sean Borman and Robert L Stevenson. “Super-resolution from image sequences-a review”. In: *Circuits and Systems, 1998. Proceedings. 1998 Midwest Symposium on*. IEEE. 1998, pp. 374–378.

[11] Philippe Thevenaz, Urs E Ruttimann, and Michael Unser. “A pyramid approach to subpixel registration based on intensity”. In: *IEEE Transactions on Image Processing* 7.1 (1998), pp. 27–41.

[12] Yashvinder S Sabharwal, Andrew R Rouse, LaTanya Donaldson, Mark F Hopkins, and Arthur F Gmitro. “Slit-scanning confocal microendoscope for high-resolution in vivo imaging”. In: *Applied Optics* 38.34 (1999), pp. 7133–7144.

[13] Christopher R King and John P Long. “Prostate biopsy grading errors: a sampling problem?” In: *International Journal of Cancer* 90.6 (2000), pp. 326–330.

[14] Subhasis Chaudhuri. *Super-resolution imaging*. Vol. 632. Springer Science & Business Media, 2001.

[15] John C Handley. “Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment”. In: *PICS*. Vol. 1. 2001, pp. 108–112.

[16] D Amnon Silverstein and Joyce E Farrell. “Efficient method for paired comparison”. In: *Journal of Electronic Imaging* 10.2 (2001), pp. 394–398.

[17] Carolyn L. Smith. “Basic Confocal Microscopy”. In: *Current Protocols in Neuroscience*. John Wiley and Sons, Inc., 2001.

[18] RECOMMENDATION ITU-R BT. “Methodology for the subjective assessment of the quality of television pictures”. In: *International Telecommunication Union* (2002).

[19] V Dubaj, A Mazzolini, A Wood, and M Harris. “Optic fibre bundle contact imaging probe employing a laser scanning confocal microscope”. In: *Journal of Microscopy* 207.2 (2002), pp. 108–117.

[20] Zhou Wang and Alan C Bovik. “A universal image quality index”. In: *IEEE Signal Processing Letters* 9.3 (2002), pp. 81–84.

[21] Nirag C Jhala, Darshana N Jhala, David C Chhieng, Mohamad A Eloubeidi, and Isam A Eltoum. “Endoscopic ultrasound–guided fine-needle aspiration: a cytopathologist’s perspective”. In: *American Journal of Clinical Pathology* 120.3 (2003), pp. 351–367.

[22] Michael K Ng and Nirmal K Bose. “Mathematical analysis of super-resolution methodology”. In: *IEEE Signal Processing Magazine* 20.3 (2003), pp. 62–74.

[23] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. “Super-resolution image reconstruction: a technical overview”. In: *IEEE Signal Processing Magazine* 20.3 (2003), pp. 21–36.

[24] Michael E Tipping, Christopher M Bishop, et al. “Bayesian image super-resolution”. In: *Advances in Neural Information Processing Systems* (2003), pp. 1303–1310.

[25] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. “Multiscale structural similarity for image quality assessment”. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[26] Barbara Zitova and Jan Flusser. “Image registration methods: a survey”. In: *Image and Vision Computing* 21.11 (2003), pp. 977–1000.

[27] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. “Advances and challenges in super-resolution”. In: *International Journal of Imaging Systems and Technology* 14.2 (2004), pp. 47–57.

[28] Georges Le Goualher, Aymeric Perchant, Magalie Genet, Charlotte Cavé, Bertrand Viellerobe, Fredéric Berier, Benjamin Abrat, and Nicholas Ayache. "Towards optical biopsies with an integrated fibered confocal fluorescence microscope". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2004, pp. 761–768.

[29] Andrew R Rouse, Angelique Kano, Joshua A Udovich, Shona M Kroto, and Arthur F Gmitro. "Design and demonstration of a miniature catheter for a confocal microendoscope". In: *Applied Optics* 43.31 (2004), pp. 5763–5771.

[30] Thomas D Wang and Jacques Van Dam. "Optical biopsy: a new frontier in endoscopic detection and diagnosis". In: *Clinical Gastroenterology and Hepatology* 2.9 (2004), pp. 744–753.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[32] Said E El-Khamy, Mohiy M Hadhoud, Moawad I Dessouky, Bassiouny M Salam, and Fathi E Abd El-Samie. "Regularized super-resolution reconstruction of images using wavelet fusion". In: *Optical Engineering* 44.9 (2005), p. 097001.

[33] Ralf Kiesslich, Martin Goetz, Michael Vieth, Peter R Galle, and Markus F Neurath. "Confocal laser endomicroscopy". In: *Gastrointestinal Endoscopy Clinics* 15.4 (2005), pp. 715–731.

[34] Kresimir Matkovic, László Neumann, Attila Neumann, Thomas Psik, and Werner Purgathofer. "Global Contrast Factor - a New Approach to Image Contrast." In: *Computational Aesthetics* 2005 (2005), pp. 159–168.

[35] Thomas D. Wang. "Confocal microscopy from the bench to the bedside". In: *Gastrointestinal Endoscopy* 62.5 (2005), pp. 696–697.

[36] Hui Ji and Cornelia Fermüller. “Wavelet-based super-resolution reconstruction: theory and algorithm”. In: *European Conference on Computer Vision*. Springer. 2006, pp. 295–307.

[37] John A Kennedy, Ora Israel, Alex Frenkel, Rachel Bar-Shalom, and Haim Azhari. “Super-resolution in PET imaging”. In: *IEEE Transactions on Medical Imaging* 25.2 (2006), pp. 137–147.

[38] Ralf Kiesslich, Liebwin Gossner, Martin Goetz, Alexandra Dahlmann, Michael Vieth, Manfred Stolte, Arthur Hoffman, Michael Jung, Bernhard Nafe, Peter R Galle, et al. “In vivo histology of Barrett’s esophagus and associated neoplasia by confocal laser endomicroscopy”. In: *Clinical Gastroenterology and Hepatology* 4.8 (2006), pp. 979–987.

[39] Hamid R Sheikh and Alan C Bovik. “Image information and visual quality”. In: *IEEE Transactions on Image Processing* 15.2 (2006), pp. 430–444.

[40] J.D. van Ouwerkerk. “Image super-resolution survey”. In: *Image and Vision Computing* 24.10 (2006), pp. 1039–1052.

[41] Tom Vercauteren, Aymeric Perchant, Grégoire Malandain, Xavier Pennec, and Nicholas Ayache. “Robust mosaicing with correction of motion distortions and tissue deformations for *in vivo* fibered microscopy”. In: *Medical Image Analysis* 10.5 (2006), pp. 673–692.

[42] Zhou Wang and Alan C Bovik. “Modern image quality assessment”. In: *Synthesis Lectures on Image, Video, and Multimedia Processing* 2.1 (2006), pp. 1–156.

[43] José M Bioucas-Dias and Mário AT Figueiredo. “A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration”. In: *IEEE Transactions on Image Processing* 16.12 (2007), pp. 2992–3004.

[44] Aggelos K Katsaggelos, Rafael Molina, and Javier Mateos. “Super resolution of images and video”. In: *Synthesis Lectures on Image, Video, and Multimedia Processing* 1.1 (2007), pp. 1–134.

[45] Ralf Kiesslich, Peter R Galle, and Markus F Neurath. *Atlas of endomicroscopy*. Springer Science & Business Media, 2007.

[46] Pierre Parot, Yves F Dufrêne, Peter Hinterdorfer, Christian Le Grimellec, Daniel Navajas, Jean-Luc Pellequer, and Simon Scheuring. "Past, present and future of atomic force microscopy in life sciences and medicine". In: *Journal of Molecular Recognition: An Interdisciplinary Journal* 20.6 (2007), pp. 418–431.

[47] Raja Atreya and Markus F Neurath. "Future trends in confocal laser endomicroscopy: Improved imaging quality and immunoendoscopy". In: *Atlas of Endomicroscopy*. Springer, 2008, pp. 93–100.

[48] Martin Goetz, Ralf Kiesslich, Markus F Neurath, and Alastair JM Watson. "Functional and molecular imaging with confocal laser endomicroscopy". In: *Atlas of Endomicroscopy*. Springer, 2008, pp. 87–92.

[49] Hayit Greenspan. "Super-Resolution in Medical Imaging". In: *The Computer Journal* 52.1 (Feb. 2008), pp. 43–63.

[50] Q. Huynh-Thu and M. Ghanbari. "Scope of validity of PSNR in image/video quality assessment". In: *Electronics Letters* 44.13 (July 2008), pp. 800–801.

[51] Quan Huynh-Thu and Mohammed Ghanbari. "Scope of validity of PSNR in image/video quality assessment". In: *Electronics Letters* 44.13 (2008), pp. 800–801.

[52] HY Liu, YS Zhang, and Song Ji. "Study on the methods of super-resolution image reconstruction". In: (2008).

[53] Tom VERCAUTEREN. "Image Registration and Mosaicing for Dynamic In Vivo Fibered Confocal Microscopy". PhD dissertation. Equipe-Projet Asclepios, INRIA Sophia Antipolis, 2008.

[54] Tom Vercauteren, Alexander Meining, François Lacombe, and Aymeric Perchant. "Real time autonomous video image registration for endomicroscopy: fighting the compromises". In: *Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XV*. Vol. 6861. International Society for Optics and Photonics. 2008, p. 68610C.

[55] Amir Beck and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202.

[56] Kerry B Dunbar, Patrick Okolo, Elizabeth Montgomery, Marcia Irene Canto, and Marcia Irene Canto. "Confocal laser endomicroscopy in Barrett's esophagus and endoscopically inapparent Barrett's neoplasia: a prospective, randomized, double-blind, controlled, crossover trial." In: *Gastrointestinal Endoscopy* 70.4 (Oct. 2009), pp. 645–54.

[57] D. Glasner, S. Bagon, and M. Irani. "Super-resolution from a single image". In: *2009 IEEE 12th International Conference on Computer Vision*. Sept. 2009, pp. 349–356.

[58] Daniel Glasner, Shai Bagon, and Michal Irani. "Super-resolution from a single image". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, pp. 349–356.

[59] Denis Kouame and Marie Ploquin. "Super-resolution in medical imaging: An illustrative approach through ultrasound". In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2009, pp. 249–252.

[60] Vorapoj Patanavijit. "Super-resolution reconstruction and its future research direction". In: *AU Journal of Technology (AU JT)* 12.3 (2009), pp. 149–163.

[61] Geoffrey A Sonn, Sha-Nita E Jones, Tatum V Tarin, Christine B Du, Kathleen E Mach, Kristin C Jensen, and Joseph C Liao. "Optical biopsy of human bladder neoplasia with in vivo confocal laser endomicroscopy". In: *The Journal of Urology* 182.4 (2009), pp. 1299–1305.

[62] Kim-Han Thung and Paramesran Raveendran. “A survey of image quality measures”. In: *2009 International Conference for Technical Postgraduates (TECHPOS)*. IEEE. 2009, pp. 1–4.

[63] Michael B Wallace and Paul Fockens. “Probe-based confocal laser endomicroscopy”. In: *Gastroenterology* 136.5 (2009), pp. 1509–1513.

[64] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. “Variational Bayesian super resolution”. In: *IEEE Transactions on Image Processing* 20.4 (2010), pp. 984–999.

[65] Christine Cavaro-Ménard, Lu Zhang, and Patrick Le Callet. “Diagnostic quality assessment of medical images: Challenges and trends”. In: *2010 2nd European Workshop on Visual Information Processing (EUVIP)*. IEEE. 2010, pp. 277–284.

[66] D. Gheonea, A. Saftoiu, T. Ciurea, C. Popescu, C. Georgescu, and A. Maloş. “Confocal laser endomicroscopy of the colon.” In: *Journal of Gastrointestinal and Liver Diseases : JGLD* 19 2 (2010), pp. 207–11.

[67] U. Gunther, S. Daum, F. Heller, M. Schumann, C. Loddenkemper, M. Grunbaum, M. Zeitz, and C. Bojarski. “Diagnostic value of confocal endomicroscopy in celiac disease”. In: *Endoscopy* 42.03 (Mar. 2010), pp. 197–202.

[68] Alain Hore and Djemel Ziou. “Image quality metrics: PSNR vs. SSIM”. In: *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE. 2010, pp. 2366–2369.

[69] Mingxing Hu, Graeme Penney, Daniel Rueckert, Philip Edwards, Fernando Bello, Michael Figl, Roberto Casula, Yigang Cen, Jie Liu, Zhenjiang Miao, et al. “A robust mosaicing method with super-resolution for optical medical images”. In: *International Workshop on Medical Imaging and Virtual Reality*. Springer. 2010, pp. 373–382.

[70] Kwang In Kim and Younghee Kwon. “Single-image super-resolution using sparse regression and natural image prior”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.6 (2010), pp. 1127–1133.

[71] Matthew Kyrish, Robert Kester, Rebecca Richards-Kortum, and Tomasz Tkaczyk. “Improving spatial resolution of a fiber bundle optical biopsy system”. In: *Endoscopic Microscopy V*. Vol. 7558. International Society for Optics and Photonics. 2010, p. 755807.

[72] Eric Cooper Larson and Damon Michael Chandler. “Most apparent distortion: full-reference image quality assessment and the role of strategy”. In: *Journal of Electronic Imaging* 19.1 (2010), p. 011006.

[73] Susan C. Lester. “6 - Operating Room Consultations”. In: *Manual of Surgical Pathology (Third Edition)*. W.B. Saunders, 2010, pp. 45–66.

[74] Peyman Milanfar. *Super-resolution imaging*. CRC press, 2010.

[75] M Dirk Robinson, Cynthia A Toth, Joseph Y Lo, and Sina Farsiu. “Efficient Fourier-wavelet super-resolution”. In: *IEEE Transactions on Image Processing* 19.10 (2010), pp. 2669–2681.

[76] François Rousseau, Kio Kim, Colin Studholme, Meriam Koob, and J-L Dietemann. “On super-resolution for fetal brain MRI”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2010, pp. 355–362.

[77] M. Salaün, G. Bourg-Heckly, and L. Thiberville. “Endomicroscopie confocale en pneumologie: de la bronche à l’alvéole”. In: *Revue des Maladies Respiratoires* 27.6 (June 2010), pp. 579–588.

[78] Anthony A. Tanbakuchi, Joshua A. Udovich, Andrew R. Rouse, Kenneth D. Hatch, and Arthur F. Gmitro. “In vivo imaging of ovarian tissue using a novel confocal microlaparoscope”. In: *American Journal of Obstetrics and Gynecology* 202.1 (2010), 90.e1–90.e9.

[79] Peng Wang, Rui Ji, Tao Yu, Xiu-Li Zuo, Cheng-Jun Zhou, Chang-Qing Li, Zhen Li, and Yan-Qing Li. “Classification of histological severity of Helicobacter pylori-associated gastritis by confocal laser endomicroscopy.” In: *World Journal of Gastroenterology* 16.41 (Nov. 2010), pp. 5203–10.

[80] J. Yang, J. Wright, T. S. Huang, and Y. Ma. “Image Super-Resolution Via Sparse Representation”. In: *IEEE Transactions on Image Processing* 19.11 (Nov. 2010), pp. 2861–2873.

[81] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. “Image super-resolution via sparse representation”. In: *IEEE Transactions on Image Processing* 19.11 (2010), pp. 2861–2873.

[82] Barbara André, Tom Vercauteren, Anna M Buchner, Michael B Wallace, and Nicholas Ayache. “A smart atlas for endomicroscopy using automated video retrieval”. In: *Medical Image Analysis* 15.4 (2011), pp. 460–476.

[83] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. “Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization”. In: *IEEE Transactions on Image Processing* 20.7 (2011), pp. 1838–1857.

[84] Q. Huynh-Thu, M. N. Garcia, F. Speranza, P. Corriveau, and A. Raake. “Study of Rating Scales for Subjective Quality Assessment of High-Definition Video”. In: *IEEE Transactions on Broadcasting* 57.1 (Mar. 2011), pp. 1–14.

[85] Weisi Lin and C-C Jay Kuo. “Perceptual visual quality metrics: A survey”. In: *Journal of Visual Communication and Image Representation* 22.4 (2011), pp. 297–312.

[86] Jonathan TC Liu, Nathan O Loewke, Michael J Mandella, Richard M Levenson, James M Crawford, and Christopher H Contag. “Point-of-care pathology with miniature microscopes”. In: *Analytical Cellular Pathology* 34.3 (2011), pp. 81–98.

[87] Caroline S Loeser, Marie E Robert, Albert Mennone, Michael H Nathanson, and Priya Jamidar. “Confocal endomicroscopic examination of malignant biliary strictures and histologic correlation with lymphatics”. In: *Journal of Clinical Gastroenterology* 45.3 (2011), p. 246.

[88] Driffa Moussata, Martin Goetz, Annabel Gloeckner, Marcus Kerner, Barry Campbell, Arthur Hoffman, Stephan Biesterfeld, Bernard Flourie, Jean-Christophe Saurin, Peter R Galle, et al. "Confocal laser endomicroscopy is a new imaging modality for recognition of intramucosal bacteria in inflammatory bowel disease in vivo". In: *Gut* 60.1 (2011), pp. 26–33.

[89] Peter E Paull, Benjamin J Hyatt, Wahid Wassef, and Andrew H Fischer. "Confocal laser endomicroscopy: a primer for pathologists". In: *Archives of Pathology & Laboratory Medicine* 135.10 (2011), pp. 1343–1348.

[90] Prateek Sharma, Alexander R. Meining, Emmanuel Coron, Charles J. Lightdale, Herbert C. Wolfsen, Ajay Bansal, Monther Bajbouj, Jean-Paul Galmiche, Julian A. Abrams, Amit Rastogi, Neil Gupta, Joel E. Michalek, Gregory Y. Lauwers, and Michael B. Wallace. "Real-time increased detection of neoplastic tissue in Barrett's esophagus with probe-based confocal laser endomicroscopy: final results of an international multicenter, prospective, randomized, controlled trial". In: *Gastrointestinal Endoscopy* 74.3 (2011), pp. 465–472.

[91] Maximilian J Waldner, Stefan Wirtz, Clemens Neufert, Christoph Becker, and Markus F Neurath. "Confocal laser endomicroscopy and narrow-band imaging-aided endoscopy for in vivo imaging of colitis and colon cancer in mice". In: *Nature Protocols* 6.9 (2011), p. 1471.

[92] M. Wallace, G. Lauwers, Y. Chen, E. Dekker, P. Fockens, P. Sharma, and A. Meining. "Miami classification for probe-based confocal laser endomicroscopy". In: *Endoscopy* 43.10 (Oct. 2011), pp. 882–891.

[93] Katherine Wu, Jen-Jane Liu, Winifred Adams, Geoffrey A Sonn, Kathleen E Mach, Ying Pan, Andrew H Beck, Kristin C Jensen, and Joseph C Liao. "Dynamic real-time microscopy of the urinary tract using confocal laser endomicroscopy". In: *Urology* 78.1 (2011), pp. 225–231.

[94] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. "FSIM: A feature similarity index for image quality assessment". In: *IEEE Transactions on Image Processing* 20.8 (2011), pp. 2378–2386.

[95] Joey M Jabbour, Meagan A Saldua, Joel N Bixler, and Kristen C Maitland. "Confocal endomicroscopy: instrumentation and medical applications". In: *Annals of Biomedical Engineering* 40.2 (2012), pp. 378–397.

[96] 1R Kiesslich, CA Duckworth, D Moussata, A Gloeckner, LG Lim, M Goetz, DM Pritchard, PR Galle, Markus F Neurath, and AJM Watson. "Local barrier dysfunction identified by confocal laser endomicroscopy predicts relapse in inflammatory bowel disease". In: *Gut* 61.8 (2012), pp. 1146–1153.

[97] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems* 25 (2012), pp. 1097–1105.

[98] Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. "Comparison of four subjective methods for image quality assessment". In: *Computer Graphics Forum*. Vol. 31. 8. Wiley Online Library. 2012, pp. 2478–2491.

[99] A Meining, RJ Shah, A Slivka, D Pleskow, R Chuttani, PD Stevens, V Becker, and YK Chen. "Classification of probe-based confocal laser endomicroscopy findings in pancreaticobiliary strictures". In: *Endoscopy* 44.03 (2012), pp. 251–257.

[100] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain". In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.

[101] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. "Making a "completely blind" image quality analyzer". In: *IEEE Signal Processing Letters* 20.3 (2012), pp. 209–212.

[102] Helmut Neumann, Michael Vieth, Raja Atreya, Martin Grauer, Jurgen Siebler, Thomas Bernatik, Markus F. Neurath, and Jonas Mudter. "Assess-

ment of Crohn's disease activity by confocal laser endomicroscopy". In: *Inflammatory Bowel Diseases* 18.12 (Dec. 2012), pp. 2261–2269.

[103] Nicolas Savoire, Barbara André, and Tom Vercauteren. "Online Blind Calibration of Non-uniform Photodetectors: Application to Endomicroscopy". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part III*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 639–646.

[104] Muhammad W. Shahid, Anna M. Buchner, Emmanuel Coron, Timothy A. Woodward, Massimo Raimondo, Evelien Dekker, Paul Fockens, and Michael B. Wallace. "Diagnostic accuracy of probe-based confocal laser endomicroscopy in detecting residual colorectal neoplasia after EMR: a prospective study". In: *Gastrointestinal Endoscopy* 75.3 (2012), 525–533.e1.

[105] W. C. Siu and K. W. Hung. "Review of image interpolation and super-resolution". In: *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. Dec. 2012, pp. 1–10.

[106] Krishnappa Venkatesh, Marta Cohen, Ashraf Abou-Taleb, Steven Thomas, Christopher Taylor, and Mike Thomson. "A new method in the diagnosis of reflux esophagitis: confocal laser endomicroscopy". In: *Gastrointestinal Endoscopy* 75.4 (Apr. 2012), pp. 864–869.

[107] Fabrice Caillol, Bernard Filoche, Monica Gaidhane, and Michel Kahaleh. "Refined probe-based confocal laser endomicroscopy classification for biliary strictures: the Paris Classification". In: *Digestive Diseases and Sciences* 58.6 (2013), pp. 1784–1789.

[108] Damon M Chandler. "Seven challenges in image quality assessment: past, present, and future research". In: *International Scholarly Research Notices* 2013 (2013).

[109] Timothy C Chang, Jen-Jane Liu, Shelly T Hsiao, Ying Pan, Kathleen E Mach, John T Leppert, Jesse K McKenney, Robert V Rouse, and Joseph C Liao. "Interobserver agreement of confocal laser endomicroscopy for bladder cancer". In: *Journal of Endourology* 27.5 (2013), pp. 598–603.

[110] Timothy C Chang, Jen-Jane Liu, and Joseph C Liao. "Probe-based confocal laser endomicroscopy of the urinary tract: the technique". In: *Journal of Visualized Experiments: JoVE* 71 (2013).

[111] Y.-Y. Dong, Y.-Q. Li, Y.-B. Yu, J. Liu, M. Li, and X.-R. Luan. "Meta-analysis of confocal laser endomicroscopy for the detection of colorectal neoplasia". In: *Colorectal Disease* 15.9 (2013), e488–e495.

[112] Florian S Fuchs, Sabine Zirlik, Kai Hildner, Juergen Schubert, Michael Vieth, and Markus F Neurath. "Confocal laser endomicroscopy for diagnosing lung cancer in vivo". In: *European Respiratory Journal* 41.6 (2013), pp. 1401–1408.

[113] Cheon-Yang Lee and Jae-Ho Han. "Elimination of honeycomb patterns in fiber bundle imaging by a superimposition method". In: *Optics Letters* 38.12 (2013), pp. 2023–2025.

[114] Z Li, XL Zuo, CQ Li, CJ Zhou, J Liu, M Goetz, R Kiesslich, KC Wu, DM Fan, and YQ Li. "In vivo molecular imaging of gastric cancer by targeting MG7 antigen with confocal laser endomicroscopy". In: *Endoscopy* 45.02 (2013), pp. 79–85.

[115] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. "Rectifier nonlinearities improve neural network acoustic models". In: *Proceedings ICML*. Vol. 30. 1. Citeseer. 2013, p. 3.

[116] Tomer Michaeli and Michal Irani. "Nonparametric blind super-resolution". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 945–952.

[117] P Su, Y Liu, S Lin, K Xiao, P Chen, S An, J He, and Y Bai. "Efficacy of confocal laser endomicroscopy for discriminating colorectal neoplasms from non-neoplasms: a systematic review and meta-analysis". In: *Colorectal Disease* 15.1 (2013), e1–e12.

[118] Salvador Villena, Miguel Vega, S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. "Bayesian combination of sparse and non-sparse priors in image super resolution". In: *Digital Signal Processing* 23.2 (2013), pp. 530–541.

[119] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index". In: *IEEE Transactions on Image Processing* 23.2 (2013), pp. 684–695.

[120] Yasser M Bhat, Barham K Abu Dayyeh, Shailendra S Chauhan, Klaus T Gottlieb, Joo Ha Hwang, Sri Komanduri, Vani Konda, Simon K Lo, Michael A Manfredi, John T Maple, et al. "High-definition and high-magnification endoscopes". In: *Gastrointestinal Endoscopy* 80.6 (2014), pp. 919–927.

[121] Marcia Irene Canto, Sharmila Anandasabapathy, William Brugge, Gary W Falk, Kerry B Dunbar, Zhe Zhang, Kevin Woods, Jose Antonio Almario, Ursula Schell, John Goldblum, et al. "In vivo endomicroscopy improves detection of Barrett's esophagus–related neoplasia: a multicenter international randomized controlled trial (with video)". In: *Gastrointestinal Endoscopy* 79.2 (2014), pp. 211–221.

[122] Shailendra S Chauhan, Barham K Abu Dayyeh MPH, Yasser M Bhat, Klaus T Gottlieb MBA, Joo Ha Hwang, Sri Komanduri, Vani Konda, Simon K Lo, Michael A Manfredi, John T Maple DO, Faris M Murad, Uzma D Siddiqui, Subhas Banerjee, Michael B Wallace MPH, and Asge Technology Committee. "Confocal laser endomicroscopy". In: *Gastrointestinla Endoscopy* (2014).

[123] Gyeong Woo Cheon, Jaepyeong Cha, and Jin U Kang. "Random transverse motion-induced spatial compounding for fiber bundle imaging". In: *Optics Letters* 39.15 (2014), pp. 4368–4371.

[124] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Learning a deep convolutional network for image super-resolution". In: *European Conference on Computer Vision*. Springer. 2014, pp. 184–199.

[125] Rolf Köhler, Christian Schuler, Bernhard Schölkopf, and Stefan Harmeling. "Mask-specific inpainting with deep neural networks". In: *German Conference on Pattern Recognition*. Springer. 2014, pp. 523–534.

[126] Pedram Mohammadi, Abbas Ebrahimi-Moghadam, and Shahram Shirani. "Subjective and objective quality assessment of image: A survey". In: *arXiv preprint arXiv:1406.7799* (2014).

[127] Kamal Nasrollahi and Thomas B Moeslund. "Super-resolution: a comprehensive survey". In: *Machine Vision and Applications* 25.6 (2014), pp. 1423–1468.

[128] Anant Shinde and Murukeshan Vadakke Matham. "Pixelate removal in an image fiber probe endoscope incorporating comb structure removal methods". In: *Journal of Medical Imaging and Health Informatics* 4.2 (2014), pp. 203–211.

[129] Manoop S Bhutani, Pramoda Koduru, Virendra Joshi, John G Karstensen, Adrian Saftoiu, Peter Vilmann, and Marc Giovannini. "EUS-Guided Needle-Based Confocal Laser Endomicroscopy: A Novel Technique With Emerging Applications." In: *Gastroenterology and Hepatology* 11.4 (Apr. 2015), pp. 235–40.

[130] Bin Cai, Wei Liu, Zhong Zheng, and Zengfu Wang. "A new similarity measure for non-local means denoising". In: *CCF Chinese Conference on Computer Vision*. Springer. 2015, pp. 306–316.

[131] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. “Texture synthesis using convolutional neural networks”. In: *arXiv preprint arXiv:1505.07376* (2015).

[132] Jae-Ho Han, Sang Min Yoon, and Gang-Joon Yoon. “Decoupling structural artifacts in fiber optic imaging by applying compressive sensing”. In: *Optik-International Journal for Light and Electron Optics* 126.19 (2015), pp. 2013–2017.

[133] Jithin Saji Isaac and Ramesh Kulkarni. “Super resolution techniques for medical image processing”. In: *2015 International Conference on Technologies for Sustainable Development (ICTSD)*. IEEE. 2015, pp. 1–6.

[134] Piyapan Prueksapanich, Rapat Pittayanon, Rungsun Rerknimitr, Naruemon Wisedopas, and Pinit Kullavanijaya. “Value of probe-based confocal laser endomicroscopy (pCLE) and dual focus narrow-band imaging (dNBI) in diagnosing early squamous cell neoplasms in esophageal Lugol’s voiding lesions.” In: *Endoscopy International Open* 3.4 (Aug. 2015), E281–8.

[135] Adam Slivka, Ian Gan, Priya Jamidar, Guido Costamagna, Paola Cesaro, Marc Giovannini, Fabrice Caillol, and Michel Kahaleh. “Validation of the diagnostic accuracy of probe-based confocal laser endomicroscopy for the characterization of indeterminate biliary strictures: results of a prospective multicenter international study”. In: *Gastrointestinal Endoscopy* 81.2 (2015), pp. 282–290.

[136] Daniela Ştefănescu, Stephen P Pereira, Margaret Keane, and Adrian Săftoiu. “Needle-based confocal laser endomicroscopy in pancreatic cystic tumors assessment.” In: *Romanian Journal of Morphology and Embryology* 56.4 (2015), pp. 1263–8.

[137] Adam S Wellikoff, Robert C Holladay, Gordon H Downie, Catherine S Chaudoir, Luis Brandi, and Elba A Turbat-Herrera. “Comparison of in vivo probe-based confocal laser endomicroscopy with histopathology in

lung cancer: A move toward optical biopsy”. In: *Respirology* 20.6 (2015), pp. 967–974.

[138] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. “Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation”. In: *arXiv preprint arXiv:1611.05250* (2016).

[139] Li Sze Chow and Raveendran Paramesran. “Review of medical image quality assessment”. In: *Biomedical Signal Processing and Control* 27 (2016), pp. 145–154.

[140] Chao Dong, Chen Change Loy, and Xiaoou Tang. “Accelerating the super-resolution convolutional neural network”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 391–407.

[141] Alessandro Fugazza, Federica Gaiani, Maria Clotilde Carra, Francesco Brunetti, Michael Levy, Iradj Sobhani, Daniel Azoulay, Fausto Catena, Gian Luigi De’Angelis, and Nicola De’Angelis. “Confocal Laser Endomicroscopy in Gastrointestinal and Pancreatobiliary Diseases: A Systematic Review and Meta-Analysis.” In: *BioMed Research International* 2016 (2016), p. 4638683.

[142] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. “Image style transfer using convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2414–2423.

[143] Andrew. Janowczyk and Anant. Madabhushi. “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases”. In: *Journal of Pathology Informatics* 7.1 (2016), p. 29.

[144] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 694–711.

[145] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. “Video super-resolution with convolutional neural networks”. In: *IEEE Transactions on Computational Imaging* 2.2 (2016), pp. 109–122.

[146] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. “Multi-class texture analysis in colorectal cancer histology”. In: *Scientific Reports* 6 (2016), p. 27988.

[147] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. “Accurate image super-resolution using very deep convolutional networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1646–1654.

[148] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Twitter. “Photo-realistic single image super-resolution using a generative adversarial network”. In: *arXiv preprint arXiv:1609.04802* (2016).

[149] Bo Li, Tianlei Zhang, and Tian Xia. “Vehicle Detection from 3D Lidar Using Fully Convolutional Network”. In: *ArXiv* abs/1608.07916 (2016).

[150] Zhen Li, Xiu-Li Zuo, Chang-Qing Li, Zhi-Yan Liu, Rui Ji, Jun Liu, Jing Guo, and Yan-Qing Li. “New classification of gastric pit patterns and vessel architecture using probe-based confocal laser endomicroscopy”. In: *Journal of Clinical Gastroenterology* 50.1 (2016), pp. 23–32.

[151] Aristeo Lopez, Dimitar V Zlatev, Kathleen E Mach, Daniel Bui, Jen-Jane Liu, Robert V Rouse, Theodore Harris, John T Leppert, and Joseph C Liao. “Intraoperative optical biopsy during robotic assisted radical prostatectomy using confocal endomicroscopy”. In: *The Journal of Urology* 195.4 (2016), pp. 1110–1117.

[152] Kouichi Nonaka, Ken Ohata, Shin Ichihara, Shinichi Ban, Yoshimitsu Hiejima, Yohei Minato, Tomoaki Tashima, Yasushi Matsuyama, Maiko Takita, Nobuyuki Matsuhashi, et al. "Development of a new classification for in vivo diagnosis of duodenal epithelial tumors with confocal laser endomicroscopy: a pilot study". In: *Digestive Endoscopy* 28.2 (2016), pp. 186–193.

[153] Yaniv Romano, John Isidoro, and Peyman Milanfar. "RAISR: rapid and accurate image super resolution". In: *IEEE Transactions on Computational Imaging* 3.1 (2016), pp. 110–125.

[154] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1874–1883.

[155] Daniela Ştefănescu, S P Pereira, M M Filip, A Saftoiu, and S Cazacu. "Advanced Endoscopic Imaging Techniques for the Study of Colonic Mucosa in Patients with Inflammatory Bowel Disease." In: *Romanian Journal of Internal Medicine* 54.1 (Jan. 2016), pp. 11–23.

[156] Andrew Sundstrom. *Replication Data for: Histological Image Processing Features Induce a Quantitative Characterization of Chronic Tumor Hypoxia*. Version V1. Harvard Dataverse, 2016.

[157] Radu Timofte, Rasmus Rothe, and Luc Van Gool. "Seven ways to improve example-based single image super resolution". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1865–1873.

[158] Ge Wang. "A perspective on deep imaging". In: *IEEE Access* 4 (2016), pp. 8914–8924.

[159] Zhong Zheng, Bin Cai, Jieting Kou, Wei Liu, and Zengfu Wang. “A Honeycomb Artifacts Removal and Super Resolution Method for Fiber-Optic Images”. In: *International Conference on Intelligent Autonomous Systems*. Springer. 2016, pp. 771–779.

[160] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. “Pixel recursive super resolution”. In: *arXiv preprint arXiv:1702.00783* (2017).

[161] Yan Huang, Wei Wang, and Liang Wang. “Video super-resolution via bidirectional recurrent convolutional networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2017), pp. 1015–1028.

[162] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. “Population based training of neural networks”. In: *arXiv preprint arXiv:1711.09846* (2017).

[163] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. “Enhanced deep residual networks for single image super-resolution”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017.

[164] Peyman Milanfar. *Super-resolution imaging*. CRC press, 2017.

[165] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. “Deep learning for health informatics”. In: *IEEE Journal of Biomedical and Health Informatics* 21.1 (2017), pp. 4–21.

[166] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. “Enhancenet: Single image super-resolution through automated texture synthesis”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4491–4500.

[167] Jaewook Shin, Bryan T Bosworth, and Mark A Foster. "Compressive fluorescence imaging using a multi-core fiber and spatially dependent scattering". In: *Optics Letters* 42.1 (2017), pp. 109–112.

[168] Ryutaro Tanno, Daniel E. Worrall, Aurobrata Ghosh, Enrico Kaden, Stamatios N. Sotiropoulos, Antonio Criminisi, and Daniel C. Alexander. "Bayesian Image Quality Transfer with CNNs: Exploring Uncertainty in dMRI Super-Resolution". In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*. Cham: Springer International Publishing, 2017, pp. 611–619.

[169] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. "Detail-revealing Deep Video Super-resolution". In: *arXiv preprint arXiv:1704.02738* (2017).

[170] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, et al. "NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. July 2017.

[171] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. "NTIRE 2017 challenge on single image super-resolution: Methods and results". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 114–125.

[172] Gian Eugenio Tontini, Jonas Mudter, Michael Vieth, Claudia Günther, Valentina Milani, Raja Atreya, Timo Rath, Andreas Nägel, Giorgia Hatem, Giacomo Carlo Sturniolo, Maurizio Vecchi, Markus F. Neurath, Peter R. Galle, Andrea Buda, and Helmut Neumann. "Prediction of clinical outcomes in Crohn's disease by using confocal laser endomicroscopy: results from a prospective multicenter study". In: *Gastrointestinal Endoscopy* (2017).

[173] Jonas Uhrig*, Nick Schneider*, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. “Sparsity Invariant CNNs”. In: *IEEE International Conference on 3D Vision*. Oct. 10, 2017. published.

[174] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising”. In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

[175] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. “Loss Functions for Image Restoration With Neural Networks”. In: *IEEE Transactions on Computational Imaging* 3.1 (Mar. 2017), pp. 47–57.

[176] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2223–2232.

[177] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. “The 2018 pirm challenge on perceptual image super-resolution”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.

[178] Alberto Breda, Angelo Territo, Andrea Guttilla, Francesco Sanguedolce, Martina Manfredi, Luigi Quaresima, Jose M Gaya, Ferran Algaba, Joan Palou, and Humberto Villavicencio. “Correlation between confocal laser endomicroscopy (Cellvizio®) and histological grading of upper tract urothelial carcinoma: a step forward for a better selection of patients suitable for conservative management”. In: *European Urology Focus* 4.6 (2018), pp. 954–959.

[179] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. “To learn image super-resolution, use a GAN to learn how to do image degradation first”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 185–200.

[180] Woosuk Choi, Mina Kim, Jae HakLee, Jungho Kim, and Jong BeomRa. "Deep CNN-based ultrasound super-resolution for high-speed high-resolution B-mode imaging". In: *2018 IEEE International Ultrasonics Symposium (IUS)*. IEEE. 2018, pp. 1–4.

[181] John P Dumas, Muhammad A Lodhi, Waheed U Bajwa, and Mark C Pierce. "A compressed sensing approach for resolution improvement in fiber-bundle based endomicroscopy". In: *Endoscopic Microscopy XIII*. Vol. 10470. International Society for Optics and Photonics. 2018, p. 1047012.

[182] Ahmed Karam Eldaly, Yoann Altmann, Antonios Perperidis, and Stephen McLaughlin. "Deconvolution of irregularly subsampled images". In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE. 2018, pp. 303–307.

[183] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. "Propagating confidences through CNNs for sparse data regression". In: *arXiv preprint arXiv:1805.11913* (2018).

[184] Jiashen Hua and Xiaojin Gong. "A Normalized Convolutional Neural Network for Guided Sparse Depth Upsampling." In: *International Joint Conferences on Artificial Intelligence*. 2018, pp. 2283–2290.

[185] Saeed Izadi, Kathleen P Moriarty, and Ghassan Hamarneh. "Can Deep Learning Relax Endomicroscopy Hardware Miniaturization Requirements?" In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 57–64.

[186] Saeed Izadi, Kathleen P. Moriarty, and Ghassan Hamarneh. "Can Deep Learning Relax Endomicroscopy Hardware Miniaturization Requirements?" In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, 2018, pp. 57–64.

[187] Daniele Ravì, Agnieszka Barbara Szczotka, Dzhoshkun Ismail Shakir, Stephen P Pereira, and Tom Vercauteren. “Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction”. In: *International Journal of Computer Assisted Radiology and Surgery* 13 (2018), pp. 917–924.

[188] Jianbo Shao, Wei-Chen Liao, Rongguang Liang, and Kobus Barnard. “Resolution enhancement for fiber bundle imaging using maximum a posteriori estimation”. In: *Optics Letters* 43.8 (2018), pp. 1906–1909.

[189] Assaf Shocher, Nadav Cohen, and Michal Irani. “Zero-shot super-resolution using deep internal learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3118–3126.

[190] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. “NTIRE 2018 challenge on single image super-resolution: Methods and results”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 852–863.

[191] Kensuke Umehara, Junko Ota, and Takayuki Ishida. “Application of super-resolution convolutional neural network for enhancing image resolution in chest CT”. In: *Journal of Digital Imaging* 31.4 (2018), pp. 441–450.

[192] Khushi Vyas, Michael Hughes, Bruno Gil Rosa, and Guang-Zhong Yang. “Fiber bundle shifting endomicroscopy for high-resolution imaging”. In: *Biomedical Optics Express* 9.10 (2018), pp. 4649–4664.

[193] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. “Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 701–710.

[194] Kai Zhang, Wangmeng Zuo, and Lei Zhang. “Learning a single convolutional super-resolution network for multiple degradations”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3262–3271.

[195] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *CVPR*. 2018.

[196] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. "Blind super-resolution kernel estimation using an internal-GAN". In: *arXiv preprint arXiv:1909.06581* (2019).

[197] Victor Cornillere, Abdelaziz Djelouah, Wang Yifan, Olga Sorkine-Hornung, and Christopher Schroers. "Blind image super-resolution with spatially variant degradations". In: *ACM Transactions on Graphics (TOG)* 38.6 (2019), pp. 1–13.

[198] John P Dumas, Muhammad A Lodhi, Batoul A Taki, Waheed U Bajwa, and Mark C Pierce. "Computational endoscopy—a framework for improving spatial resolution in fiber bundle imaging". In: *Optics Letters* 44.16 (2019), pp. 3968–3971.

[199] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. "Blind super-resolution with iterative kernel correction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1604–1613.

[200] Shuhang Gu, Martin Danelljan, Radu Timofte, Muhammad Haris, Kazutoshi Akita, Greg Shakhnarovic, Norimichi Ukita, Pablo Navarrete Michelini, Wenbin Chen, Hanwen Liu, et al. "Aim 2019 challenge on image extreme super-resolution: Methods and results". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 3556–3564.

[201] Saeed Izadi, Zahra Mirikharaji, Mengliu Zhao, and Ghassan Hamarneh. "WhiteNNer-Blind Image Denoising via Noise Whiteness Priors". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019.

[202] Saeed Izadi, Darren Sutton, and Ghassan Hamarneh. “Image Super Resolution via Bilinear Pooling: Application to Confocal Endomicroscopy”. In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2019, pp. 236–244.

[203] Margaret G Keane, Natascha Wehnert, Miguel Perez-Machado, Giuseppe K Fusai, Douglas Thorburn, Kofi W Oppong, Nicholas Carroll, Andrew J Metz, and Stephen P Pereira. “A prospective trial of CONfocal endomicroscopy in CYSTic lesions of the pancreas: CONCYST-01”. In: *Endoscopy International Open* 7.9 (2019), E1117.

[204] Bingzhen Li, Jiaojiao Gu, and Wenzhi Jiang. “Artificial Intelligence (AI) chip technology review”. In: *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. IEEE. 2019, pp. 114–117.

[205] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Manuel Fritsche, Shuhang Gu, Kuldeep Purohit, Praveen Kandula, Maitreya Suin, AN Rajagoapalan, Nam Hyung Joon, et al. “Aim 2019 challenge on real-world image super-resolution: Methods and results”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 3575–3583.

[206] Daniele Ravì, Agnieszka Barbara Szczotka, Stephen P Pereira, and Tom Vercauteren. “Adversarial training with cycle consistency for unsupervised super-resolution in endomicroscopy”. In: *Medical Image Analysis* (2019).

[207] Mathieu Salaün, Florian Guisier, Stéphane Dominique, Anne Genevois, Vincent Jounieaux, Emmanuel Bergot, Caroline Thill, Nicolas Piton, and Luc Thiberville. “In vivo probe-based confocal laser endomicroscopy in chronic interstitial lung diseases: Specific descriptors and correlation with chest CT”. In: *Respirology* 24.8 (2019), pp. 783–791.

[208] Jianbo Shao, Junchao Zhang, Xiao Huang, Rongguang Liang, and Kobus Barnard. "Fiber bundle image restoration using deep learning". In: *Optics Letters* 44.5 (2019), pp. 1080–1083.

[209] Jianbo Shao, Junchao Zhang, Rongguang Liang, and Kobus Barnard. "Fiber bundle imaging resolution enhancement using deep learning". In: *Optics Express* 27.11 (2019), pp. 15880–15890.

[210] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. "Flickr1024: A large-scale dataset for stereo image super-resolution". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019.

[211] Lizzy Wijmans, Joey Yared, Daniel M de Bruin, Sybren L Meijer, Paul Baas, Peter I Bonta, and Jouke T Annema. "Needle-based confocal laser endomicroscopy for real-time diagnosing and staging of lung cancer". In: *European Respiratory Journal* 53.6 (2019), p. 1801520.

[212] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. "Handheld multi-frame super-resolution". In: *ACM Transactions on Graphics (TOG)* 38.4 (2019), pp. 1–18.

[213] Xin Yi, Ekta Walia, and Paul Babyn. "Generative adversarial network in medical imaging: A review". In: *Medical Image Analysis* 58 (2019), p. 101552.

[214] Jin Zhu, Guang Yang, and Pietro Lio. "How can we make gan perform better in single medical image super-resolution? A lesion focused multi-scale approach". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1669–1673.

[215] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. "Real-World Super-Resolution via Kernel Estimation and Noise Injection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 466–467.

[216] Margaret G Keane, Kofi W Oppong, and Stephen P Pereira. "Response to letter: Endoscopic ultrasound-guided confocal endomicroscopy requires high-quality imaging and interpretation for diagnostic evaluation of pancreatic cystic lesions". In: *Endoscopy International Open* 8.3 (2020), E312.

[217] Tess Kramer, Lizzy Wijmans, Martijn De Bruin, Peter Bonta, and Jouke Annema. *Bronchoscopic needle based confocal laser endomicroscopy (nCLE) as a real-time detection tool for peripheral lung cancer*. 2020.

[218] Somashekar G Krishna. "Endoscopic ultrasound-guided confocal endomicroscopy requires high-quality imaging and interpretation for diagnostic evaluation of pancreatic cystic lesions". In: *Endoscopy International Open* 8.03 (2020), E310–E311.

[219] Y. Li, B. Sixou, and F. Peyrin. "A review of the deep learning methods for medical images super resolution problems". In: *IRBM* (2020).

[220] Jialin Liu, Wei Zhou, Baoteng Xu, Xibin Yang, and Daxi Xiong. "Honeycomb pattern removal for fiber bundle endomicroscopy based on a two-step iterative shrinkage thresholding algorithm". In: *AIP Advances* 10.4 (2020), p. 045004.

[221] Jialin Liu, Wei Zhou, Baoteng Xu, Xibin Yang, and Daxi Xiong. "Restoration for fiber bundle endomicroscopy using a fast iterative shrinkage-thresholding algorithm". In: *AOPC 2020: Optical Spectroscopy and Imaging; and Biomedical Optics*. Vol. 11566. International Society for Optics and Photonics. 2020, p. 1156612.

[222] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. "NTIRE 2020 challenge on real-world image super-resolution: Methods and results". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 494–495.

[223] Antonios Perperidis, Kevin Dhaliwal, Stephen McLaughlin, and Tom Vercauteren. "Image computing for fibre-bundle endomicroscopy: A review". In: *Medical Image Analysis* 62 (2020), p. 101620.

[224] Carlos Renteria, Javier Suárez, Alyssa Licudine, and Stephen A Boppart. "Depixelation and enhancement of fiber bundle images by bundle rotation". In: *Applied Optics* 59.2 (2020), pp. 536–544.

[225] Amanjot Singh and Jagroop Singh. "Survey on single image based super-resolution—implementation challenges and solutions". In: *Multimedia Tools and Applications* 79.3 (2020), pp. 1641–1672.

[226] Tzu-An Song, Samadrita Roy Chowdhury, Fan Yang, and Joyita Dutta. "Super-resolution PET imaging using convolutional neural networks". In: *IEEE Transactions on Computational Imaging* 6 (2020), pp. 518–528.

[227] Agnieszka Barbara Szczotka, Dzhoshkun Ismail Shakir, Daniele Ravì, Matthew J Clarkson, Stephen P Pereira, and Tom Vercauteren. "Learning from irregularly sampled data for endomicroscopy super-resolution: a comparative study of sparse and dense approaches". In: *International Journal of Computer Assisted Radiology and Surgery* (2020).

[228] Z. Wang, J. Chen, and S. C. H. Hoi. "Deep Learning for Image Super-resolution: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1.

[229] Zhihao Wang, Jian Chen, and Steven CH Hoi. "Deep learning for image super-resolution: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[230] Baoteng Xu, Jialin Liu, Wei Zhou, Daxi Xiong, and Xibin Yang. "Fiber bundle image restoration using conditional generative adversarial network". In: *AOPC 2020: Display Technology; Photonic MEMS, THz MEMS, and Metamaterials; and AI in Optics and Photonics*. Vol. 11565. International Society for Optics and Photonics. 2020, p. 115650X.

[231] *Biopsy*. `https://www.nhs.uk/conditions/biopsy/`. 1 June 2018, Accessed: Dec 3, 2020.

[232] *Cellvizio system*. `http://www.maunakeatech.com/en/cellvizio/11-cellvizio-system`. Accessed: 14 Jan 2018.

[233] *Confocal microscopy*. http://www.wikilectures.eu/w/Confocal_microscopy. Accessed: 2018-01-14.