



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2266*

Representation Learning and Information Fusion

Applications in Biomedical Image Processing

ELISABETH WETZER



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023

ISSN 1651-6214
ISBN 978-91-513-1802-8
URN urn:nbn:se:uu:diva-500386

Dissertation presented at Uppsala University to be publicly examined in Polhemsalen, 10134, Ångström, Lägerhyddsvägen 1, Uppsala, Monday, 12 June 2023 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Fred Hamprecht (Heidelberg University, Department of Physics and Astronomy).

Abstract

Wetzer, E. 2023. Representation Learning and Information Fusion. *Applications in Biomedical Image Processing. Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2266. 85 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1802-8.

In recent years Machine Learning and in particular Deep Learning have excelled in object recognition and classification tasks in computer vision. As these methods extract features from the data itself by learning features that are relevant for a particular task, a key aspect of this remarkable success is the amount of data on which these methods train. Biomedical applications face the problem that the amount of training data is limited. In particular, labels and annotations are usually scarce and expensive to obtain as they require biological or medical expertise. One way to overcome this issue is to use additional knowledge about the data at hand. This guidance can come from expert knowledge, which puts focus on specific, relevant characteristics in the images, or geometric priors which can be used to exploit the spatial relationships in the images. This thesis presents machine learning methods for visual data that exploit such additional information and build upon classic image processing techniques, to combine the strengths of both model- and learning-based approaches. The thesis comprises five papers with applications in digital pathology. Two of them study the use and fusion of texture features within convolutional neural networks for image classification tasks. The other three papers study rotational equivariant representation learning, and show that learned, shared representations of multimodal images can be used for multimodal image registration and cross-modality image retrieval.

Keywords: Representation Learning, Texture Descriptors, Equivariant Neural Networks, Contrastive Learning, Image Classification, Image Registration, Image Retrieval, Digital Pathology

Elisabeth Wetzer, Department of Information Technology, Computerized Image Analysis and Human-Computer Interaction, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

© Elisabeth Wetzer 2023

ISSN 1651-6214

ISBN 978-91-513-1802-8

URN urn:nbn:se:uu:diva-500386 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-500386>)

*Dedicated to all the young girls who dream of becoming a scientist one day
when they grow up*

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I **E. Wetzer**, J. Lindblad, I.M. Sintorn, K. Hultenby, N. Sladoje. "Towards automated multiscale imaging and analysis in TEM: Glomerulus detection by fusion of CNN and LBP maps", Proceedings of the European Conference on Computer Vision (ECCV) Workshops 2018
- II **E. Wetzer**, J. Gay, H. Harlin, J. Lindblad, N. Sladoje. "When texture matters: texture-focused CNNs outperform general data augmentation and pretraining in oral cancer detection", 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
- III N. Pielawski*, **E. Wetzer***, J. Öfverstedt, J. Lu, C. Wählby, J. Lindblad, N. Sladoje. "CoMIR: Contrastive Multimodal Image Representation for Registration", Advances in Neural Information Processing Systems 33 (NeurIPS 2020)
- IV E. Breznik*, **E. Wetzer***, J. Lindblad, N. Sladoje. "Cross-Modality Sub-Image Retrieval using Contrastive Multimodal Image Representations", submitted manuscript under review
- V **E. Wetzer**, J. Lindblad, N. Sladoje. "Can representation learning for multimodal image registration be improved by supervision of intermediate layers?", submitted manuscript under review

Reprints were made with permission from the publishers.

* Authors contributed equally

Summary of Contributions

The Roman numerals correspond to the numbers in the list of papers.

- I I am the primary author. I acquired the image data together with I.M. Sintorn. I labelled and designed the dataset. Medical expertise regarding the data was provided by K. Hultenby. The experimental design and method development was a joint collaboration with I.M. Sintorn, J. Lindblad and N. Sladoje. All authors participated in the writing of the manuscript.

- II I am the primary author. I implemented the methods together with J. Gay and H. Harlin. All authors contributed to the study's design and participated in the writing of the manuscript.

- III I contributed equally with N. Pielawski to the study's conception, design, conduct, analysis, and the interpretation of results. I wrote the code to create the registration dataset with input from J. Öfversted and performed the analysis of the image registration techniques. N. Pielawski wrote the code for the representation learning, J. Lu trained the GAN models in the study. All authors participated in discussions and the writing of the manuscript.

- IV I contributed equally with E. Breznik to the study's design, conduct, analysis, and the interpretation of results. All authors participated in discussions. E. Breznik and I wrote the paper with input from the co-authors.

- V I am the primary author. I conceived the main idea, implemented the code and analyzed the results. All authors participated in discussions. I wrote the paper with input from the co-authors.

Related Work

In addition to the papers included in this thesis, I have also contributed to the following works:

Peer-reviewed Abstracts

- R1 **E. Wetzer**, E. Breznik, J. Lindblad, N. Sladoje. "Re-Ranking Strategies in Cross-Modality Microscopy Retrieval", International Symposium on Biomedical Imaging (ISBI) 2022

Non peer-reviewed Conference Papers

- R2 **E. Wetzer**, J. Lindblad, N. Sladoje. "Partial Dimensional Collapse in Contrastive Learning using Intermediate Layers", Swedish Symposium on Image Analysis (SSBA) 2023
- R3 N. Pielawski, J. Öfverstedt, **E. Wetzer**. "Global Multimodal Image Registration using Gaussian Processes", Swedish Symposium on Image Analysis (SSBA) 2023
- R4 **E. Wetzer**, N. Pielawski, J. Öfverstedt, J. Lu, C. Wählby, J. Lindblad, N. Sladoje. "Rotationally Equivariant Representation Learning for Multimodal Images", Swedish Symposium on Image Analysis (SSBA) 2022
- R5 E. Breznik, **E. Wetzer**, J. Lindblad, N. Sladoje. "Label-Free Reverse Image Search of Multimodal Microscopy Images", Swedish Symposium on Image Analysis (SSBA) 2022
- R6 J. Gay, H. Harlin, **E. Wetzer**, J. Lindblad, and N. Sladoje. "Oral Cancer Detection: A Comparison of Texture Focused Deep Learning Approaches", Swedish Symposium on Image Analysis (SSBA) 2019

Non peer-reviewed Conference Abstracts

- R7 **E. Wetzer**, N. Pielawski, E. Breznik, J. Öfverstedt, J. Lu, C. Wählby, J. Lindblad, and N. Sladoje. "Contrastive Learning for Equivariant Multimodal Image Representations", "The Power of Women in Deep Learning" at the Mathematics of Deep Learning Programme at the Isaac Newton Institute 2021
- R8 **E. Wetzer**, N. Pielawski, J. Öfverstedt, Jiahao Lu, C. Wählby, J. Lindblad, N. Sladoje. "Registration of Multimodal Microscopy Images using CoMIR - learned structural image representations", Correlated Multimodal Imaging in Life Sciences (COMULIS) Conference 2021
- R9 **E. Wetzer**, N. Pielawski, J. Öfverstedt, J. Lindblad, I. Floroiu, A. Dumitru, M. Costache, R. Hristu, S.G. Stanciu, N. Sladoje. "Cross-modal Representation Learning for Efficient Registration of Multiphoton and Brightfield Microscopy Images of Skin Tissue", Network of European Biomimage Analysts (NEUBIAS) Conference 2020
- R10 N. Koriakina, N. Sladoje, **E. Wetzer**, and J. Lindblad. "Uncovering hidden reasoning of convolutional neural networks in biomedical image classification by using attribution methods", Network of European Biomimage Analysts (NEUBIAS) Conference 2020
- R11 **E. Wetzer**, J. Gay, H. Harlin, J. Lindblad and N. Sladoje. "Texture-based oral cancer detection: A performance analysis of deep learning approaches." Network of European Biomimage Analysts (NEUBIAS) Conference 2019
- R12 **E. Wetzer**, J. Lindblad, I.M. Sintorn, K. Hultenby and N. Sladoje. "Towards automated multiscale Glomeruli detection and analysis in TEM by fusion of CNN and LBP maps", Network of European Biomimage Analysts (NEUBIAS) Conference 2019
- R13 **E. Wetzer**, J. Lindblad, I.M. Sintorn, K. Hultenby and N. Sladoje. "Towards automated multiscale imaging and analysis in TEM: Glomeruli detection by fusion of CNN and LBP maps", Swedish Symposium on Deep Learning (SSDL) 2018
- R14 I.M. Sintorn, A. Suveer, A. Dragomir, K. Hultenby, **E. Wetzer**, K. Lidayová, N. Sladoje, J. Lindblad, M. Ryner. "Facilitating Ultrastructural Pathology through Automated Imaging and Analysis", Proceedings of the 14th European Congress on Digital Pathology (ECDP) 2018

Contents

Acronyms	11
1 Acknowledgements	13
2 Introduction	15
2.1 Objectives	16
2.2 Thesis Outline	19
3 Background	20
3.1 Microscopy	20
3.1.1 Visible-light Microscopy	20
3.1.2 Electron Microscopy	22
3.1.3 Datasets used for Evaluation	23
3.2 Image Processing Tasks	27
3.2.1 Image Classification	27
3.2.2 Image Registration	28
3.2.3 Image Retrieval	32
3.2.4 Image Generation	34
3.3 Image Representations	37
3.3.1 Learned Features vs. Hand-Crafted Features	37
3.3.2 Deep Learning	37
3.3.3 Texture Features	38
3.3.4 Geometric Deep Learning	41
3.4 Learning Embedding Spaces	46
3.4.1 Multidimensional Scaling	46
3.4.2 Contrastive Learning	47
4 Contributions	49
4.1 Short Summary of Papers	49
4.1.1 Paper I	49
4.1.2 Paper II	49
4.1.3 Paper III	49
4.1.4 Paper IV	50
4.1.5 Paper V	50
4.2 Representation Learning with Applications to Multimodal Image Registration and Retrieval – Papers III-V	51
4.2.1 Contrastive Multimodal Image Representations for Multimodal Registration and Cross-Modality Image Retrieval – Paper III - V	52

4.3	Information Fusion – Regulating Feature Properties to Aid Data-Driven Approaches	58
4.3.1	Focus on Texture Features – Papers I & II	58
4.3.2	Equivariant and Invariant Features – Papers II - V	60
4.3.3	Fusion Strategies for Heterogeneous Input – Papers I & V	65
5	Conclusions	69
5.1	Successive Related and Future Work	70
	Sammanfattning på Svenska	73
	Bibliography	75

Acronyms

α -AMD	α -cut based Average Minimal Distance
AI	Artificial Intelligence
ANN	Artificial Neural Network
BF	Bright-Field
BoW	Bag of Words
CBIR	Content-Based Image Retrieval
CDF	Cummulative Distribution Function
CDIR	Cross-Domain Image Retrieval
CL	Contrastive Learning
CMIR	Cross-Modality Image Retrieval
CNN	Convolutional Neural Network
CoMIR	Contrastive Multimodal Image Representation
CT	Computed Tomography
CycleGAN	Cycle-consistent Generative Adversarial Network
DCLGAN	Dual Contrastive Learning Generative Adversarial Network
DL	Deep Learning
DRIT	Disentangled Representation for Image-to-Image Translation
EM	ElectroMagnetic
FID	Fréchet Inception Distance
FM	Fluorescence Microscopy
FN	False Negatives
FoV	Field of View
FP	False Positives
GAN	Generative Adversarial Network
Gc-GAN	Geometry-consistent Generative Adversarial Network
GPU	Graphics Processing Unit
H&E	Hematoxylin-Eosin
I2I	Image-to-Image
InfoNCE	Info Noise Contrastive Estimation

KID	Kernel Inception Distance
LBCNN	Local Binary Convolutional Neural Network
LBP	Local Binary Pattern
MDS	MultiDimensional Scaling
MI	Mutual Information
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NIR	Near-InfraRed
NN	Neural Network
Pap	Papanicolaou
PCC	Pearson Correlation Coefficient
QPI	Quantitative Phase Imaging
RANSAC	RANdom SAmples Consensus
ReLU	Rectified Linear Unit
RGB	Red Green Blue
RIS	Reverse Image Search
SHG	Second Harmonic Generation
SIFT	Scale-Invariant Feature Transform
SSIM	Structural Similarity Index Measure
STEW	Shorter-Than-Excitation Wavelength
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
TEM	Transmission Electron Microscopy
TMA	Tissue MicroArray
TN	True Negatives
TP	True Positives
WSI	Whole Slide Image
XAI	Explainable Artificial Intelligence

1. Acknowledgements

First of all, I want to thank my supervisors Nataša Sladoje, Joakim Lindblad and Ida-Maria Sintorn for their continuous support, many insightful discussions, intellectual guidance, and the many hours they spent with me brainstorming, considering ideas, giving feedback, sharing their wisdom on paper writing, reviewing and supervising. Thank you for always being there and supporting – even in the middle of the night!

I would like to pay my special regards to the entire MIDA group – its current members as well as former ones: Nataša Sladoje, Joakim Lindblad, Johan Öfverstedt, Nadezdha Koriakina, Swarnadip Chatterjee, Marc Fraile Fabrega, Love Nordling, Karl Bengtsson Bernander, Teo Asplund, Jiahao Lu, and Jo Gay. I will miss our Monday meetings very much!

I'm extremely grateful to have had the chance to work with so many fantastic researchers on joint studies: Johan Öfverstedt, Nicolas Pielawski, Eva Breznik, Jiahao Lu, Jo Gay, Hugo Harlin, Kjell Hultenby, Carolina Wählby, Ida-Maria Sintorn, Joakim Lindblad, and Nataša Sladoje. It was a pleasure to work together on the studies that constitute this thesis!

I'd also like to acknowledge the help and patience of Kristina Lidayová, Ida-Maria Sintorn, Kjell Hultenby and Håkan Wieslander in hours of imagining together at Vironova AB and Karolinska Institute, which I enjoyed to the fullest extent.

I am grateful to the seniors in our division who create such a friendly and social work environment, and always take the time to share their insight and knowledge: Robin Strand, Carolina Wählby, Ida-Maria Sintorn, Nataša Sladoje, Joakim Lindblad, Ingela Nyström, Filip Malmberg, Gunilla Borgefors, Ewert Bengtsson, Åsa Cajander, Orçun Göksel, Petter Ranefall, Anna Klemm, Christophe Avenel, Lars Oestreicher, Mikael Laaksoharju, Anders Hast, Stefan Seipel and Christer Oscar Kiselman.

Special thanks goes to my fellow Ph.D. Students, many of whom already graduated: Eva Breznik, Johan Öfverstedt, Raphaela Heil, Nicolas Pielawski, Håkan Wieslander, Axel Andersson, Ankit Gupta, Nadezdha Koriakina, Marc Fraile, Love Nordling, Swarnadip Chatterjee, Andrea Behanova, Can Deniz Bezek, Teo Asplund, Leslie Solorzano, Erik Hallström, Eduard Chelebian, Karl Bengtsson Bernander, Gabriele Partel, Amit Suveer, Kalyan Ram Ayyala-somayajula, Sajith Kecheril Sadanandan, Anders Persson, Anindya Gupta, Kimmo Kartasalo, Fredrik Nysjö, Thomas Wilkinson, Natalya Calvo, Mengyu

Zhong, Philip Harrison, and especially to my office mates, for their brainstorming sessions at the white boards, coding help, honest feedback, encouragement and many cookies. I'd also like to acknowledge Giorgia Milli, Mariëlle Jansen, Bedour Alshaigy, Diana Malakhova and Johan Snider for the company during this exciting journey.

I want to shout out to my Ph.D. thesis writing group: Raphaela Heil, Eva Breznik, Nicolas Pielawski, Virginia Grande Castro, Natalia Calvo Barajas and Ragnar Seton. It was lovely to share this experience with you, thank you for all the tips and tricks along the way!

I am very grateful for the good times I have had writing and designing the SSBA newsletter for multiple years with Raphaela Heil, Teo Asplund, Fredrik Nysjö, Andrea Behanova and Can Deniz Bezek.

I'd also like to thank Robin Strand, Åsa Cajander, Kajsa Örvjåvik, and Victor Kuismin who I had the pleasure to work with during my time as the division's communication officer for interesting aspects you taught me regarding the departmental organization and visibility of the division.

Furthermore I'd also like to express my gratitude to CIM, not only for supporting my PhD, many conference participations, organizing interesting CoSy seminars, and events with fellow CIM PhD students, but in particular Jörgen Östensson for his support and the insight to academic administration I have gotten by joining many CIM board meetings.

I'd also like to thank Katharina Schäfer and Paul Marjoram for their mentoring.

Thank you, Sean Searle, Edvin Norén, Li Hedenmalm and Igor Tominec for your company and encouragement in this period of my life!

Lastly, I want to thank my family for their unlimited support and belief in me, especially my mother who always supported my decisions on this adventure and never ceased to let me know she was proud of me, my sisters, Beate, and my family in Sweden who have continuously encouraged me and shown interest in my research and studies.

I would like to express my deepest gratitude to Regina Wetzer and Dean Pentcheff for introducing me to science, always believing in me, supporting me, and being my biggest role models.

Thank you, Ragnar – without you, this endeavor would not have been possible. I have no words to express my gratitude to all the ways in which you have supported me in those years – you are my favorite person in the world to discuss sciency stuff with, make research plans with, discuss programming languages with, push through all-nighters with, and learn coding tricks from.

Finally, I want to thank the most special person in my life – Eíra who finally started to sleep a bit better during the nights just as I started writing this thesis. You are my inspiration and have made me stronger and happier than I have ever been. Your encouragement ("Jätteduktig, Mama!") is just what I need on a tough day.

2. Introduction

During the time of writing this thesis, breakthroughs in artificial intelligence (AI) and machine learning (ML) are making broad news ever so regularly. While the fear among some people outside the scientific community that AI will take over the world has pertained since science fiction movies in the 80s such as Terminator came out, scientists have always put these concerns into perspective and reassured that we were far from any AI becoming "conscious". Although this unarguably holds true, it is also a fact that AI-generated data more and more often can fool not just ordinary people, but also experts and scientists, which poses new risks and dangers for society as a whole. To name some recent examples: AI-generated videos of politicians, such as Kyiv's current mayor Vitali Klitschko, have been used in video calls to other politicians in the European Union to obtain sensitive information, a fake that was not immediately recognized by all parties involved (e.g. in a talk with Vienna's mayor Michael Ludwig [100]). Another example is the recent release of Chat-GPT. As it turns out, this AI text-generating tool can produce abstracts that even fool scientists into believing they were written by humans, thereby shaking the grounds of current research verification procedures such as peer review [23].

The reason for the remarkable results of these AI systems lies in the amount of data they train on. To perform well, a video-generating AI has to train on thousands or even millions of gigabytes of videos of people, and particularly of the person it should mimic. This large amount of training data is usually available, especially for people of public interest. There is a nearly unlimited amount of visual data available online resulting from the millions of photos and videos uploaded to Instagram, TikTok, and other social media, which can be used to train machine learning algorithms, may it be legally obtained, with respect to copyright, or not. Similarly, the amount of text available to train on is effectively unbounded, as vast amounts are produced by billions of users of social media around the world every second.

Despite these advances in state-of-the-art methods, there is a significant gap in exploiting the full potential of machine learning, and in particular deep learning, for biomedical applications. The main reason is the limited availability of training images in biomedical applications. Obtaining annotations and labels for the data-hungry models is very expensive as they require expert knowledge, i.e., have to be provided by cytologists, radiologists, or pathologists. This is not only expensive but also time-consuming – time that could be spent on a patient. Another limiting factor is that medical images can often

only be shared in a restricted way as they may be hard to render truly anonymous (e.g. in the case of magnetic resonance imaging head scans [117] or genetic data). Nonetheless, there is great potential in employing data-driven methods in life sciences and medicine. One way to overcome the issue of limited access to annotated images and the data greed of machine learning is to use additional information about the data at hand to assist the AI and make the learning problem easier and therefore less data hungry. This additional help can include using expert knowledge to focus on specific characteristics in the images which are relevant to the task or to exploit spatial relationships in the images and regulate the search space of functions to fit the problem. We have a lot of prior knowledge stemming from classic image processing methodology, which sometimes tends to be overlooked within the excitement of the extremely rapidly changing and evolving field of deep learning.

2.1 Objectives

This work aims to establish machine learning methods for visual data that exploit prior knowledge and context of the task at hand to combine the strengths of classic image processing methods with the power of machine learning. A critical aspect of these methods is their general applicability for a variety of different data, i.e., that they are not tailor-made to solve one particular problem only.

This thesis comprises five papers, of which two address image classification tasks (Papers I & II) and three rotational equivariant representation learning for multimodal images (Papers III-V). All studies share the aspect of finding a good balance between constraining the properties of learned image features and emphasizing properties of prior known importance. The methods developed and studied in this thesis are evaluated on (but not limited to) biomedical datasets, in particular, on images acquired by different microscopy techniques.

A primary application for such methods is in digital pathology. In digital pathology microscopy images of specimens are digitized, and the resulting digital slides can be viewed and analyzed by human experts, but also by using image processing for data extraction [11, 30, 58]. The field is an intricate interplay between medical experts, imaging specialists, and computer scientists. Computational means can alleviate more efficient handling of large datasets or make them searchable based on keywords or similarities between images. They can also directly optimize the image acquisition or even make automated acquisition feasible in the first place by detecting relevant regions to image. Automated imaging and preprocessing are key to installing large-scale screening programs for diseases such as cancer by flagging patients whose samples differ from the mass of healthy patients in one way or another. Healthcare resources and expert time can then be focused on these patients under suspicion of abnormalities. Furthermore, explainable AI (XAI) even has the potential

to isolate patterns in images relevant to a diagnosis which doctors and clinicians have so far overlooked. When developing image processing and machine learning methods for digital pathology, it is essential to collaborate closely with medical experts for reasonable interpretation of results and to guide the design choices of the methods with their expertise. Smart model design is particularly crucial when datasets are limited, as typical in this application field. Such models can be formulated by exploiting structural relationships present in the images, as done in geometric deep learning, or by adopting reliable image feature descriptors from classic image processing such that they can either be extracted inside networks, or be fused with learned features in a suitable way. The work in this thesis addresses such synergies for different tasks in digital pathology, which are in need of solutions to reduce manual labor of tedious work to allow experts to spend their time on relevant clinical diagnoses that cannot be replaced safely by a machine.

Figure 2.1 gives a graphical overview of the interdisciplinary field of digital pathology, its workflows, and contact points on which the papers in this thesis aim to provide solutions to current needs. The papers address tasks in the workflow in the following order:

- Paper I proposes a method aimed towards automated image acquisition of regions of interest in nephropathology by using a transmission electron microscope incorporating *a priori* knowledge about texture,
- Paper IV uses learned representations for multimodal images and proposes a pipeline combining them with powerful image descriptors from classic image processing to provide a pipeline that retrieves corresponding images acquired in one modality given an image in another modality. This cross-modal image retrieval task can serve as a first step in a registration pipeline to find matching pairs of multimodal images that show the same area of the imaged sample,
- Paper III introduces these aforementioned multimodal image representations and shows that they can be used to align images captured by different sensors to fuse their information and extract complementary information not present in a single modality. This alignment step usually serves as a preprocessing step for further analysis of the acquired images with respect to subsequent biomedical interpretation (e.g., diagnosis or cancer grading),
- Paper II, uses *a priori* knowledge about texture for image classification to identify potential oral cancer patients based on images of non-invasive oral swabs to roll out cancer screening programs with the help of AI,
- Paper V builds upon the theoretical machine learning aspects used in Paper III, aiming to improve the multimodal image representations used in Papers III & IV for subsequent downstream tasks.

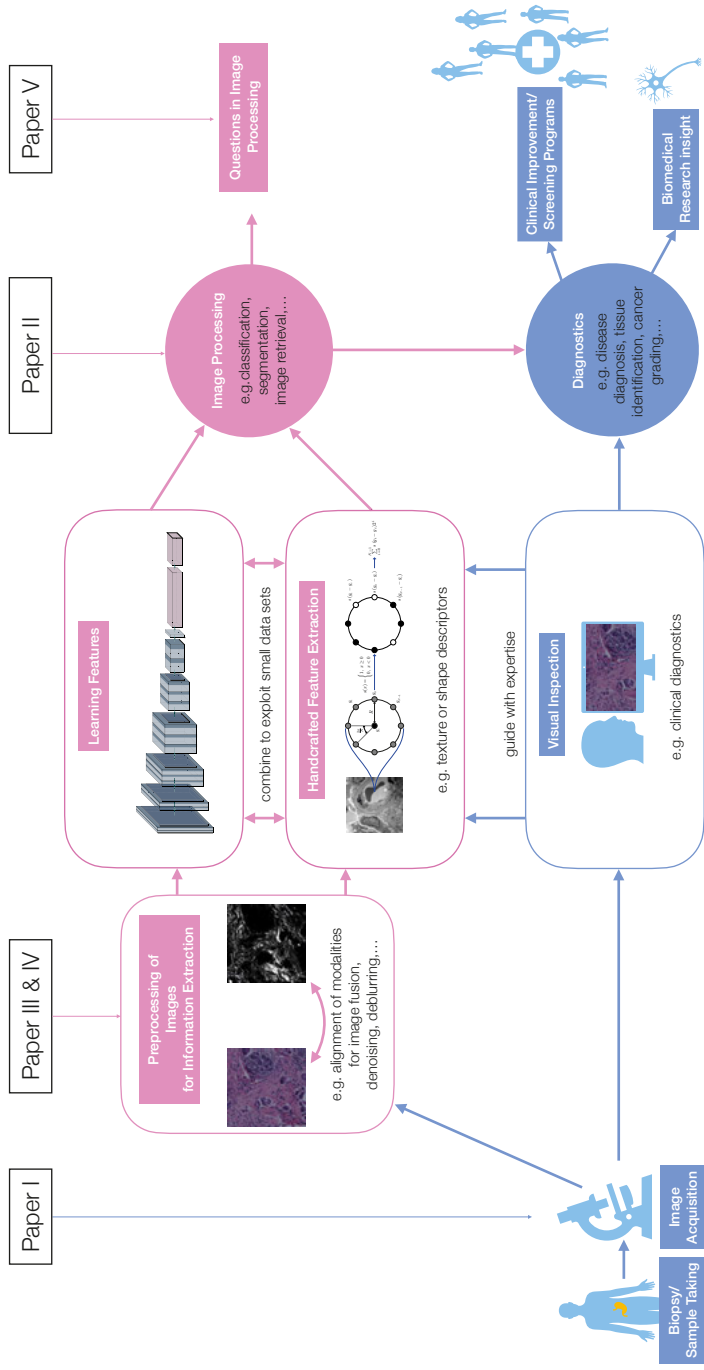


Figure 2.1. Graphical overview of the interdisciplinary field of digital pathology. In blue, the pathways are marked which heavily depend on humans – from collecting a biological sample, over its imaging, to the visual inspection and interpretation of the images to form a diagnosis, prognosticate, or derive other information relevant for decision-making. In pink, the parts of the workflow are highlighted which can typically be assisted by computational methods – preprocessing of the images, selection of regions of interest, and feature extraction for image processing tasks, which can, in turn, provide derived and compressed information to aid the pathologist with the tasks at hand.

2.2 Thesis Outline

The thesis starts with an introduction of the imaging modalities and datasets used to evaluate the proposed methods. In the following sections, I introduce the image processing tasks performed in the papers included in this thesis, and give a short review of different image representations, fusion strategies of heterogeneous input for learning-based methods, and embedding learning strategies used in the contributed work. In Chapter 4, I briefly summarize the papers that make up this thesis, discuss the methodology they have in common, and formulate overarching conclusions they share concerning representation learning and information fusion. Finally, in Chapter 5, I discuss the conclusions of this thesis and future perspectives of my research.

3. Background

In this chapter, I give an overview of the types of images I have worked with in the studies carried out in connection with this thesis, introduce the datasets used for evaluation and point out their challenges. I then provide an introduction to the image processing tasks carried out in Papers I-V, as well as to image representations and features originating from classic image processing and machine learning. Finally, I introduce two methods to learn embedding spaces, which I use in the studies included in this thesis to fuse learned and handcrafted features or images originating from different sources.

3.1 Microscopy

Optical microscopy has been the major driving force of biological discoveries since it was first used to study tissue and cells in the 17th century [136]. Until relatively recently, visual discoveries could only be shared using hand drawings. Even after it became feasible to acquire analog photographs of the magnified samples, analysis was primarily subject to qualitative and descriptive means, and it was not until the use of digital cameras combined with advances in optics, electronics, and computational methods, that quantitative microscopy could be performed, most often relying on image processing techniques.

3.1.1 Visible-light Microscopy

Visible-light microscopy, often referred to as optical microscopy, or light microscopy, is the type of microscopy that uses light in the visible spectrum and a system of lenses to generate magnified images of small objects [19].

Bright-field Microscopy

Bright-field (BF) microscopy is an absorption-based, simple technique requiring only basic equipment that is relatively easily accessible [132] and can be used for live cell imaging. The main disadvantage is that most biological samples have low contrast in this imaging technique, as the contrast results from the sample's light absorbance. This holds true especially when the sample has no color of its own, as is the case for many human cells. Contrast defines how well a particular object of interest can be distinguished from the

background [106]. Hence, samples often have to be stained to increase the contrast. A popular choice of stain for histological samples is the hematoxylin and eosin (H&E) stain. This is the gold standard to visually study tissue samples in histopathology [29]. For cytological samples, a Papanicolaou (Pap) stain is a common choice, e.g., used to detect cervical cancer in Pap smear tests [116].

Fluorescence Microscopy

Fluorescence refers to the emission of light by a substance that has absorbed light, a form of luminescence. The image is created by illuminating a sample with light that is absorbed by fluorophores, which are fluorescent chemical compounds. Fluorophores are sensitive to specific wavelengths and the wavelength that is used to illuminate a sample can be filtered with special emission filters. Fluorescence Microscopy (FM) allows to study processes within a specimen as some cellular components can be labeled with a fluorescent stain or antibody which binds to a particular molecule, or a fluorescent protein can be genetically linked to the protein of interest [98]. Sometimes this labeling can influence the function or the localization of the protein, but it allows us to study where a protein is found inside a cell [98] and is currently the primary imaging tool in cell biology [29].

Quantitative Phase Imaging

Quantitative phase contrast methods acquire an image of a sample by phase shift interferometry, creating a phase image in addition to the intensity image. Quantitative Phase Imaging (QPI) can be used on unlabeled specimens, which means no chemical or fluorescent labels are required [89], making QPI a so-called label-free imaging technique. This manifests in lower phototoxicity than FM and no photobleaching [106]. QPI is emerging as a powerful imaging technique for cells and tissue as it builds upon favorable aspects of microscopy, holography, and light-scattering techniques. As such, QPI is sensitive to morphology on a nanoscale and non-destructive to transparent structures [98].

Second Harmonic Generation

Second Harmonic Generation (SHG) is a shorter-than-excitation wavelength (STEW) process – a light-emitting phenomenon in which the emitted photons have a shorter wavelength than the incident photons [46]. In the case of SHG, this refers to photons of exactly twice the incident frequency – or half the wavelength – of the excitation laser [22]. The incoming light interacts with the electrons and atomic nuclei of the sample, resulting in different optical effects. Most of these interactions are nonlinear and induce a polarization consisting of multiple components, including electric dipole polarization, magnetization, and electric quadruple polarization [118]. Generally, the total

induced polarization is difficult to find. However, for sufficiently weak electromagnetic (EM) fields, which explain most nonlinear optical phenomena, it can be expanded into a power series expansion of the EM field, whose n^{th} -term is responsible for the n^{th} -order nonlinear optical effect. The second-order nonlinear optical susceptibility is nonzero only in noncentrosymmetric molecules and results in the signal associated with SHG. Examples of biomaterials that have a strong SHG signal are microtubules and myosin [63], as well as collagen in the extracellular matrix, which supports epithelial cells and is highly altered in cancer, connective tissue diseases, autoimmune disorders, and cardiovascular diseases [22], making this image modality a very interesting topic in cancer research. SHG is a label-free imaging technique that does not damage the sample [19] and can also be used in-vivo and on stained samples.

3.1.2 Electron Microscopy

While the advantage of using visible light is that it is generally not very damaging to the tissue sample, can easily be focused, and the fact that the human eye acts as a well-trained detector, it also comes at the cost of long wavelengths which cannot resolve small objects. Electrons, however, have a short wavelength (in the pm range), orders smaller than a single atom and can also easily be focused. This makes electron microscopy an important tool in material sciences, as it can resolve between atoms and allows for precise studies of material irregularities. In biology, the use of electron microscopy faces fundamental challenges regarding sample preparation. In order to focus the electron beam, the electrons and the sample have to be in an ultra-high vacuum, which does not allow for live imaging and requires tissue sample preparation to remove all water which would otherwise evaporate in the vacuum and lead to high scattering effects of the electrons. The sample preparation for a tissue sample consists of a dehydration process and plastic embedding, which can change the structures of the tissue, followed by cutting it into thin slices and a coating of the sample with metal to increase the contrast and to allow higher electron dosage [98]. Moreover, many biological structures such as the cytoskeleton, the Golgi complex, or endomembrane system are remodeled within fractions of a second after tissue anoxia, i.e., cell death as induced for sample preparation. This happens on a scale only relevant to electron microscopy imaging, but means that the image we eventually observe is altered from the true tissue composition before cell death and sample preparation.

Furthermore, biological imaging is subject to dose limitations as the high-energy electrons can break the covalent bonds in the sample. This limited exposure in turn, limits the contrast of images and lowers the signal observed relative to noise [98].

Hence, while the resolution power of electron microscopes is on an atomic scale, the resolution is reduced to nanometers in biological applications due

to artifacts in cell and tissue specimen introduced by sampling, fixatives, dehydration, staining, and section thickness. One fundamental research question therefore concerns how to preserve cell structures in ultra-high vacuum, the fact that the resulting image will be a projection of the 3D biological sample, and sample preparation itself as samples can only be 30-70 nm thin slices.

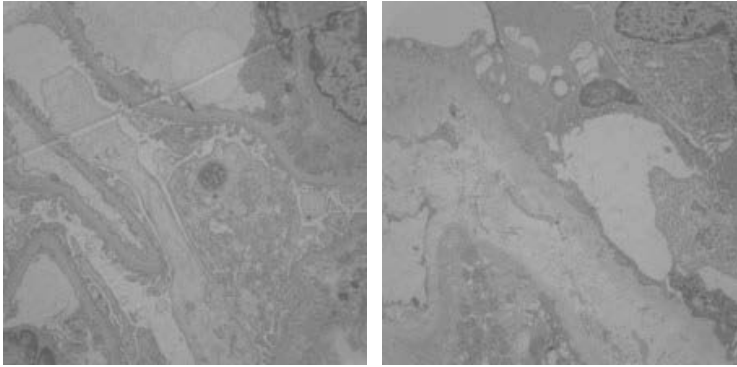
The unscattered and scattered electrons passing through and interacting with the sample are captured by Transmission Electron Microscopy (TEM). As such, the resulting image is a projection of the 3D sample, similar to X-Ray or Computed Tomography (CT) in medical imaging [98].

3.1.3 Datasets used for Evaluation

In the following section, I briefly introduce the datasets used in the papers of this thesis. Most illustrations in the thesis feature representative images from these datasets. Apart from one dataset comprised of remote sensing images, all datasets consist of microscopy images capturing either histological content, i.e., tissue samples, or cytological content, i.e., individual cells. There are some properties characteristic to microscopy images. As such, histological images generally do not have salient regions in the same way natural images do. They resemble texture images which is discussed in Sec. 3.3.3. Histological and cytological images are inherently symmetric under rotation and reflection [42, 127] in a mathematical sense, i.e. the objects in the images remain unchanged under these transformations (a cell or tissue can occur in any rotation). In digital histopathology, for example, digital slides of a stained specimen are analyzed. In the workflow of the slide preparation the tissue resection is done arbitrarily, meaning that the structures within the tissue section can have any orientation [70]. Similarly, cytological images are images of collected and stained cells spread out on a glass slide [44] and are hence arbitrarily oriented. Equivariant networks can exploit these properties as is discussed in Sec. 3.3.4. Most of these properties are also intrinsic to remote sensing images or other above-head imaging systems [88], which is why many methodologies developed for remote sensing are applicable to microscopy images and vice versa.

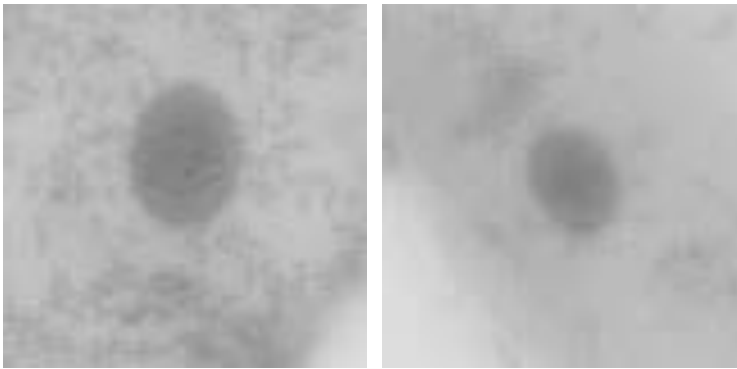
TEM Dataset – Paper I

The TEM dataset was acquired as part of the study in Paper I and consists of 494 intensity images of size 2048×2048 px captured using MiniTEM, a desktop, low-voltage (25 keV) TEM. The field of view (FoV) covered by one image is $16 \mu\text{m}$, yielding a pixel size of 7.8 nm. The histological images of nephrological tissue were annotated manually by assigning one of two labels – either containing glomerulus-specific structures, or containing other kidney tissue. An example is shown in Fig. 3.1.



(a) Kidney tissue with glomerulus (b) Other kidney tissue

Figure 3.1. Examples of TEM images as used in Paper I.



(a) Healthy oral cell (b) Cell from oral cancer patient

Figure 3.2. Examples of BF microscopy images of cells with a Papanicolaou stain from the oral cavity.

BF Dataset – Paper II

The BF dataset used in Paper II was acquired using a BF microscope with a 570 nm bandpass filter and first introduced in [139]. It consists of cytological images showing single cells with a Pap stain collected from the oral cavity of healthy and cancer patients. These intensity images are of size 80×80 px, each showing a cell nucleus in its center. There are 7755 healthy cells and 2519 cells from cancer patients, with a patient-level class label. Example images of cells from healthy and cancer patients are shown in Fig. 3.2.

BF & SHG Dataset – Papers III-V

The BF & SHG dataset is a histological registration [31] benchmark dataset which was created in the course of Paper III. It originates from tissue microarray (TMA) core image pairs provided in [32]. Patches were cropped from the

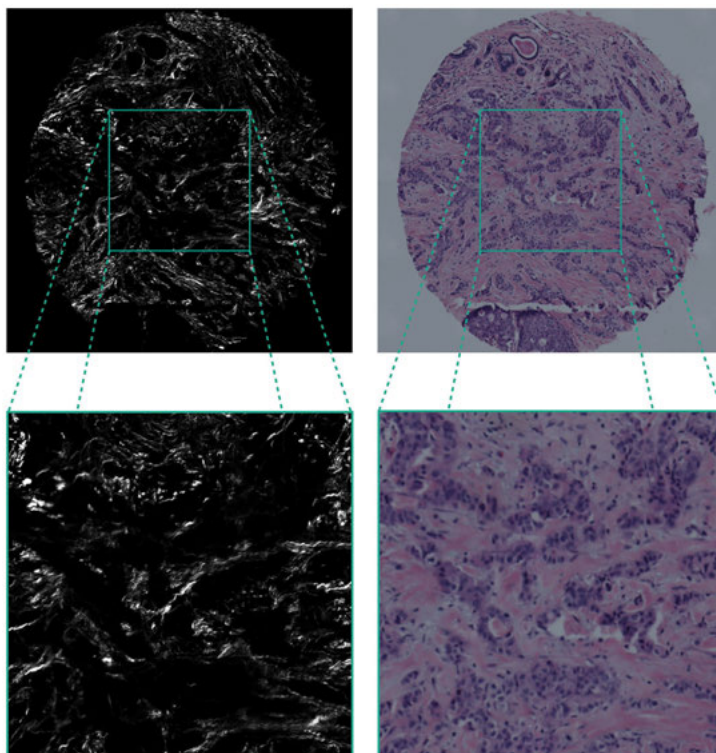
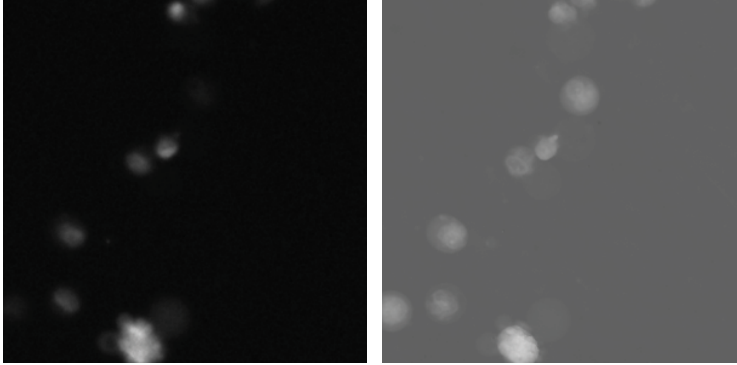


Figure 3.3. Figure from Paper III. Example of a SHG microscopy image (left) and the same sample with an H&E stain imaged by BF microscopy (right). The teal squares delineate the patches extracted from the TMA cores to create the multimodal registration dataset available at [31].

TMA core images. The patch size was chosen such that the circular borders of the TMA core were not present in the patches. An example of two corresponding images of this multimodal dataset is shown in Fig. 3.3. There are 206 image pairs of size 834×834 px in this dataset. Both images in a multimodal pair originate from the same sample slice, i.e., there are no deformations between the images, but the images are acquired in different microscopes and are aligned manually using landmarks [64].

QPI & FM Dataset – Paper V

The QPI & FM dataset is a publicly available multimodal cytological registration benchmark dataset [84] of simultaneously acquired correlative time-lapse QPI [128] and FM images [129] of three prostatic cell lines exposed to cell death-inducing compounds. The original images are acquired by a multimodal holographic microscope, meaning they are co-aligned by acquisition, i.e., there are no deformations or misalignments when using this dataset as



(a) Fluorescence microscopy image (b) Quantitative phase image

Figure 3.4. Example of a multimodal image pair in the QPI & FM dataset [84].



(a) RGB image (b) Near-infrared image

Figure 3.5. Example of the multimodal image pair from the remote sensing dataset acquired by RGB & NIR [131].

ground truth for registration as done in Paper V. An example of two corresponding images of this multimodal dataset is shown in Fig. 3.4.

Remote Sensing Dataset – Paper III

The publicly available remote sensing dataset [131] consists of 20 satellite images of the city of Zurich each with a size of about 930×940 px. The multimodal images are captured with the same sensor in identical resolution in RGB and Near-InfraRed (NIR), i.e., the images are co-aligned by acquisition and share a lot of common structures. An example of two corresponding images of this multimodal dataset is shown in Fig. 3.5.

3.2 Image Processing Tasks

In this section, I give an overview of the image processing tasks performed in the papers of this thesis.

3.2.1 Image Classification

Image classification is the process of assigning a label or category class to an image, for example, whether an imaged cell originates from a cancer patient or a healthy one, as done in Paper II. The mapping function from an image to its label is most often expected to be invariant to many transformations, such as translations, rotations, intensity, or illumination changes, to name a few. This resides in the fact that generally, an object stays within the same category, irrelevant of its orientation. Exceptions are objects characterized by their orientation, e.g., the digits 6 and 9. Invariant and equivariant functions are discussed in more detail in Sec. 3.3.4.

Prior to the advent of deep learning (DL), image classification pipelines usually consisted of a feature extraction step, in which high-level descriptors (the features) are extracted; and a classification step, in which a dedicated algorithm (the classifier) maps the features to their final labels. Typical classifiers include random forests, and support vector machines (SVMs).

In the first step, characteristic features, e.g., texture or shape descriptors, are handcrafted. The raw pixel intensities themselves could also be used as features for a classifier. However, even for a relatively small image of, e.g., 256×256 px, this would result in 65 536 features, which in turn means that a very large number of training images are needed to learn a reliable decision boundary and combat the curse of dimensionality. Another issue of this naive approach is that spatial relationships of the pixels in the image are lost when each pixel is considered an independent feature from its neighboring pixels. Hence, the design of a well performing feature descriptor used to be the backbone for any image classification task. Today, features are often learned from the data itself by training a convolutional neural network (CNN). This combines the feature extraction and classification steps into a single end-to-end learning procedure, in which the features are directly optimized for their discriminative power in the classification task. This topic returns in Sec. 3.3.1, but the interplay of handcrafted and learned features is revisited throughout this entire thesis.

Similarly, image classification is a task so intrinsic to the entire field of image processing that we encounter it in each of the papers of this thesis. In Papers I & II, in form of binary image classification and in Papers III-V in the form of a categorical cross-entropy loss which is used to identify (i.e., classify) an anchor's positive sample amongst a set of negative samples (see Sec. 3.4.2).

Evaluation of Image Classification

Classification performance is straightforward to quantify. Several measures have been formulated. In Paper I we report accuracy and in Paper II we report accuracy and F1-Score. In both papers we solve binary classification tasks, i.e., only two class labels exist.

Accuracy is defined as

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{all classifications}} = \frac{TP+TN}{TP+TN+FP+FN}, \quad (3.1)$$

where TP are true positives, FP are false positives, TN are true negatives, and FN are false negatives. The F1-Score is given by

$$\text{F1-Score} = \frac{2TP}{2TP+FP+FN}. \quad (3.2)$$

It depends on the data and task which measure is the most important to report, e.g., in diagnostics and screenings, a balance has to be found between sensitivity and specificity as often higher sensitivity means lower specificity and vice versa. It is disease (application) dependent if it is more harmful to miss a patient (false negative) or expose a healthy patient to treatment (false positive).

3.2.2 Image Registration

Image registration is the task of finding a spatial transformation between two images, such that corresponding points assume the same coordinates [40]. An illustration for a pair of multimodal images is given in Fig. 3.6.

Registration techniques can be categorized in multiple ways based on the following:

- rigid, affine, or deformable registration – depending on the type of transformation;
- monomodal vs. multimodal – depending on if the images originate from the same modality or not;
- intensity-based vs. feature-based – depending on if the whole image or a subset of sparse features are used for alignment.

Many different methods have been developed over time, and the choice of technique depends on the data and task at hand and the needs for the application (robustness versus quality or speed). Several review papers provide a thorough taxonomy of existing methods, particularly for the medical imaging domain [147, 130], including recent deep learning approaches [18].

Geometric Transformations

Image registration methods can be categorized based on the transformation model they support. The choice depends on the expected geometric displacement between the images. *Rigid* transformation models are composed of translations and rotations; *nonreflective similarity* transformations of translations,

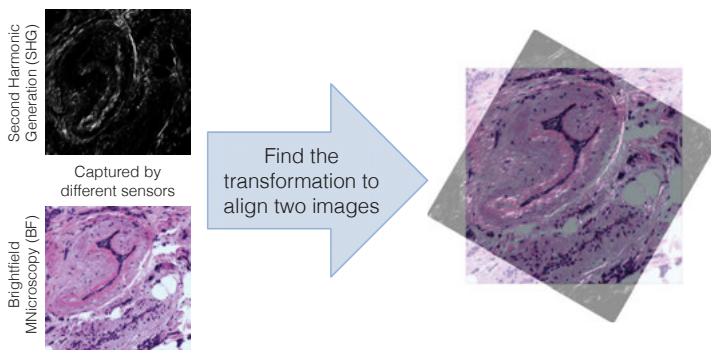


Figure 3.6. Multimodal image registration is the task of finding the geometric transformation to align two images captured by different sensors (here SHG and BF as used in Papers III-V), which can differ strongly in their appearance.

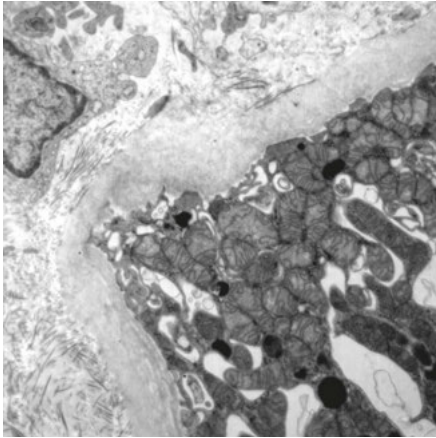
rotations, and scaling; *affine* transformations of translations, rotations, scaling, and shearing; and *deformable* (also called nonrigid or nonlinear) transformations are the most flexible models which allow each pixel to move freely (under certain conditions) and require regularization to confine the movement.

Rigid registration is the most used registration approach in the medical domain and clinical procedures, whereas research is focusing mostly on nonlinear methods [130]. Often rigid registration is performed prior to deformable registration to ensure a good initial displacement for local optimization.

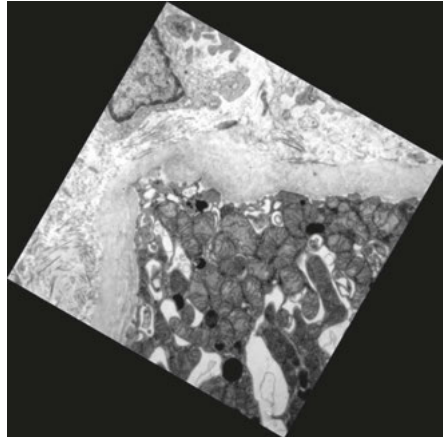
Figure 3.7 shows an example image of the TEM dataset used in Paper I (see Sec. 3.1.3) and transformed versions after applying either a rigid, affine or nonlinear (deformable) transformation.

Monomodal versus Multimodal Registration

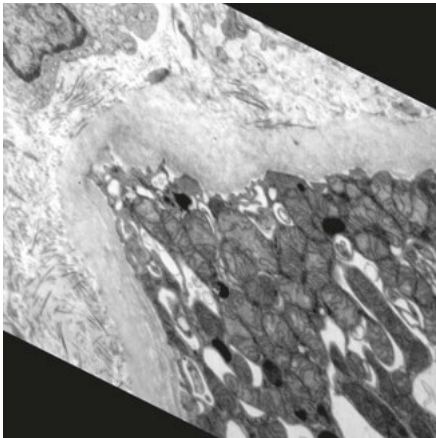
Monomodal image registration refers to alignment tasks in which both images stem from the same imaging domain. The images can be subject to occlusions or intensity differences (e.g., different weather conditions), but generally it can be assumed that objects present in both scenes result in similar signal expressions in the images. This makes finding a relationship between pixel intensities or correspondences between the images generally easier than in the multimodal case. For the latter, the images are formed by different imaging techniques, and many structures visible in one modality are not captured by the other modality at all. Fig. 3.8 shows an example of the BF & SHG dataset (see Sec. 3.1.3). The two images are aligned and show the exact same tissue sample in the same field of view and scale. It can be seen that while the BF image is dense and rich in information, the signal in the SHG image is very sparse. Clearly, registration methods that rely on matching pixels between the images based on the images' intensities in such a multimodal setting, i.e., intensity-based registration techniques, face a nearly impossible task. Mutual information (MI) is a widespread choice as a similarity measure to quantify



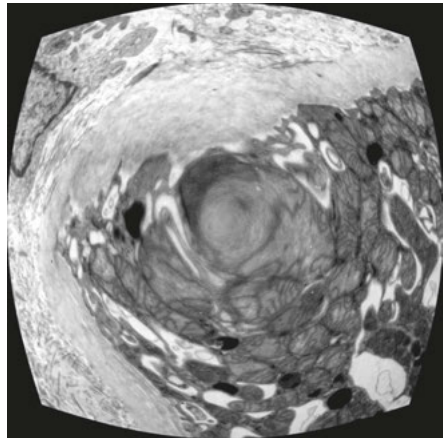
(a) Image from the TEM dataset



(b) Rigidly transformed image (a)



(c) Affinely transformed image (a)



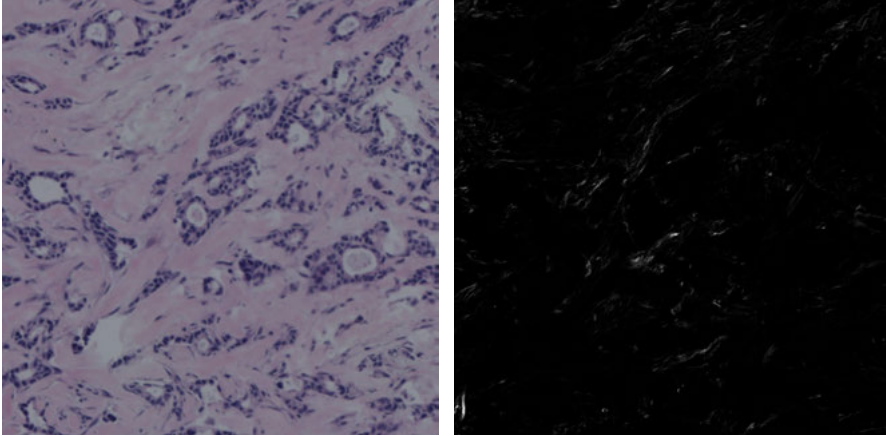
(d) Non-linearly transformed image (a)

Figure 3.7. Examples of geometric transformations encountered in registration tasks. All transformed images are scaled to fit for illustration purposes.

the affinity between two images. A thorough introduction of MI as an image similarity measure for registration is given in [109]. One major drawback of using MI is that it is commonly optimized by local optimization methods, which are subject to getting stuck in local minima, particularly if the initial displacement between the images is large. Öfverstedt et al. have recently successfully alleviated this issue in [96].

Intensity-based and Feature-based Registration

While intensity-based (or area-based) registration methods rely on the similarity of the entire images (sometimes using subsampling for efficiency), methods that aim to locate prominent, sparse, and salient features, such as corner



(a) BF image of tissue with H&E stain (b) SHG image of the sample in (a)

Figure 3.8. Tissue can have a dense signal expression in one modality (here BF), while a sparse signal is captured by another modality (here SHG).

points, are referred to as feature-based registration techniques. Feature-based methods rely on a feature extraction and a feature matching step in which the geometric correspondences between extracted features of the images are established. The most popular feature detector and descriptor used in registration is the Scale-Invariant Feature Transform (SIFT, [81]) due to several favorable properties: it is scale, rotation, translation invariant and to some degree insensitive to affine projections and illumination variations. The feature matching step to estimate the transformation is often done by RANdom SAMple Consensus (RANSAC,[33]).

In Paper III, we propose a representation learning technique to generate image representations of multimodal images which solve the multimodal rigid registration task by transforming both images into a shared space in which monomodal registration methods can be used. In follow-up work by Nordling et al. [94], it was shown that this approach could also be used for multimodal deformable registration. In Papers III & V, we evaluate the quality of learned multimodal representations by the downstream task of registration, using both feature-based [81] and intensity-based [95] registration methods and compare to local optimization of MI [90].

Evaluation of Image Registration Methods

In order to evaluate a registration technique, it is essential to quantify its performance reliably [112]. The most straightforward and trustworthy way to do so is by using landmarks corresponding to salient features in both images. In the case of rigid registration (as is the task in all registration experiments performed in this thesis), a suitable and simple choice is to use the corner points

of the images as landmarks and define the registration error as

$$err = \frac{1}{n} \sum_{i=1}^4 \left\| C_i^{Ref} - C_i^{Reg} \right\|_2, \quad (3.3)$$

where C_i^{Ref} , C_i^{Reg} , $i \in \{1, 2, 3, 4\}$, are the corner positions of the reference image and the one resulting from the registration, respectively.

3.2.3 Image Retrieval

Content-Based Image Retrieval (CBIR) is the task in which features of an image are used to search for the closest match to a query from a large set of images (often called repository) based on *content*. The query can be provided in different forms, often keywords, class labels, or images. If images are used as queries, the CBIR is referred to as Reverse Image Search (RIS) [126] or *query-by-example*. Most often CBIR systems consist of feature extraction followed by feature matching based on a suitable similarity measure [62, 92].

A well-established technique to perform CBIR is to accumulate local feature descriptors into a so-called *bag-of-words* (BoW, also referred to as bag-of-features) [21, 107, 121], for which the most descriptive features (words) form a vocabulary and each image is assigned a histogram of words. The retrieval step is then based on histogram comparison, typically using cosine similarity. Fig. 3.9 shows a graphic sketch of how a BoW is formed. First, local features are extracted on a set of images (e.g., SIFT features or using a pretrained CNN). Next, the entirety of extracted features are clustered (in Paper IV using K -means) i.e., the features are partitioned into K clusters. These clusters represent individual words of the vocabulary (illustrated by the differently shaped and colored markers in Fig. 3.9). These words then summarize the locally extracted features, and each image can be represented by one global descriptor – a histogram that counts the occurrences of each word in the image. The histograms can be matched in computationally efficient ways.

The relevance of CBIR is high in digital pathology, due to the increased use of whole slide image (WSI) scanners, which has led to the creation of large datasets of microscopy images. This data has great potential to unravel patterns usable for early disease diagnoses such as cancer if computational systems can make the data searchable [51, 66, 102].

However, methods that perform *cross-modality image retrieval* (CMIR) or *cross-domain image retrieval* (CDIR), i.e., retrieval of images in one modality or visual domain provided a query in another, are still sparse, and subject to ongoing research. A recent review on CDIR is provided in [145] and presents work in multimodal domains for person re-identification, remote sensing, and sketch & natural image retrieval. Generally, the methods for cross-modal retrieval can be categorized into two approaches to solve the domain gap. They either perform *feature space migration* by extracting features in images of the

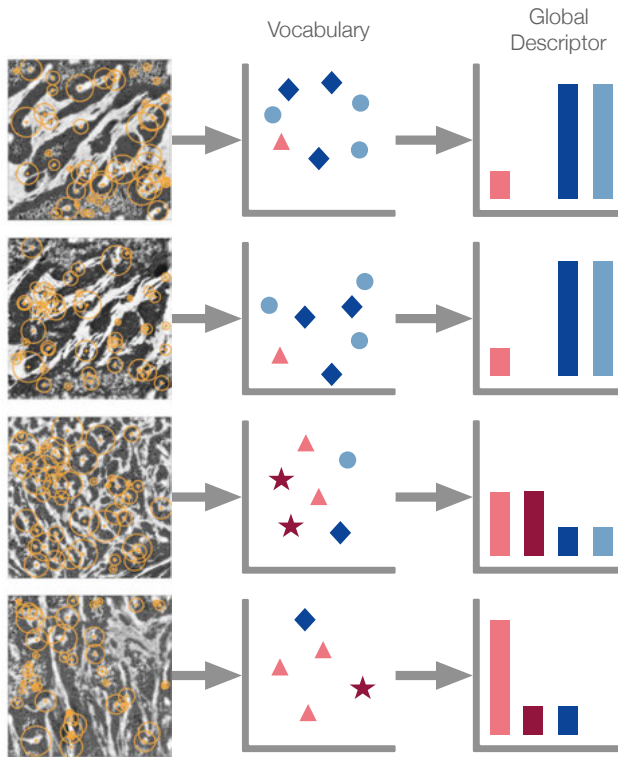


Figure 3.9. Simplified illustration of a BoW. Many local features are found in the feature extraction step, clustered into words (here differently colored and shaped markers), which form a vocabulary. The occurrences of each of these words are counted per image (corresponding colored bars) and summed up as a histogram which acts as a global image descriptor.

different modalities and learn a function to relate them across the domains, or they perform *image domain migration*, most often using generative networks to translate one modality into the other.

In Paper IV, we evaluate if the learned image representations of multimodal images proposed in Paper III, can be used to perform domain migration successfully, such that given a query image in *modality A*, its counterpart in *modality B* can be found within a set of images in *modality B* as shown in Fig. 3.10.

Evaluation of Image Retrieval Tasks

Different evaluation criteria are used in assessing CBIR systems. Their choice depends on the particular use case of the retrieval. In the evaluation of Paper IV, we define retrieval success based on if the correct match is found within the top-K results. We report the retrieval success rate as the percentage of images

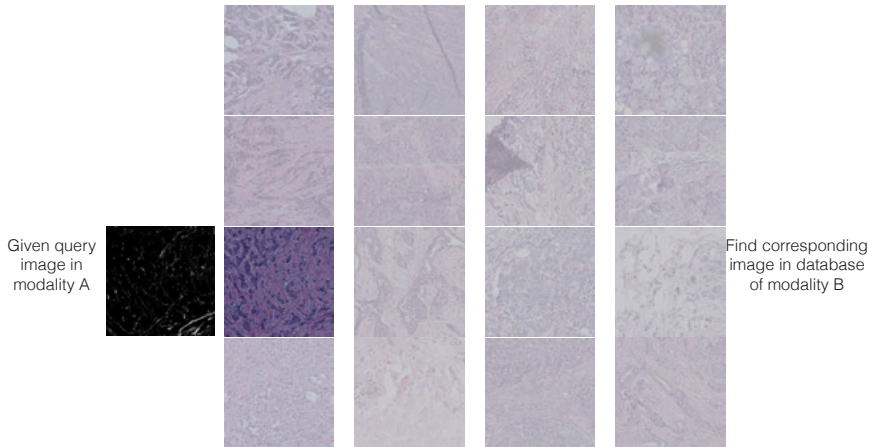


Figure 3.10. In the multimodal image retrieval task in Paper IV, the query is given in the form of a query image in *modality A*, and the aim is to find its match in a set of images in *modality B*. The match in BF to the SHG query image here is marked as opaque.

in the test set whose matches have been successfully found within the top-K results.

3.2.4 Image Generation

Image Generation, also referred to as *image synthesis*, is the task of generating new images by learning relevant features, structures, and statistics from an image dataset. In recent years a large number of models have been proposed for this task, many of them are based on Generative Adversarial Networks (GANs). While the image synthesis results are impressive in domains such as the generation of faces, the development for reliable systems in biomedical applications is lagging behind. One of the main reasons is that GANs generally require extensive training data, which is often unavailable in clinical settings. Furthermore, to be useful in biomedical settings, the generated images have to *reliably* produce reasonable and accurate output, which is still hard to ensure [82].

Generative Adversarial Networks

GANs [39] consist of at least two networks – a generator and a discriminator. The generator and discriminator compete in a zero-sum game, in which the generator generates a representation often called *fake image*, and the discriminator learns to discriminate between these fake and real images simultaneously. Ideally, this will result in a generator producing fake images indistinguishable from real ones. This principle can be used to transform an image from one domain or modality to another in a process called Image-to-Image

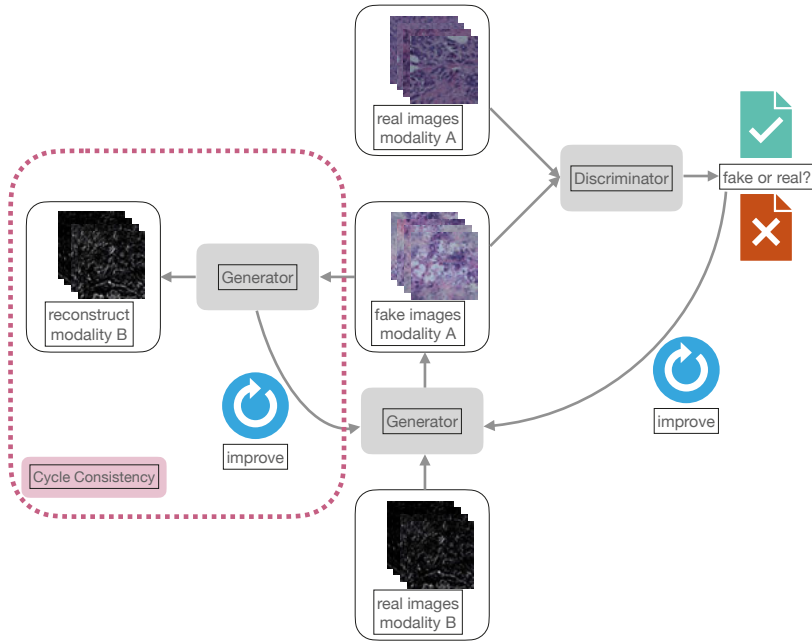


Figure 3.11. Schematic overview of GAN-based I2I translation. A generator generates fake images in *modality A*, given images in *modality B*. The discriminator is trained to classify if these generated images are real or fake, its feedback making the generator stronger. In the case of CycleGAN, an additional generator is introduced, which reconstructs the images in *modality B* from the fake images and gives feedback to the I2I-performing generator (Cycle Consistency).

(I2I, [56]) translation. Paired image-to-image translation as, for example, performed by pix2pix [56] (used in Papers III & IV) uses an adversarial loss in combination with a reconstruction loss between the fake and target image to find a suitable mapping. Some GANs can achieve I2I also in an unpaired manner, e.g., CycleGAN [146] (used in Papers III & IV) and Disentangled Representation for Image-to-Image Translation (DRIT [73, 74], used in Paper III). They do so by also learning the inverse mapping back from the target domain to the input domain using another reconstruction loss. This is referred to as cycle consistency and is currently the backbone of most unpaired I2I networks [105]. A graphical overview of a GAN is given in Fig. 3.11.

Contrastive Learning

In Paper III, we propose to use contrastive learning (see Sec. 3.4.2) to generate common 2D, image-like representations of multimodal image pairs, which we name *Contrastive Multimodal Image Representations* (CoMIRs). Unlike GAN-based I2I, CoMIRs do no attempt to transform one modality to the other,

but rather to find one abstract modality between the two that extracts common structures.

Evaluation of Generated Images

The performance of image-generating methods is less straightforward to quantify than of methods performing classification, registration, or retrieval. Often evaluation is done qualitatively, i.e., by visual inspection and judgment of the generated images. This is obviously time-consuming for large datasets, can be subjective and biased to the person performing the qualitative evaluation, and may even require expert knowledge in the case of biomedical images. A few quantitative measures have been proposed to evaluate, in particular, GAN generated representations. Most prominent are the Fréchet Inception Distance (FID, [53]) and Kernel Inception Distance (KID, [15])[17, 91]. FID calculates the Fréchet (or Wasserstein-2) distance between multivariate Gaussians fitted to the embedding space of the Inception-v3 network [124] of generated and real images [17]. It thereby compares the distribution of generated images with the distribution of the real images. Some studies showed that FID correlates well with human visual assessments of generated images [26, 104], but counter-examples are reported in [80].

In Paper V, we evaluate CoMIRs by reporting their FID, as well as several image metrics and similarities from classic image analysis, such as the pairwise mean squared error, correlation, structural similarity index measure (SSIM, [135]), and α -AMD distance [78, 95]. However, we only find weak relations between all these measures, except for correlation, and the CoMIRs' usefulness for the downstream task of rigid, feature-based registration.

3.3 Image Representations

3.3.1 Learned Features vs. Hand-Crafted Features

Learned features are features extracted from the data by a neural network, in the case of images, usually a CNN. The features are often obtained by training classifiers, either by providing labels of the data or in a self-supervised manner. Handcrafted features, on the other hand, are designed independently from the data and require no training. This means they can be applied to individual images and hence can outperform learned features when only small datasets are available [77]. Learned features are very flexible but also run the risk of learning shortcuts in the data [72] and are prone to overfitting for small datasets. It is often assumed that CNNs can learn all necessary features for a downstream task from the data themselves, but their interpretability is limited compared to handcrafted features. This problem is amplified when the loss function of a CNN acts as a surrogate function for the actual downstream task, and the learned features are not directly optimized for that task (e.g., in case a pretrained CNN is used a feature extractor on different, unlabelled data, or in the use of the contrastive loss as in Papers III-V to generate representations for a different downstream task).

Several studies show the benefit of fusing the power of CNNs with feature engineering, i.e., creating features which are transformations of primitive features that are readily available (e.g., pixel intensities) [77, 85]. Similarly, in Papers I & II, we study if the performance of CNNs can be boosted by providing particular focus on texture features in the images for two binary classification tasks – one on the TEM dataset and one on the BF dataset.

3.3.2 Deep Learning

AI is comprised of a number of interdisciplinary technologies aiming to solve a complex problem. It is tightly connected to ML, which is usually considered a subfield of AI. ML aims to learn a particular task based on training data, from which it extracts features and patterns relevant to solving the task.

One particular class of learning methods is neural networks (NN), also referred to as artificial neural networks (ANN). Their central concept is to form linear combinations of features extracted from the input and model the target output as a nonlinear function of these derived features [47].

NN models have unknown parameters, usually called weights, and optimization techniques are used to find values that optimize the loss function, which in turn gives feedback to the network about how relevant the features extracted from the data are for the downstream task. This iterative updating of weights and features is what constitutes learned features, differing from handcrafted features, which, if poorly chosen, can be irrelevant to the task. Key to the quality of learned features is a careful selection or design of the loss function with respect to the machine learning task, as different losses are used

for different tasks. For classification tasks, for example, the loss is most often chosen as the cross entropy

$$\mathcal{L} = - \sum_{i=1}^k y_i \log(p(y_i|x_i; \theta)), \quad (3.4)$$

where k is the number of classes, y_i the label vector, x_i the input image, θ the network weights, and $p(y_i|x_i; \theta)$ the output of the softmax layer in the network.

While all NNs are compositions of linear and nonlinear functions (layers), CNNs are characterized by using convolutions instead of general matrix multiplications in some of their layers. Using convolutions is particularly suitable for processing image data. If many layers are hierarchically combined, more complicated functions can be modelled as compositions of simpler ones, a concept fostered in the research area referred to as Deep Learning (DL).

VGG [120] (used in Paper I) is a relatively simple CNN architecture composed of blocks of convolutional layers with a very small receptive field and max-pooling layers that reduce the spatial dimensions. One problem VGG faces is that especially for deeper versions, i.e., many of those convolutional blocks stacked on top of one another, gradients can become too small. Consequently, the weights fail to be updated, a problem referred to as vanishing gradients [37]. Using residual connections, a type of skip connection, as used in ResNet [50] relieves this issue (used in Papers I, II & IV). They allow the network to connect input from earlier layers with the output of layers deeper in the network. Another network architecture used in Papers III-V is a Tiramisu network [59] based on the U-Net[114] architecture. U-Net has been designed to process images at various resolution levels and combine these multi-resolution features to perform image segmentation. The Tiramisu architecture uses additional skip connections as compared to the original U-Net architecture.

3.3.3 Texture Features

No formal definition of texture exists [38], though its quantification is considered an important region descriptor and characteristic to identify objects or regions of interest in an image, regardless of its domain [38, 46]. Texture describes smoothness, coarseness, and regularity and can be fine, rippled, mottled, irregular, or lined. Texture is a characteristic property of surfaces, i.e., the weave of a fabric or the patterns of crop fields, and carries important information on the structural arrangement of surfaces [46]. Hence texture descriptors have been an essential tool in image analysis related to material science and satellite imagery. Generally, the techniques to quantify texture are categorized into statistical, structural, and spectral approaches [38, 46]. While statistical approaches characterize textures as smooth, coarse or grainy, structural techniques focus on the arrangement of image primitives, e.g., equidis-

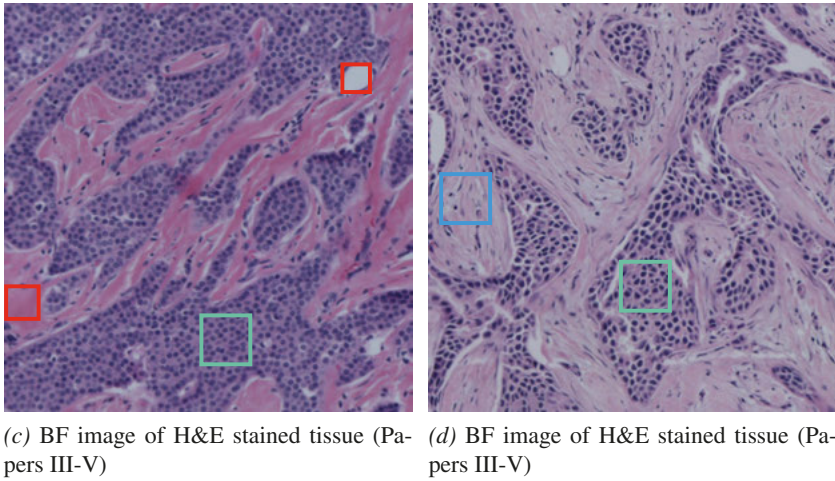
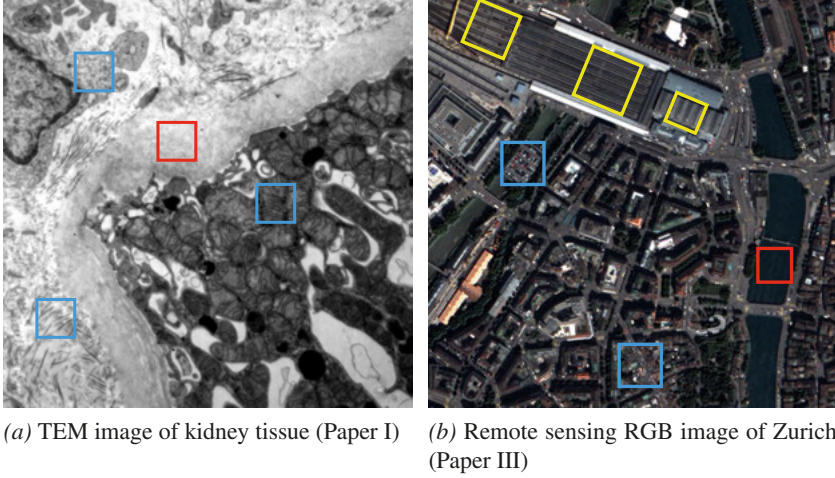


Figure 3.12. Examples of occurrences of different textures in the datasets used in this study. Blue frames mark coarse textures, red frames smooth textures, yellow frames regular textures and green mark semi-structured textures.

tantly spaced parallel lines, and spectral techniques identify periodic patterns in the Fourier spectrum.

Some of the most successful structural texture descriptors in image analysis are Local Binary Patterns (LBP, [48, 133]) [99]. A number of variants have been proposed [79], and found useful for numerous applications [79, 99, 103, 108].

The LBP code $LBP_{r,p}(c)$, for a pixel c with intensity value g_c , is given by

$$LBP_{r,p}(c) = \sum_{i=0}^{p-1} s(g_i - g_c) 2^i, \quad s(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0, \end{cases} \quad (3.5)$$

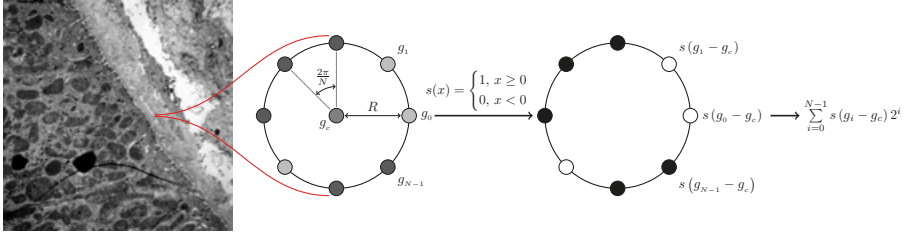


Figure 3.13. Illustration of $LBP_{R,8}$ code generation for a pixel c with intensity g_c of an image from the TEM dataset.

where g_i , are the intensities of points sampled equidistantly on a circle of radius r in the neighborhood of g_c for $i \in \{0 \dots p - 1\}$.

LBP codes are usually binned over the full image into a histogram, thereby providing a feature vector of size 2^p for the entire image. This approach is similar to the descriptors introduced in Sec. 3.2.3 in which local features are clustered into BoW-descriptors.

LBP Maps

Levi et al. [75] propose an approach to use LBPs as input for CNN training. Instead of binning the codes into a global descriptor for an image, they suggest keeping each LBP code per pixel, thereby creating a dense map of values in the range $[0, 2^p - 1]$, the same size as the image itself. LBP codes themselves are not suited as input to CNNs, because CNNs are based on discrete convolutions, similar to a weighted average of their input. LBP codes, however, are an unordered set of binary codes, and codes with similar numeric values do not necessarily relate to similar patterns. Hence averaging two codes is not reasonable.

Multidimensional Scaling (MDS) can be used to map the LBP codes into a metric space and render them usable for CNN training. MDS is introduced in more detail in Sec. 3.4.1. The authors in [75] use MDS to turn the dense map of LBP codes into an LBP map suitable as CNN input and show that 3-channel LBP maps work best in their application of emotion recognition and motivates our choice in Papers I & II. Examples of such maps are shown in Fig. 3.14 for a varying radius, capturing texture at different scales.

Learning LBP-like features

Instead of extracting texture features first and then process them in a CNN, some attempts have been made to directly extract LBP-like features inside CNN architectures. Some of these methods integrate modules within the CNN architecture to extract features that resemble LBPs but are differentiable functions and therefore can be optimized (are learnable) by backpropagation. One such model is called Local Binary CNNs (LBCNN) [60]. In LBCNNs, LBPs are modeled by N fixed sparse filters, which are randomly initialized with ± 1

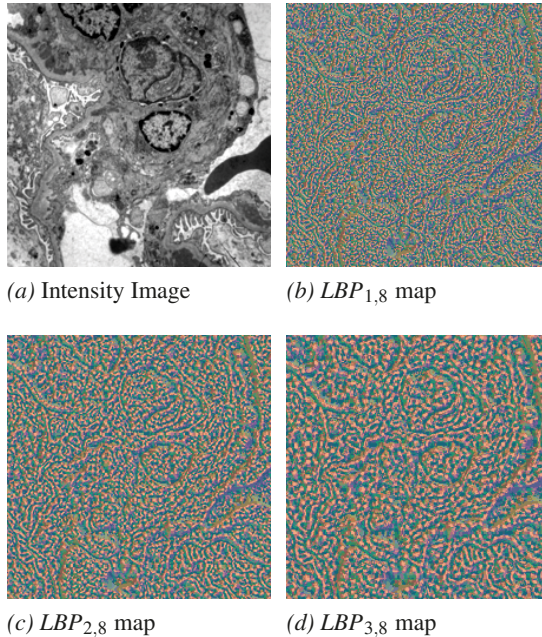


Figure 3.14. 3-channel LBP maps as introduced in [75]. Here, they are shown in different scales, indicated by different sizes of the radius for the 1-channel image in (a) from the TEM dataset.

at a chosen sparsity level, followed by a nonlinear activation function in the form of a rectified linear unit (ReLU) corresponding to the soft thresholding in the classic LBP function. A kernel with trainable weights transforms the input image into an LBP-like feature map which can be passed on to the next LBCNN layer. These LBCNN layers can be used as a substitution for regular convolutional layers in any CNN architecture.

Another attempt is made in [76] by proposing a network containing four modules, including an LBP extraction module within the network. In Paper II we evaluate this end-to-end learning approach, as well as LBCNN and LBP maps.

3.3.4 Geometric Deep Learning

Learning in high dimensions is intractable as the number of samples needed grows exponentially with the dimension of the feature space, which is commonly referred to as the curse of dimensionality [13]. Images are indeed high dimensional data if each pixel is considered an independent feature, but their spatial relations and underlying geometries can be exploited to simplify the issue. This can be done by, e.g., using graph neural networks which capture the relationships between pixels, superpixels, feature points, or regions. Another

approach that has proven very useful is to use geometric priors like symmetry, which decreases the complexity of the hypothesis class (i.e. the set of candidate models that can be learned) by admitting only equivariant neural networks without discarding useful hypotheses. Instead of fitting an unrestricted model, the hypothesis space can be reduced to the quotient space under the used symmetry group [138].

Invariance and Equivariance

The concepts of equivariance and invariance are tightly connected to groups. A group (G, \cdot) is an algebraic structure that consists of a set G and a binary operation (the group product), that is closed, associative, has an identity element, and there exists an inverse element for each element in G w.r.t. to the operation. Some of the groups most commonly considered are the cyclic groups \mathcal{C}_n consisting of rotations by multiples of the angle $360^\circ/n$, the dihedral groups \mathcal{D}_n consisting of \mathcal{C}_n and reflections, the Euclidean group $E(2)$ consisting of isometries of the plane \mathbb{R}^2 , i.e., translations, rotations, and reflections, which in turn can be constructed from the translation group $(\mathbb{R}^2, +)$ and the orthogonal group $O(2)$ (both are subgroups to $E(2)$), the special orthogonal group $SO(2)$, the group $(\{\pm 1\}, *)$ consisting of the reflections along a given axis.

An abstract group can be turned into a group that is identified with a concrete set of transformations by the notion of group actions given by the set $\mathcal{T} = \{T_g : A \rightarrow A\}_{g \in G}$, a set of invertible transformations that is compatible with the group [24].

Equivariance defines the property of a function to commute with the action of a symmetry group G when that symmetry group acts on its domain and codomain. A function or operator $f : \Omega \mapsto Y$ is called equivariant under a family of transformations \mathcal{T} if for any transformation $T \in \mathcal{T}$, there exists $T' \in \mathcal{T}$ s.t.

$$f(T(X)) = T'(f(X)) \quad \forall X \in \Omega, \quad T, T' \in \mathcal{T}, \quad (3.6)$$

i.e., the function f commutes with actions on the group that is acting on the space of the input and output in the following way

$$\begin{array}{ccc} X & \xrightarrow{T} & T(X) \\ f \downarrow & & \downarrow f \\ f(X) & \xrightarrow{T'} & Y \end{array}$$

Note that T and T' can, but do not have to be the same group actions.

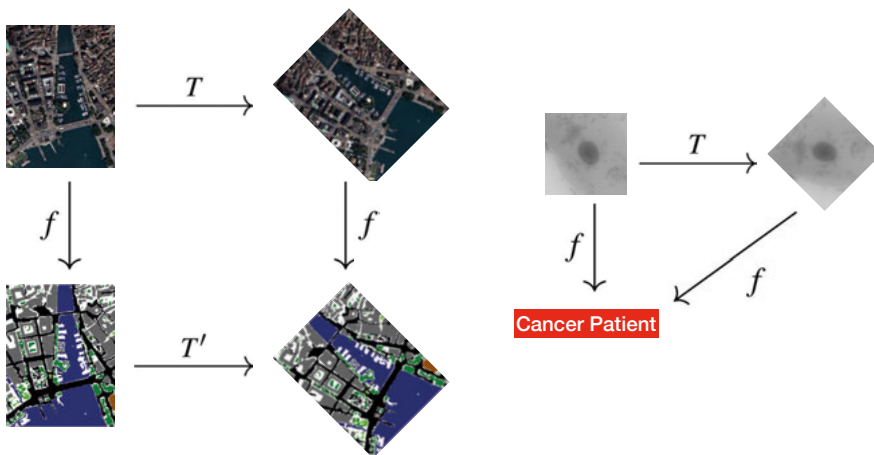
Invariance is a special case of equivariance when T' is the identity of the group, i.e., $f(T(X)) = f(X) \quad \forall X \in \Omega, T \in \mathcal{T}$. This means f and $T \circ f$ behave

in the following way

$$\begin{array}{ccc}
 X & \xrightarrow{T} & T(X) \\
 f \downarrow & & \swarrow f \\
 Y & &
 \end{array}$$

A function learning a classification label should ideally be invariant to many transformations – a cancer cell should be recognized as a cancer cell in any orientation and at any position in the image. But in tasks such as segmentation, registration, or computational imaging, in which the image of the CNN function is two-dimensional, translational and rotational equivariance in particular are highly beneficial. Figure 3.15 contrasts this interplay between T , a rotation by 45° , and a function f learned by a CNN to perform classification on either an image level (*image classification*) or on a pixel-level to create a 2D label map (*segmentation*).

Equivariance of features can be achieved through different approaches. Most commonly, it is either implemented through the network architecture design or is learned in the training process [24].



(a) Segmentation of an image from the Zurich dataset.

(b) Classification of an image from the oral cancer dataset.

Figure 3.15. Rotation equivariance versus invariance: for tasks such as segmentation, a function f learned by a CNN to map each pixel of the image to a label (given by different colors in the segmentation map, e.g., blue for water) has to commute with rotation, such that the segmentation labeling map of a rotated image equals a rotated segmentation labeling map. In classification, however, rotated versions of the input should all be mapped to the same label, i.e., be invariant.

Equivariance by design

Equivariance by design refers to designing a network architecture such that each mapping from layer to layer is an equivariant function. As a composition of equivariant mappings, the entire function learned by the network is also equivariant. As mentioned earlier, neural networks consist of linear and nonlinear functions. There are two main approaches to turn the linear maps inside a network into equivariant functions. One is based on *lifting convolutions*, and the other is based on *steerable convolutions*. The lifting approach is first introduced in [27] and generalizes the Euclidean convolution to a group convolution by defining the invariant Haar measure on the group G , which either turns into the Lebesgue measure if $G = \mathbb{R}^d$, or a counting measure if G is discrete. The equivariant convolutions on the group can be discretized and parametrized with learnable parameters just like common Euclidean convolutions, which are a special case of group convolutions when $G = \mathbb{R}^2$.

The challenge of this approach is that images are not automatically given in a form on which group convolutions can be applied but need to be *lifted* to the group by, e.g., using a linear map with learnable parameters. Furthermore, the resulting signal on the group has to be mapped back to the original image domain [24].

The steerable filter approach circumvents these challenges, given that the symmetry group is a subgroup of the affine group (as is the case for rotations, for example). The authors in [137] propose a way to constrain the filter kernels to products of radial functions and specific circular harmonics to use common Euclidean convolutions [24].

Additionally to the linear maps inside the network, the nonlinearities must be considered when designing an equivariant network architecture. If the group actions act only on the domain space of the signals, they are equivariant. However, if the group actions act on the range of the input signal, particular nonlinearities have to be considered [24, 137].

Equivariance by learning

Equivariance can be learned from data to some extent by using data augmentations, a method mainly used to enlarge the training data by generating new similar training images.

Implementation of learned equivariance is more flexible and can be more easily adopted to a wider class of transformations [24] than equivariance by design. However, it demands a high learning capacity which can easily lead to overfitting [138].

In Paper II, we implement Vector Field Networks [88] for the classification task of oral cancer cells. This model enforces equivariance by design as it convolves the image with rotated versions of each filter in a shallow network followed by a global pooling over orientations. It uses bicubic interpolation for the filter rotations, which results in a finer resolution of the orientations

but also comes at the cost of interpolation artefacts [138]. Unlike other approaches based on equivariance by design, such as [27, 28, 138], which store the full response of rotated filters, in the Vector Field Networks [88] only the maximum response is stored. In Paper III, we propose another approach to obtain rotational equivariant image representations of multimodal images, which is a type of equivariance by learning. However, it is a stronger approach than simply using data augmentations. It ensures that the network commutes with rotation and thereby learns that rotated input versions are not independent data samples.

3.4 Learning Embedding Spaces

An embedding is a representation of a topological object, such that algebraic properties are preserved. All papers associated with this thesis touch upon the topic of learning embeddings either to: (i) embed handcrafted features into a metric space, such that these embedded features can be used as CNN input while preserving the relevant extracted information by the handcrafted descriptor (Papers I & II), (ii) understand the underlying structure of high-dimensional CNN features (Paper V), (iii) or to learn embeddings for heterogeneous multimodal input such that corresponding samples in a multimodal image pair are mapped to points close together in that representation space (Papers III-V).

3.4.1 Multidimensional Scaling

MDS [16] is a common technique in data visualization. Given (dis-)similarities of data points, MDS can map the data from an unordered set into a metric space or find low-dimensional embeddings for high-dimensional features while maintaining the relative relationships between the data points in a numerically optimal manner. It is hence often used to embed data points for which no actual locations in a higher dimensional space are known, but their pairwise relationship can be quantified. An illustrative example is shown in Fig. 3.16: a set of cities can be mapped into 2D solely based on their pairwise distances on the sphere modeling Earth.

To realize this embedding, the dimensionality reduction is formulated as an optimization problem that minimizes an objective function called stress. Over time many different stress functions have been formulated. In this thesis, we use two different ones: Kruskal’s normalized stress-1 criterion [68] in Paper I and Sammon’s stress [115] in Paper V.

In Paper I, we perform non-metric, in Paper V metric MDS on the dissimilarity matrix $\Delta = (\delta_{ij}) \in \mathbb{R}_+^{n \times n}$ to find 2D points whose distances $\mathbf{D} = (d_{ij}) \in \mathbb{R}_+^{n \times n}$ approximate the dissimilarities in Δ .

In Paper I, we use a so-called representation function $f(\delta_{ij})$, which specifies the relation between the dissimilarities and their corresponding values $\mathbf{D} = (d_{ij}) \in \mathbb{R}_+^{n \times n}$ which lie in a Euclidean space and approximate a monotonic transformation of δ_{ij} . We use non-metric stress normalized by the sum of squares of the inter-point distances, which is given by the following Kruskal’s normalized stress-1 criterion, [68]:

$$\text{Stress-1} = \sqrt{\frac{\sum (f(\delta_{ij}) - d_{ij})^2}{\sum d_{ij}^2}}, \quad (3.7)$$

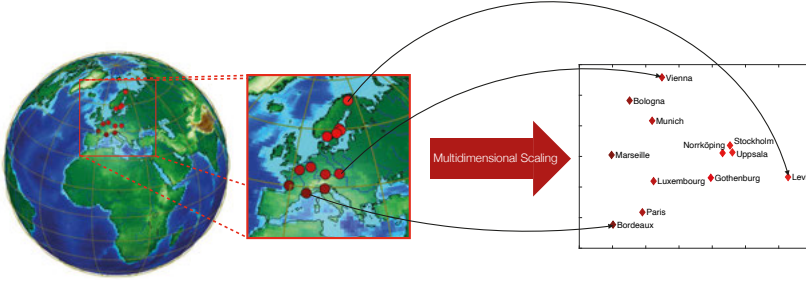


Figure 3.16. The concept of multidimensional scaling (MDS) illustrated on a set of European cities. Given the pairwise geodesic distances of the cities on the sphere, MDS can be used to map them into a lower dimensional space (here 2D), such that their pairwise distances are optimally preserved w.r.t. to optimization criteria. The tone of red indicates the cities' proximity to the northernmost city in this set – Levi, in Finland.

In Paper V we use Sammon's nonlinear stress criterion [115], given by

$$\text{Stress} = \frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}}, \quad (3.8)$$

where δ_{ij} denotes the distance between sample i and j in a high-dimensional feature space and d_{ij} the distance in the 2D projection space. In Paper V, Δ contains the pairwise MSE between high-dimensional CNN features.

To apply MDS to the set of LBP codes in Papers I & II, we use one of the dissimilarity measures between the codes suggested in [75]:

$$\delta_{ij} = \delta(P_i, P_j) = \min \left\{ \tilde{\delta}(P_i^0, P_j^0), \tilde{\delta}(\text{rev}(P_i^0), P_j^0), \tilde{\delta}(P_i^0, \text{rev}(P_j^0)) \right\}. \quad (3.9)$$

Here, $\tilde{\delta}(P_i, P_j) = \|CDF(P_i) - CDF(P_j)\|_1$, where P^0 is the concatenation of the binary string P and an additional bit of 0, $\text{rev}(P)$ the rearrangement of a string P in reverse order, and $CDF(P)$ is the cumulative distribution function (CDF) of bit values, approximating the Earth Mover's Distance between the strings [75].

3.4.2 Contrastive Learning

Contrastive learning (CL) describes learning an embedding function and space such that the embeddings of similar input samples are mapped close together. In contrast, dissimilar ones are mapped to points far away. The similarity between input images can be defined through labels (supervised CL), such that images from the same class are mapped to points close together, or in an unsupervised manner, such that similar inputs are versions of the same image [25, 49, 101, 142]. The latter is one of the most powerful techniques used

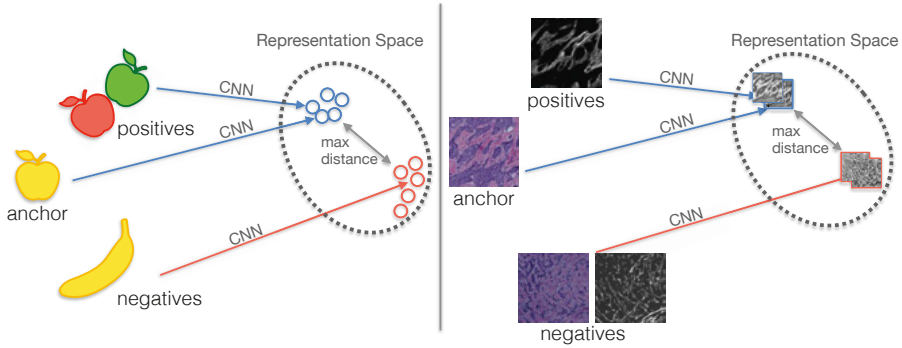


Figure 3.17. Contrastive learning finds an embedding in a representation space such that a data point that serves as an *anchor* is mapped to a representation that is close to the one of the so-called *positive*, which usually is either a representative of the same class (i.e., the class "apples" vs. "bananas"), or a different augmentation of the same object (i.e., different colors, or orientations of the same object), while the representation of a *negative* is far away to the representation of the anchor, as seen on the right. In Paper III, this is extended by learning 2D representations for a multimodal image pair, in which one modality serves as an anchor while its corresponding image in another modality acts as a positive.

in self-supervised learning. Generally, during training, one data sample is designated as an anchor in each iteration, and a positive and negative are defined with respect to this anchor. The positives can be other representatives from the same class, different views, or augmented versions of the anchor. Over the last few years, a family of methods [9, 52, 54, 101, 105, 123, 125, 141] which maximize the MI (or a lower bound thereof) has been established as a very effective tool. These methods are based on noise contrastive estimation (InfoNCE) [43]. Early work in this area mainly focused on sampling one negative example (triplet loss), but several studies showed improved performance using multiple samples or hard examples [9, 25, 54, 110, 125].

The most common usage of CL is to learn abstract 1D embeddings of the input data, i.e., high-dimensional vectors, which can be used in downstream tasks such as clustering or classification. In Paper III, however, we adjust the training principle to create 2D embeddings instead, matching the dimensions of the input images. The aim is to learn representations of multimodal images such that the representations preserve relevant structures. To generate CoMIRs, an image in one modality acts as the anchor, while its corresponding patch showing the same image but captured in the other modality is considered its positive. Any other image in the dataset in either of the modalities can serve as a negative. The two images in different modalities of one sample thereby act as different views of one sample in self-supervised learning. The principle of CL and its modified application in Paper III is shown in Fig. 3.17.

4. Contributions

In this chapter, I briefly summarize Papers I-V and the image processing tasks and applications they address. I discuss the methodologies I used in the papers, what they have in common and the overarching conclusions they share with respect to representation learning and information fusion in a non-chronological order.

4.1 Short Summary of Papers

4.1.1 Paper I

In Paper I, we propose an approach to automate the image acquisition process of kidney tissue samples in TEM, localizing particular structures called glomeruli. We suggest performing image classification on low-resolution TEM images to decide if they contain the structures of interest and should subsequently be imaged in high resolution or be discarded. In previous studies [119], it was shown that texture is an important characteristic of these structures and can be used for their recognition. Based on this observation, we show that classification performance can be boosted by incorporating the texture information in the form of LBP maps [75] over the use of the intensity images only.

4.1.2 Paper II

In Paper II, we investigate the feasibility of oral cancer screening programs by performing binary image classification of cells acquired by BF microscopy, originating from healthy and cancer patients. For this task, we evaluate different deep learning approaches incorporating texture information in their training regime [60, 75, 76], comparing their performance with general-purpose CNNs using different levels of data augmentation, pretraining, as well as a rotation equivariant network [88]. We show that CNNs which use texture information based on LBPs outperform the other evaluated approaches.

4.1.3 Paper III

In Paper III, we show that contrastive learning can be used to generate 2D representations of multimodal image pairs, such that the learned representations

are similar in intensity and share common structures. We call these representations *Contrastive Multimodal Image Representations* (CoMIRs) and show that they can be used for the downstream task of multimodal image registration. We evaluate them on a challenging dataset of BF and SHG microscopy images (see Sec. 3.1.3) and compare the representations' usability to bridge the gap between modalities to commonly used I2I methods. Furthermore, we introduce a hyperparameter-free modification to the contrastive loss to constrain the network to learn rotationally equivariant representations, a property of CoMIRs required for successful rigid registration.

4.1.4 Paper IV

In Paper IV, we address CoMIRs' usefulness for cross-modality image retrieval, i.e., if given an image in *modality A*, we can find its corresponding counterpart in *modality B*, given a repository of images in *modality B*. We propose to generate CoMIRs of the images in both modalities and then create a BoW by extracting SURF features, clustering them into global image histogram descriptors using k-means, and matching them using cosine similarity. To boost the retrieval performance, we propose re-ranking among the top results. We evaluate the pipeline on the set of BF & SHG microscopy images also used in Papers III & V, and perform a replacement study of different parts of the pipeline, demonstrating the necessity for rotationally equivariant image representations, and rotationally invariant feature extractors, if the images in the different modalities are not aligned.

4.1.5 Paper V

In Paper V, we evaluate if additional contrastive supervision on intermediate layers in the networks generating CoMIRs can improve the representations' quality for the downstream registration task, based on previous success on biomedical image classification in [61] using a similar learning regime. We test multiple approaches and critic functions for including such supervision on the bottleneck layer of the U-Net generating CoMIRs of the BF & SHG microscopy dataset used in Papers III & IV as well as QPI & FM images. We come to the conclusion that leaving the intermediate layers in the network unconstrained results in CoMIRs more suitable for feature-based, rigid registration.

4.2 Representation Learning with Applications to Multimodal Image Registration and Retrieval – Papers III-V

In Paper III, we introduce *Contrastive Multimodal Image Representations*, short CoMIRs, which are learned representations intended to overcome the semantic gap between multiple heterogenous modalities by extracting common structures and information between a multimodal image pair. We evaluate their usefulness for multimodal rigid registration in Paper III, for cross-modal image retrieval in Paper IV, and address potential ways to improve their quality for the downstream task of rigid registration in Paper V.

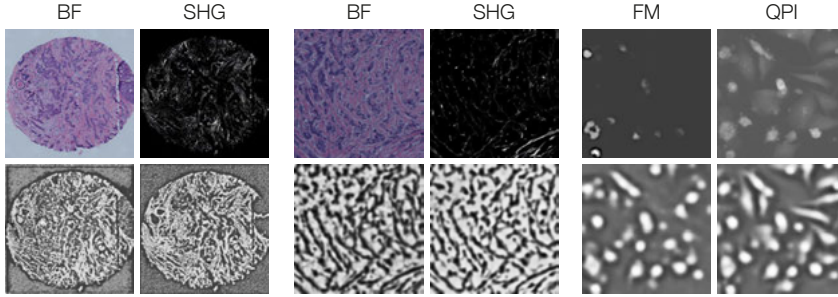
The main concept is to have a pseudo-Siamese network consisting of two U-Nets, one for each modality, connected by a contrastive loss acting on the final layer. The setup generates 2D, image-like, dense representations for both input images of different modalities. These CoMIRs are similar w.r.t. to a chosen similarity measure and capture shared structures across modalities. Some examples of aligned, multimodal image pairs and their corresponding CoMIRs can be seen in Fig. 4.1: in (a) a pair of BF and SHG images from TMA cores [32] is shown; in (b) an example of BF & SHG images used in Papers III-V is shown, which was created by cropping patches from the TMA pair in (a) as part of the registration dataset [31]; and in (c) an example of the QPI & FM dataset [84] used in Paper V is shown. The upper row of Fig. 4.1 shows the original images, and the lower row their corresponding CoMIRs.

To generate CoMIRs, patches of typically 128×128 px or 256×256 px are sampled randomly from the images which are processed during training. A trained model can however be applied to images of different sizes, i.e., the same trained model was used to generate the CoMIRs displayed in Fig. 4.1 in (a) and (b). The patch size for training has to be chosen considering a trade-off between contextual information, while being small enough to avoid averaging over too many pixels and fitting in the graphics processing unit (GPU).

The contrastive learning objective to generate CoMIRs can be formalized as follows. Let $\mathcal{D} = \{(\mathbf{x}_i^1, \mathbf{x}_i^2)\}_{i=1}^n$ be an independent and identically distributed set of n data points in form of aligned multimodal image pairs, where \mathbf{x}^j is an image in modality j , and f_{θ_j} the network processing modality j with respective parameters θ_j for $j \in \{1, 2\}$, s.t. $\mathbf{y}^j = f_{\theta_j}(\mathbf{x}^j)$ is the output of the network given \mathbf{x}^j . We use a contrastive loss function based on InfoNCE [101], which is given for an arbitrary multimodal image pair $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{D}$ by

$$\mathcal{L}_{\theta}(\mathcal{D}) = -\frac{1}{n} \sum_{i=1}^n \left(\log \frac{e^{h(\mathbf{y}_i^1, \mathbf{y}_i^2)/\tau}}{e^{h(\mathbf{y}_i^1, \mathbf{y}_i^2)/\tau} + \sum_{\mathbf{y}_j^1, \mathbf{y}_j^2 \in \mathcal{D}_{neg}} e^{h(\mathbf{y}_j^1, \mathbf{y}_j^2)/\tau}} \right), \quad (4.1)$$

where the exponential of a critic function $h(\mathbf{y}^1, \mathbf{y}^2)$ computes the similarity between CoMIRs $\mathbf{y}^1 = f_{\theta_1}(\mathbf{x}^1)$ and $\mathbf{y}^2 = f_{\theta_2}(\mathbf{x}^2)$ for the scaling parameter



(a) BF & SHG TMA Cores (b) BF & SHG images (c) QPI and FM images
cropped from TMA cores as shown in (a)

Figure 4.1. Examples of multimodal image pairs $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ in the top row and their respective CoMIRs $(\mathbf{y}_i^1, \mathbf{y}_i^2)$ in the bottom row.

$\tau > 0$ with respect to a chosen similarity measure and \mathcal{D}_{neg} a set of negative samples. The main novelty from how InfoNCE has previously been used is that we create 2D representations instead of abstract 1D vectors. Additionally, we modify the training regime based on this loss to learn rotationally equivariant representations. We require CoMIRs to be rotationally equivariant representations in order to use them for downstream tasks such as rigid image registration as is discussed in more detail in Sec. 4.3.2. We achieve this by commuting the network training and transformations from the finite, cyclic, symmetry group of multiples of 90° rotations \mathcal{C}_4 , i.e., we compute $h(\cdot, \cdot)$ in Eqn. 4.1 for

$$h(T'_1(f_{\theta_1}(T_1(\mathbf{x}_i^1))), T'_2(f_{\theta_2}(T_2(\mathbf{x}_i^2))))), \quad (4.2)$$

where $T_i, T'_i \in \mathcal{C}_4$ are sampled randomly at each iteration (here $T'_i := T_i^{-1}$). Limiting the group of transformations to \mathcal{C}_4 has the advantage of avoiding any interpolation of input images and CoMIRs.

4.2.1 Contrastive Multimodal Image Representations for Multimodal Registration and Cross-Modality Image Retrieval – Paper III - V

In Paper III, we show that CoMIRs can be used to overcome the challenges of multimodal image registration by enabling the use of monomodal image registration methods. We propose to generate CoMIRs of both images that are subject to registration and find the optimal transformation between the CoMIR pair, which can then be applied to the original multimodal images. This idea is outlined in Fig. 4.2.

To evaluate the CoMIRs' suitability for the use of monomodal registration methods, we create a multimodal registration dataset of BF & SHG images

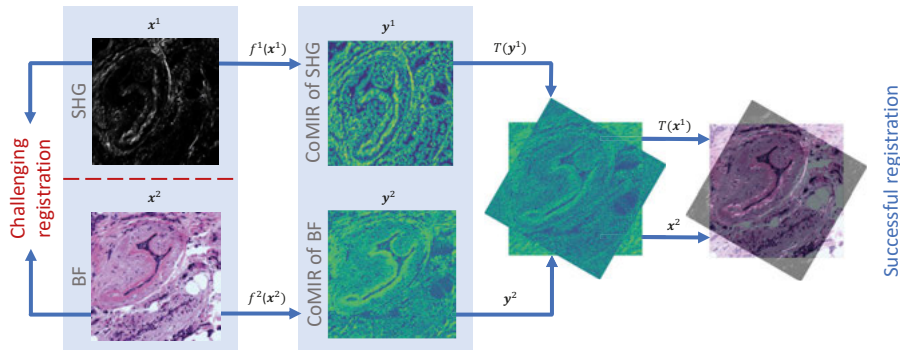


Figure 4.2. Figure from Paper III. We propose to generate shared representations – referred to as CoMIRs – by means of contrastive learning, which are similar in intensity and extract common features between a multimodal image pair, such that monomodal registration methods can be used to find the transformation between CoMIRs, which subsequently can be applied to the original images for multimodal alignment.

and provide it in [31] as described in Sec. 3.1.3 (example shown in Fig 4.1 (b)). The images were synthetically, rigidly transformed and registration success is calculated based on the average of the corner displacement between the reference and registered image, as explained in Sec. 3.2.2. Registration is successful if the error stays below a certain threshold. This threshold is determined considering the image size and the precision of the ground truth alignment of the images. In particular, for the BF & SHG dataset, we know that we cannot expect zero-pixel registration error as the images were not acquired in the same machine, and the alignment was done manually. We therefore also include a small study of human annotators in Paper III, highlighting the challenges of aligning this particular dataset (see the manual registration error in Fig. 4.3).

We show that both intensity-based monomodal registration methods, such as α -AMD [95] and monomodal feature-based registration methods, such as using SIFT features [81] with RANSAC [33] for feature matching perform well on the BF & SHG registration dataset [31] used in Papers III-V. The results of the registration experiments of this study are shown in Fig. 4.3 from Paper III. It shows the empirical cumulative distribution function of the successful registrations as a function of the error for α -AMD (with single and multi-start), as well as using SIFT. $A \rightarrow B$ denotes the instance that BF serves as a reference image and SHG as a floating image (or their corresponding CoMIRs), $B \rightarrow A$ that SHG served as the reference and BF as the floating image, respectively. It can be seen that registration using either α -AMD or SIFT on CoMIRs outperforms CurveAlign, a state-of-the-art registration method for BF & SHG in particular, as well as local optimization of MI.

In Paper V, in which we study modifications to the training configuration to generate CoMIRs, we furthermore evaluate feature-based rigid registration

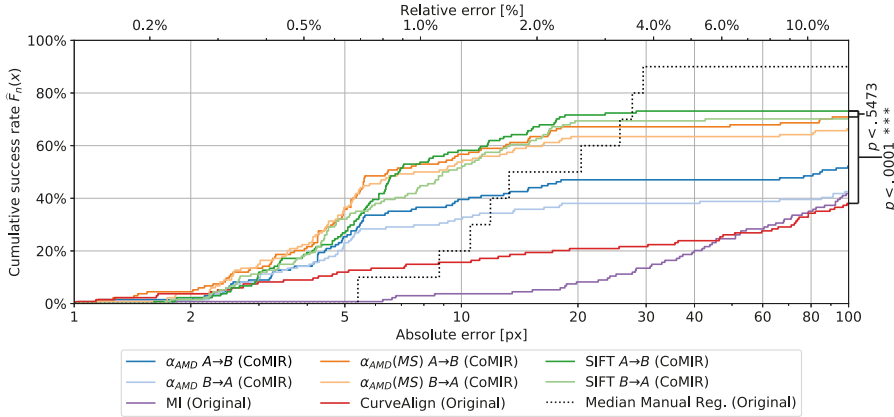


Figure 4.3. Figure from Paper III showing the cumulative number of successful registrations for the increasing error shown as an empirical cumulative distribution function of the different registration methods over the histological test set of BF & SHG images. The results are compared to the median of the error of six independent manual registrations on a subset of ten images. A Wilcoxon signed-rank test highlights the statistical differences between CurveAlign, SIFT, and α -AMD.

of CoMIRs using SIFT on the cytological dataset of QPI & FM images [84]. We reach a 62% registration success rate on this dataset, though a better performance on this dataset using CoMIRs is reached in [82].

In Paper IV, we address if CoMIRs can be used for image retrieval across modalities of non-aligned image sets in the form of a reverse image search. The multi-stage pipeline is depicted in Fig. 4.4. We propose to generate CoMIRs of both the query and the searchable repository and in a second step, create a CBIR as described in Sec. 3.2.3, by extracting Speeded up robust features (SURF,[12]) and bin them to global image descriptors by forming a histogram using K -means clustering, which can be matched using cosine similarity. As a final refinement step of the search, we suggest performing reranking among the top-30 retrieved images by repeating the steps to create a bag of words among those top candidates. We evaluate the proposed pipeline on the BF & SHG dataset used in Papers III-V. We show that using CoMIRs in the first stage of the pipeline is superior to other image translation techniques. Comparison to two other state-of-the-art image retrieval methods shows the task is very challenging. The proposed pipeline achieves 75.4% Top-10 retrieval success to retrieve BF images within a set of SHG images and 83.6% Top-10 retrieval success to retrieve SHG images within BF images, vs. 35.8% and 43.4%, respectively, for the best performing competing method.

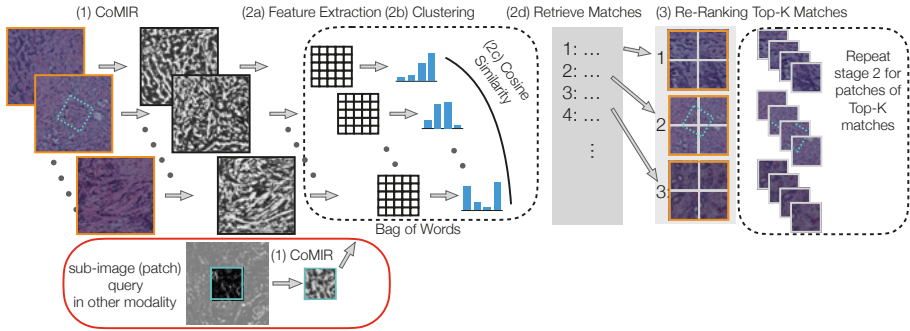


Figure 4.4. Figure from Paper IV showing the three stages of the s-CBIR pipeline. In Stage I, CoMIRs are generated for the images in the repository and the query (either as a full-sized image or in the form of a patch), followed by sparse feature extraction in Stage II. The sparse SURF features are binned into single global descriptors for each image using K -means clustering, creating the vocabulary for a BoW. The histogram descriptors are matched using cosine similarity. In Stage III, the Top-K matches are split into patches, and a new BoW is computed for Re-Ranking.

Comparison to Generative Adversarial Networks for Multimodal Images

In both Papers III & IV, we compare CoMIRs’ ability to bridge between different modalities to that of modality transfer methods in the form of GAN-based I2I. Figure 4.6 from Paper III shows examples of the representations generated by the GANs pix2pix, CycleGAN, and DRIT (described in Sec. 3.2.4) for a particular SHG & BF image pair from the test set, as well as corresponding CoMIRs. The green arrows indicate which pairs are subject to registration performed in Paper III. We find in Paper III that the image pairs resulting from I2I lead to poor registration performance using both intensity- and feature-based registration methods, an observation further confirmed in an extensive study by Lu et al. [82].

In Paper IV, we evaluate the image representations generated by CycleGAN and pix2pix in the CBIR in the form of a replacement study to substitute CoMIRs in the modality transfer stage of the pipeline on the same challenging dataset.

In one of the experiments in Paper IV, we evaluate their ability in image domain migration by performing a reverse image search using a query given in a *fake modality A*, searching the repository of aligned images in *modality A*, and vice versa. In the case of CoMIRs, the query is the CoMIR of the query image in *modality A*, and the searchable repository is the set of aligned CoMIRs of *modality B*. This is a synthetic experiment to highlight the difference in modality migration potential between CoMIRs, CycleGAN, and pix2pix representations, not considering any rigid transformations between the query and its match. The results of this experiment are shown in Fig. 4.5.

We find that neither pix2pix nor CycleGAN results in representations of sufficient quality for the cross-modality retrieval task, even for this less chal-

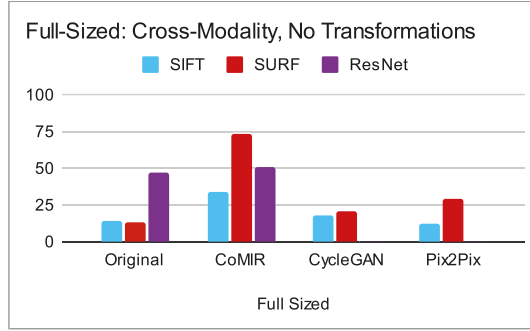


Figure 4.5. Results from Paper IV showing the superiority of CoMIRs to facilitate domain migration compared to CycleGAN and pix2pix for the histological dataset of BF & SHG images. It shows the percent of successfully retrieved images within the Top-10 matches across modalities when the query and its match in the other modality are aligned. These results do not include re-ranking and are averaged over retrieval directions (BF query within SHG and SHG query within BF or their respective representations) for different combinations of images or their learned representations and feature extractors.

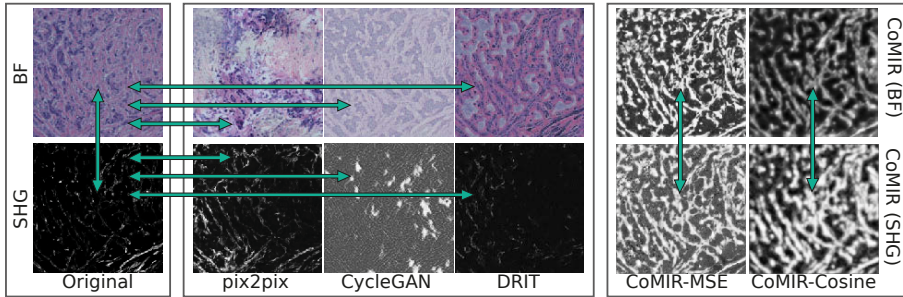


Figure 4.6. Figure from Paper III showing an example BF & SHG image pair from the test set and its image representations generated by pix2pix, CycleGAN, DRIT, and CoMIR using MSE and cosine similarity as a critic function. The arrows indicate the different image pairs which were registered in Paper III.

lenging task which does not consider rigid transformations between the query and its match. Fig. 4.5 shows the retrieval results without any re-ranking. As can be seen, the retrieval performance using CycleGAN or pix2pix is insufficient, even if the image pairs in the datasets are aligned.

Furthermore, we find that, in particular, CycleGAN suffers from mode collapse. In Fig 4.7 from Paper IV, three examples are shown for which the fake BF modalities (middle image in rows 2,4 and 6) are highly similar, although they originate from different input images (indicated by the blue arrows). The structures of the real BF images are not preserved, but the discriminator accepts the generated texture as reasonable BF tissue. In the third row of Fig. 4.7, the reconstructed images based on these generated fake BF images are shown,

highlighting that the cycle consistency was fulfilled, even though the fake image shows generic BF imitating texture, lacking significant structures present in the real counterpart.

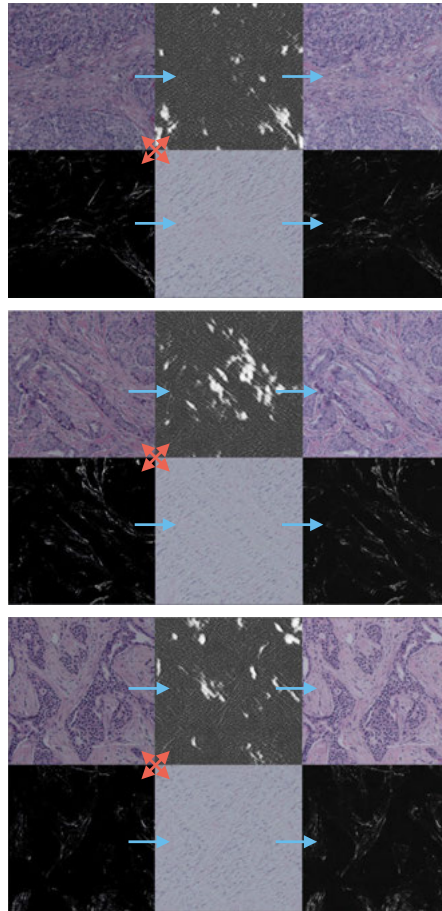


Figure 4.7. Figure from Paper IV, showing that CycleGAN suffers from mode-collapse. The first column shows original BF & SHG images, the second shows CycleGAN's representations in the respective other modality given those originals (indicated by the blue arrow). The third column shows the reconstructed originals. Red arrows indicate image pairs which should be similar. It shows that the fake BF images (even rows, middle) do not preserve the structure and appearance of the corresponding real BF images but appear similar, independent of the content of the SHG images they are generated from.

4.3 Information Fusion – Regulating Feature Properties to Aid Data-Driven Approaches

The essence of data-driven approaches is that they learn relevant features for a downstream task, such as classification from the data itself. Deep neural networks are able to learn hierarchical, i.e., increasingly abstract and complex features. However, the features' robustness and ability to generalize highly depend on the amount of data available to train on and learn from, which is often very limited in the biomedical domain. In order to exploit these limited datasets to the fullest extent, several measures can be taken: (i) data augmentation can be applied to enlarge the dataset artificially; (ii) pretraining on other datasets or tasks can be used to learn useful features inherent to images of many domains; (iii) feature properties can be guided and regularized based on knowledge about the data at hand, as it has been done in handcrafted features before the advent of deep learning.

Histological image properties resemble those of texture images as described in Sec. 3.1.3. This similarity and the fact that texture was the feature of choice in classic image processing on similar tasks is the intuition behind the additional utilization of texture descriptors to complement CNN training in histological image classification in Papers I and II. The fact that histological images are inherently symmetric under rotation and reflection [42, 127] motivates the use of equivariant networks, which have shown to improve performance in many biomedical classification [42, 70] and segmentation [42, 70, 88, 127, 138] tasks. The impact of imposing, in particular, rotational equivariance on learned features for different downstream tasks is studied in Papers II, III, and IV.

4.3.1 Focus on Texture Features – Papers I & II

Paper I presents a deep learning approach for Glomerulus detection in TEM images of kidney tissue by performing binary classification of low-resolution images. Previous studies [119] showed that glomeruli in BF microscopy images could be recognized and detected by a variant of LBP texture descriptors described in Sec. 3.3.3. Inspired by this finding, we evaluate if texture features can be used in combination with CNN features for the glomeruli classification task. In order to use LBP features within CNN training, we create LBP maps as discussed in Sec.3.3.3. Following [8, 75], we create 3-channel dense LBP maps, which can be fused with the intensity TEM images for CNN training. We show that explicitly providing these texture feature maps to the network can boost classification performance over using only intensity images, improving the classification accuracy from 91% to 97% in the case of VGG16 and 96% to 98% in the case of ResNet50.

In Paper II, we address the task of cell nuclei classification of BF microscopy images for oral cancer detection. It has been shown that for cytologi-

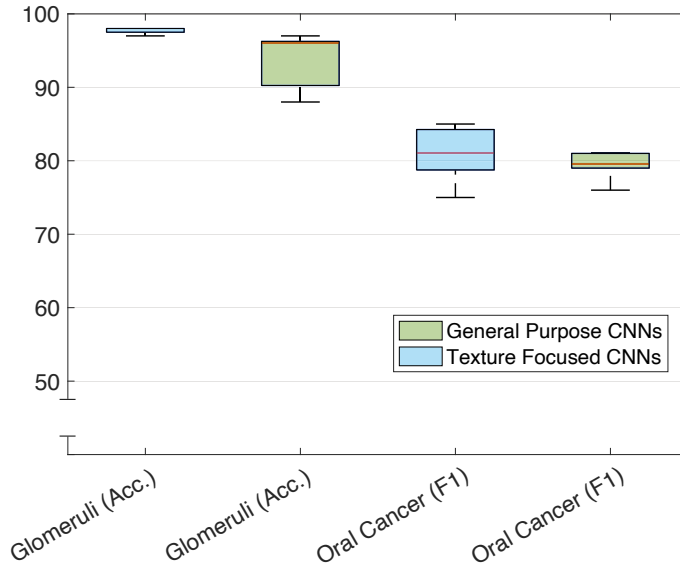


Figure 4.8. Summary of results from Papers I and II using general-purpose CNNs for classification that only use intensity information versus texture-focused CNN approaches.

cal cancer classification, that chromatin texture is among the most discriminative features [57, 144]. Building on work of [139], which has demonstrated the ability of general-purpose CNNs, such as VGG [120] and ResNet [50], to discriminate between cells from the oral cavity originating from cancer or healthy patients, we evaluate a number of approaches to incorporate the knowledge of the texture’s importance into the CNN training: (i) the approach of using LBP maps as proposed in Paper I, i.e., training two CNNs, one using the intensity images as input, the other the corresponding LBP maps, concatenating their softmax output and training a linear SVM; (ii) LBCNN as proposed in [60], in which LBP computation is modeled by N fixed sparse and thresholded filters, combined with trainable weights, extracting LBP-like features and serving as a drop-in replacement for any convolutional layer; (iii) and a CNN with an explicit LBP extraction module as proposed in [76], which extracts LBPs within the network in a differentiable way.

We show that texture-focused approaches outperform general networks on oral cancer cell classification with respect to F-1 Score, even when pretraining and data augmentation are used. The best performing texture-incorporating approach achieves 81% accuracy and 85% F-1-Score, in comparison to the highest accuracy among all general-purpose CNN configurations evaluated, reaching 80% (VGG16, as reported in [139]), and 81% F-1 Score using ResNet50 with data augmentation of flips and rotations by multiples of 90° ,

avoiding interpolation of the images.

In both studies, conducted in Paper I and Paper II, we observe improved classification performance when combining the pure learning approaches of CNNs with a particular focus on texture features. Figure 4.8 shows a summary of the performed experiments in Papers I & II contrasting CNN training on intensity images with general-purpose CNNs exclusively versus additional training on LBP maps or including texture extracting modules (texture-focused CNNs) within the CNNs. For the Glomeruli classification task, accuracy is reported. For the oral cancer cell classification, the F1-score is reported. In the case of Glomeruli classification, the plot shows the results of the experiments for VGG16 and ResNet50 using transfer learning and training from scratch as *general-purpose CNNs*, and all eight experiments of late fusion of intensity and LBP maps of different LBP scales summed up as *texture-focused CNNs*.

4.3.2 Equivariant and Invariant Features – Papers II - V

Images are high-dimensional inputs. To counteract the curse of dimensionality, we have geometric priors as discussed in Sec. 3.3.4, i.e., additional knowledge about the local structures and spatial connectivity between pixels in the image. A small translational shift of an image in x or y would result in a very different flattened 1D vector, while basically all the spatial connectivity would be preserved in the 2D image. To compensate and learn this translational invariance from the data for classification tasks, the network has to be provided with very many samples to learn that all shifted versions of one data point belong to the same class – or the equivariant and invariant properties of the data at hand can be taken into consideration for the model design.

In Paper II, we do not only evaluate general-purpose and texture-focused CNNs, but also a rotation equivariant Vector Field Network proposed in [88], in which standard convolutional filters are modified to become rotationally equivariant. Since the downstream task is classification, i.e., the features of two rotated versions of the input map to the same label, the network model becomes a rotation invariant function, as discussed in Sec. 3.3.4. In the task of cell classification studied in Paper II, we did not observe an improvement when using rotation equivariant vector field networks over general-purpose CNNs, or CNNs incorporating texture information. Generally, it has been shown that equivariant networks which learn features to respect the symmetries of a group, generalize better when these symmetries are present in the data [113, 134, 137], as is the case in the data of Paper II. However, the model proposed in [88], which we adopted for the oral cancer cell classification task, was relatively small compared to the evaluated general-purpose CNN approaches ResNet50 and VGG16. It was originally trained in [88] for the classification of images of size 28×28 px and was adopted for the cell images

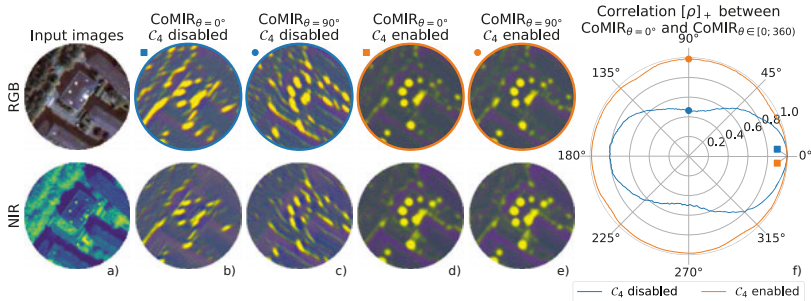


Figure 4.9. Figure from Paper III showing that rotational equivariance for arbitrary degrees beyond multiples of 90° is achieved by the proposed method.

used in this study, which were 80×80 px. The reduced capacity in comparison to the general-purposed networks may explain the lower performance w.r.t. to accuracy and F1-Score.

In Papers III and IV, we show the explicit need for rotation equivariant features in the downstream tasks of image registration and retrieval. In Paper III, we show that contrastive learning can be used to generate common 2D image-like representations for input images of different imaging modalities, which we call *Contrastive Multimodal Image Representations* (CoMIRs). In Paper III, we show their usefulness for multimodal registration, in Paper IV for cross-modality image retrieval, and Paper V builds upon improving the representation quality for the downstream task of registration. The tasks of image registration and retrieval differ from the classification task in Paper II, as that in the case of classification, the codomain of the CNN function is a set of scalar valued points in the form of class labels, which ultimately turns the CNN function invariant, even if all layers within the network are equivariant mappings. In contrast, in the case of CoMIRs, we generate 2D representations which generally can preserve the equivariance for the entire mapping.

Unlike many other rotational equivariant networks, which formulate a network as an equivariant mapping as a composition of equivariant layers, we do not constrain the network to learn equivariant features in each layer, but propose to modify the training regime in the contrastive learning of CoMIRs, such that the two streams of the network see differently rotated versions of the corresponding multimodal images and invert the rotation in the final layer before evaluating the loss on these features. This modification to the contrastive loss imposes that the CNN commutes with the rotation without requiring any additional hyperparameters. Note that this differs from using data augmentation to achieve equivariance. In segmentation tasks, for example, a certain level of rotational equivariance of the features can be achieved by rotating the input and segmentation labels. This equivariance learning approach works because the target output of the network (the segmentation label) is rotated with the

input, establishing a relationship between rotated samples of the same image. In contrastive learning however, there is no label map which can be rotated with the input. Instead a rotated input image pair will be presented to the network as a sample independent to an unrotated version of the same image pair. Hence, rotation equivariance of the contrastively learned representations can only be achieved either by network design or adjusting the training regime as we propose in Paper III, given in Eqn. 4.2.

We limit the rotations to multiples of 90° to avoid any interpolation of the input or feature maps. We observe that this choice results in sufficient rotational equivariance for angles beyond multiples of 90° , which can be seen in Fig. 4.9, in which the effect of the rotational equivariance constraint on the learned CoMIRs of an image pair (RGB and NIR) of the Zurich dataset [131] is shown. Column **a**) shows the two corresponding, aligned input images in RGB and NIR. The images in column **b**) show the CoMIRs of these input images at $\theta = 0^\circ$. Column **c**), shows the CoMIRs for the same input pair when the input images are rotated by 90° and the network is unrestricted regarding rotational equivariance. For visualization purposes, the CoMIRs are rotated back to 0° , i.e., are presented in a stabilized view to align with the CoMIRs in the remaining columns. The CoMIRs between columns **a**) and **b**) vary greatly because the network learned the connection between samples across modalities but not the relation to the same sample pairs in other orientations. In contrast, columns **c**) and **d**) show the resulting CoMIRs of the unrotated and 90° -rotated input when the rotation equivariance is maintained during the training by commuting the network and rotations. As can be seen, the CoMIRs of rotated input are rotated CoMIRs of unrotated input, i.e., rotationally equivariant feature maps. Finally, in column **f**) of Fig. 4.9, we measure the positive correlation between the CoMIR of an input image and the CoMIR of the rotated input image for all $\theta \in [0^\circ, 360^\circ)$, for which the generated CoMIR has been rotated back to 0° . The correlation for the unconstrained network is shown in blue, reaching 1 only at 0° . For the rotational equivariance constrained network, it is shown in orange, resulting in a correlation very close to one for all degrees.

The downstream task studied in Paper III is rigid registration of multimodal images, i.e., the images in the test set are expected to be rigidly misaligned, and the consistency of representations across these transformations is essential for the performance.

In Paper IV, we study if CoMIRs can bridge between modalities in the task of cross-modality image retrieval. Retrieving an image in *modality A*, given its corresponding counterpart in *modality B*, can serve as a first step in image registration pipelines to find the images or regions which should be registered since many registration methods are based on the assumption that the identities of the reference-target image pairs are known, which is not always the case in microscopy [6]. As the images are assumed to be captured by different machines, they can be expected to be rigidly misaligned.

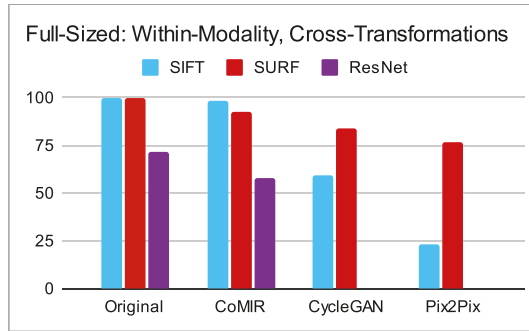


Figure 4.10. Results from Paper IV showing the percent of successfully retrieved images within the Top-10 matches without re-ranking for different combinations of images or their learned representations and feature extractors when the query image and the images in the searchable repository are *not rigidly aligned*, but in the *same modality*. The drop in performance using ResNet as a feature extractor is due to it not being rotationally invariant, unlike SIFT and SURF. The drop in performance using SIFT and SURF for the I2I approaches is due to the lack of rotational equivariance of the GAN-generated 2D image representations.

If the searchable repository consists of images in *modality A* (in the experiments in Paper IV, they are BF or SHG microscopy images) and the query image is in the same modality and is aligned with its match in the repository (i.e., we retrieve the identical image from the repository), we report a 100% top-10 retrieval success rate for all combinations of representations (original modalities, CoMIRs, I2I approaches) and feature extractors (SIFT, SURF, ResNet). However, if the query is rigidly misaligned from its corresponding match in the repository while still in the same modality as the images in the repository, we observe the necessity for rotational equivariant representations and invariant feature extractors. In Fig. 4.10 we see this experiment’s top-10 retrieval success rate. Extracting SIFT and SURF features which are by design rotational invariant, on the original microscopy images, remains at 100% top-10 retrieval success. However, using a pretrained ResNet as a feature extractor results substantially drops to 72%. We hypothesize that this is caused by the ResNet features differing between rotated versions of an input image.

Furthermore, we see in Fig. 4.10 that the retrieval success for the evaluated I2I approaches drops drastically when querying the fake image of a rigidly transformed query image within the repository of fake images of the same modality. The performance in this experiment drops even when using the rotation invariant SIFT and SURF features. This means that the performance drop is attributed to the lack of rotational equivariance in the fake representations. The I2I approaches, in combination with a ResNet feature extractor, resulted in fewer retrieval matches than random selection.

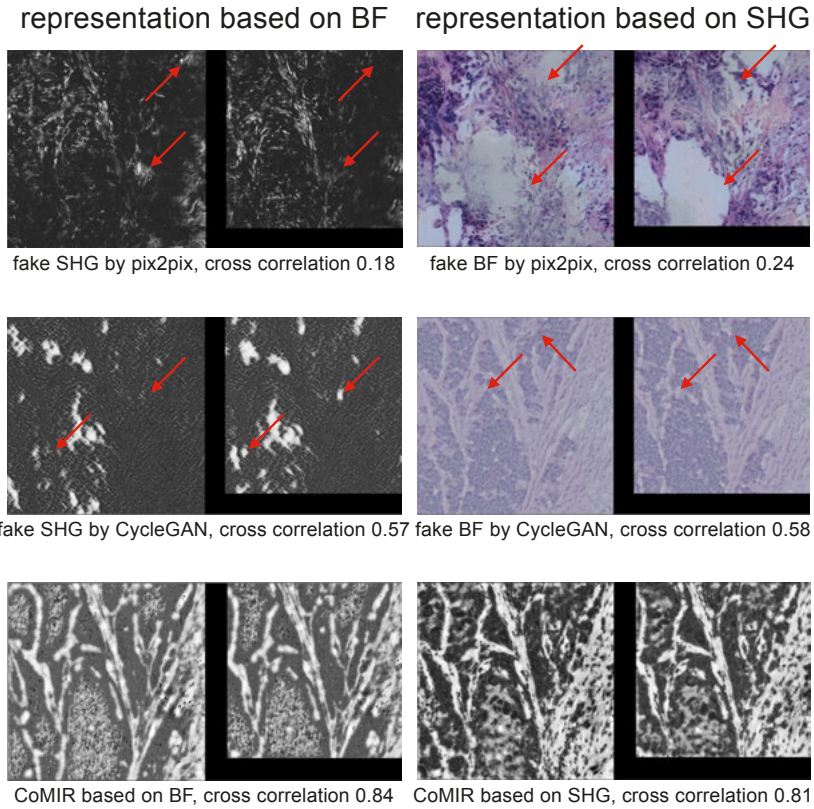


Figure 4.11. Figure from Paper IV showing image representations generated by pix2pix, CycleGAN, and CoMIR. The image pairs to the left show the learned representations for one BF image in the test set, and the image pairs to the right the representations of the SHG image. The first image in each pair originates from an image that has not undergone any transformation. The second originates from rigidly transformed input and has been aligned to the representation of the untransformed input. The 2D correlation coefficient for their overlapping area is given. Red arrows point at locations in the images in which structures clearly differ.

Figure 4.11 shows an example of the GAN and CoMIR representations for an example pair of a BF and SHG test image pair. The representations were generated by inference on an unrotated and a rigidly transformed image pair, which were aligned manually after their generation. The figure shows the representations of the BF input to the left and the representations of the SHG input on the right. The red arrows mark areas which differ between the representations of the untransformed and transformed input. The Pearson Correlation Coefficient (PCC) is calculated between the pairs of representations of untransformed and transformed input calculated on the overlap after alignment (i.e., discarding the border effects). It shows that consistent features are preserved in CoMIRs over rigid transformations in the input data.

4.3.3 Fusion Strategies for Heterogeneous Input – Papers I & V

Although the focus in this section is on Papers I & V, we touch upon strategies to combine heterogenous image information to improve the performance of the downstream task in all papers. In Paper I, we evaluate an approach to process LBP maps together with their intensity counterparts (also used in Paper II). Considering these LBP maps as an abstract and generated "modality", the setting is comparable to the multimodal learning we face in Paper III (and the following work based on this in Papers IV-V), in which we process multiple imaging modalities.

In both Papers I and III, we use two networks that are trained in parallel on the two heterogenous modalities to produce features that combine relevant information about the downstream task present in each of the respective modalities. However, the corresponding modalities in the two papers differ, as the LBP maps are computed from the intensity input and are hence aligned and ready to be fused. In Paper III, and followingly Paper V, the two modalities are captured by different sensors and need to be aligned prior to image fusion. The task addressed in Papers III & V is precisely to solve this image alignment by learning representations that are easier to align than the original modalities. The two studies hence differ in the training regime and downstream task: supervised binary classification in Paper I and contrastive learning to produce representations for registration in Papers III & V, as well as the used network architectures. In Paper I, one dimensional feature vectors are produced by VGG16 and ResNet architectures. In Papers III & V, two dimensional feature maps are generated using U-Nets. Despite the differences in the objective and downstream tasks, we observe in both Paper I and Paper V that early fusion of heterogenous input can be detrimental to the learning.

Early, Mid and Late Fusion

In Paper I, we test multiple fusion strategies to combine the precomputed LBP maps with the intensity input: early, mid and late fusion.

In the **early fusion** approach, the intensity image is stacked with 3-channel LBP maps and processed together by the networks.

In the **middle fusion** approach, we perform *probabilty fusion*, i.e., we use a two-stream architecture, in which both networks (VGG16 or ResNet50) are trained independently on the intensity images and LBP maps. Once the networks are trained, the 4096-dimensional feature vectors at the end of the network are concatenated, and a linear support vector machine (SVM) is trained for the classification of the resulting 8192-dimensional feature vector.

In the **late fusion** approach, we perform *feature fusion*, i.e., two CNNs are trained independently in parallel on the intensity images and LBP maps, and the output probabilities of the softmax layer are concatenated into a 4-dimensional vector, classified by a linear SVM.

We observe very consistently over all experiments that extracting features for both modalities independently and fusing them at a late stage in the network (as done in both the middle or late fusion approach) yields better classification behavior than processing both modalities jointly in one network. While the classification performance for early fusion ranges between 72% and 86% in accuracy among all experiments on the glomeruli classification task in Paper I, the middle fusion approach improves the accuracy to 97%, and the late fusion to the range of 97% to 98%.

In Paper V, we also study the effect of fusing the information of the two input modalities at different stages of the dual-stream network. We investigate approaches to improve CoMIRs by using additional contrastive losses on the bottleneck layer of the U-Nets used in the training. The idea of additional contrastive losses is based on recent work done by Kaku et al. [61], which showed on three biomedical (monomodal) datasets that bringing the representations of intermediate layers of an image closer together earlier in the network, improves the momentum contrastive method in self-supervised learning. We evaluate multiple approaches on how to incorporate additional supervision of the latent representation in the bottleneck of the U-Net, among them the *alternating loss* and *summed loss* approach, in which the InfoNCE loss used in the contrastive learning is not only evaluated on the final layer (the CoMIRs) as $\mathcal{L}_C(\mathcal{D})$, but also on the latent bottleneck features as $\mathcal{L}_{BN}(\mathcal{D})$. In case of the *alternating loss*, the two losses are evaluated in an alternating manner, in case of the *summed loss* $\mathcal{L}_C(\mathcal{D})$ and $\mathcal{L}_{BN}(\mathcal{D})$ are added in a weighted manner. Thereby the features of the heterogenous input are being contrasted already in the bottleneck layer of the U-Net, fusing the information present in the different modalities earlier to learn common representations. In the performed experiments, we evaluate three different critic functions to evaluate $\mathcal{L}_{BN}(\mathcal{D})$ on the BF & SHG dataset also used in Papers III & IV, and use the best performing critic function to confirm our main observations on the QPI & (FM) dataset. The quality of the CoMIRs is measured in particular by their usefulness for the downstream task of registration. We observe that leaving the features unregulated until the final layer of the network results in representations most useful for registration.

In Fig. 4.12, the summary of some of the experiments of Papers I & V are shown, which relate to fusing the information of the dual stream networks in the early or late stages of the training. It shows the accuracy on the glomeruli classification task in Paper I using either early fusion or late fusion for the experiments with VGG16, ResNet50, and all considered LBP scales (in total 6 experiments for early fusion and 6 experiments for late fusion averaged over seven runs). It also summarizes the results of Paper V in terms of registration success rate (RSR), defined as the percentage of images in the test set whose CoMIRs were successfully registered, where success is defined by a registra-

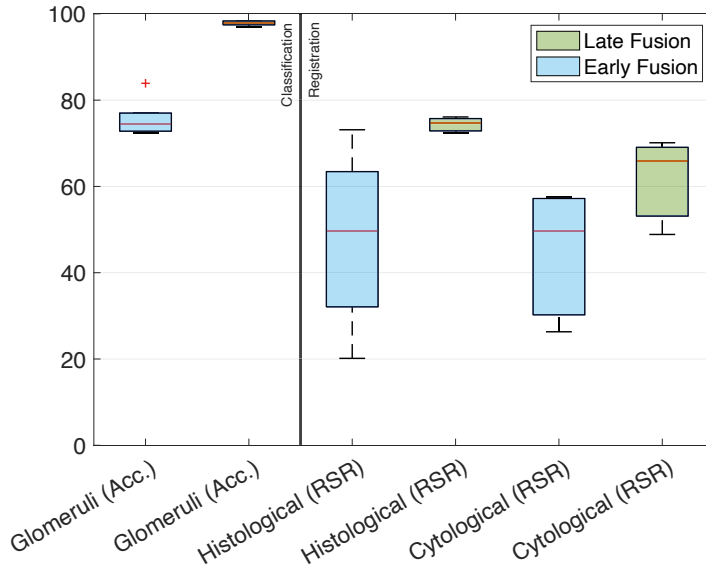


Figure 4.12. Summary of results from experiments from Papers I & V studying different timings of fusing heterogeneous input modalities in CNNs. For the glomeruli classification task of Paper I, classification accuracy (Acc.) is reported, while for the registration tasks in Paper V, the learned representations are evaluated by feature-based rigid registration, and performance is reported as registration success rate (RSR).

tion error smaller than a chosen (dataset-dependent) threshold. *Late Fusion* in the CoMIR context refers to CoMIRs as learned in the original publication of Paper III, i.e., in which the two parallel networks are trained in a true pseudo-Siamese fashion and are unregulated in their features throughout the network until the last layer on which the contrastive loss is acting. *Early Fusion* in this context refers to the experiments conducted within Paper V, in which the attempt is made to combine information from the two input modalities already in the bottleneck of the U-Net by having additional contrastive losses on this intermediate layer (*alternating loss* and *summed loss*). The RSR shown in Fig. 4.12 on the histological dataset refers to the experiments performed on the BF & SHG dataset and encompasses the experiments with three different critic functions (in total 6 experiments with 3 runs each). The results on the cytological dataset refer to the QPI & FM dataset and entail experiments using the best-performing critic function from the experiments on the histological dataset (in total two experiments on three different folds of the dataset each).

In Paper V, we go further and visualize the relationship of the intermediate features. We use MDS optimizing Sammon’s stress criteria to map the high dimensional bottleneck features into 2D. Figure 4.13 shows the bottleneck features of three runs to generate CoMIRs of BF & SHG images which do not have any additional loss on the bottleneck (i.e., are generated as intro-

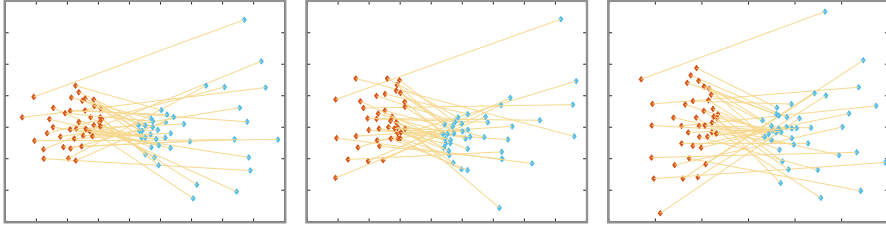


Figure 4.13. Metric MDS embeddings of bottleneck features during three independent CoMIR trainings without intermediate losses inside the network, i.e., features are only brought together at the last layer. BF features are marked by blue diamonds, SHG features by red diamonds, corresponding samples are connected by a yellow line.

duced in Paper III). Blue diamonds mark BF features, SHG features by red diamonds, and yellow lines connect corresponding samples. We can see that the networks tend to learn features that are more similar to other features of the same modality rather than those originating from the same sample in the other modality.

We see clear indications in two different learning regimes that early fusion of heterogenous input can be detrimental to the performance of the downstream task. We stress that representation learning approaches have to be developed for a particular application in mind and cannot be expected to generalize to multimodal settings and/or other downstream tasks.

5. Conclusions

Deep Learning has contributed significantly to the progress and spread of AI tools in many application domains, especially in computer vision, where it has excelled in object detection and recognition tasks. Biomedical applications, however, suffer from small datasets and limited annotations, and fail to feed these data hungry models sufficiently to fully exploit their potential. In this thesis, we present several ways to still take advantage of machine learning methods by using additional information about the data at hand, as was the custom in image processing method development prior to the advent of deep learning.

In Paper I, we discuss how handcrafted texture descriptors extracted from images can be transformed to serve as additional CNN input. We evaluate this approach on a classification task of kidney tissue captured by TEM. We explore different fusion strategies of the information residing in the images and the derived texture representations. We conclude that late fusion results in superior classification performance.

In Paper II, we compare the approach from Paper I with network architectures that learn similar texture features within the CNN training and evaluate these approaches on a classification task of cells from the oral cavity to recognize cancer patients. We show that texture-focused CNNs outperform general-purpose CNNs, even when allowing extensive data augmentation and pretraining. In both studies, we observe that using additional texture information results in superior image classification performance compared to leaving it up to general-purpose CNNs to learn relevant features entirely by themselves in an unconstrained way.

In Paper III, we introduce a representation learning technique to learn common representations for images originating from different sensors. We show that these representations are useful in reducing the challenging task of multi-modal image registration to a monomodal one. We present a simple and clever modification to the contrastive learning regime to exploit geometric properties within the dataset, such that the function learned by the network generates rotationally equivariant representations.

Paper IV shows that these representations can also be used for domain migration in cross-modality image retrieval when combined with rotationally invariant feature extractors stemming from classic image processing.

In Paper V, we attempt to improve these representations for the downstream task of registration by further regulating the features within the network and fusing information from the two modalities earlier on in the training. However, we learn that this approach is detrimental to the registration performance.

Further analysis of the representations shows that fusion of the heterogeneous network input is preferable later in the training, similar to observations made in a different context in Paper I.

The work presented in this thesis demonstrates that deep learning networks can learn very useful features from the raw data themselves but often perform best in connection with content-aware design to exploit the underlying data structures or by regulating feature properties in order to emphasize image characteristics which are known to be important based on human expertise or classic image processing concepts.

5.1 Successive Related and Future Work

In this section, I discuss some independent, related work that was developed either at the same time or following the studies performed during this thesis. This work either confirms the observations discussed in the studies of this thesis, describes advancements on related topics or presents developments that provide interesting potential future work following the papers presented in this thesis.

Texture-focused Deep Learning

Several studies followed after the publication of Papers I & II, which confirmed our observations that deep learning based methods can benefit from incorporating texture information explicitly.

The work by Majtner et al. [86] incorporates texture information in the form of hand-crafted granulometry-based descriptors [122] to CNN training and showed improved classification performance on texture datasets such as KTH-TIPS2-b [87] and the Virus Texture dataset [69]. The fusion strategy we propose in Paper I performs well in their approach but is outperformed by multiplying corresponding softmax probabilities for the intensity image and its corresponding texture descriptor.

Another study confirming that texture descriptors carry complementary information to learned features is presented in [10]. The authors propose using an improved, fuzzy version of LBP descriptors, namely α -LBP features [99], and present a way to use the resulting fuzzy patterns as CNN input without the need to map them into a low dimensional metric space using MDS as is proposed in [75] and applied in Papers I & II. The information from the intensity and texture sources are fused by creating a weighted ensemble method using Nelder–Mead simplex search [71], boosting classification performance over learning features from only intensity input on three datasets for human epithelial type 2 (HEp-2) cell image classification [34, 55, 140].

In another study on this subject, the authors in [5] fuse LBP maps as used in Paper I and RGB images by adding the two inputs and processing them together (versus stacking the inputs as done in the early fusion approach pre-

sented in Paper I) and improve classification performance for diabetic foot ulcer detection.

Also, the work presented in [7] confirms our findings that LBP maps as we use them improve accuracy over using image input alone. Like us, the authors compare early fusion by stacking the heterogenous input and late fusion. They show that late fusion performs consistently better than early fusion on three scene datasets [65, 111, 143]. It is worth noting that in this study the datasets are much larger than the aforementioned biomedical datasets, varying between 15 324 and 108 754 images. The improvement in classification performance is nonetheless significant when using the texture information, which indicates that it is not only small datasets that can benefit from being provided complimentary information abstracted in a semi-handcrafted manner.

Oral Cancer Screening

A number of studies have followed the initial work of oral cancer cell classifications in [139] and Paper II. The dataset has since then been extended, and a fully automated pipeline for single cell extraction, focus level selection, and cell classification is presented in [83]. According to [44], a review on prospective screening methods for oral cancer, the work in [83] shows great promise to reduce human labor significantly and, at the same time, to make the analysis more accurate. Further work by Bengtsson Bernander et al. [14] focuses on using rotation equivariant networks to improve classification accuracy and reduce overfitting on the oral cancer classification task. Koriakina et al. [67] study which cells are most relevant for the machine-aided diagnosis and thereby pave the way to unravel diagnostic patterns in the images even medical experts were previously not aware of. Data collection on this study is in full swing within Sweden [1, 2, 3], an essential next step to ensure robustness and generalization of the trained AI models to unseen data.

Representation Learning

Representation Learning is a highly active field of research – in 2023 alone, the International Conference on Learning Representations (ICLR) received 4922 submissions for peer review [4]. The community is gaining new insights into the underlying dynamics quickly. One effect of GANs, such as CycleGAN and pix2pix, becoming a mainstream tool in various application domains is that many limitations or potentials are discovered and quickly addressed in further developments of these methods.

In Paper IV, we thoroughly discuss the limitations of pix2pix and CycleGAN arising from the rotationally non-equivariant representations these GANs generate. This issue is also addressed for CycleGAN in [35], and a geometry-consistent generative adversarial network (Gc-GAN) is proposed. Gc-GAN uses a loss term additional to the adversarial loss to enforce the generating function to commute with vertical flipping and 90°clockwise rotations.

This loss term requires an additional hyperparameter to be tuned, unlike the equivariance constraint proposed in Paper III for contrastive learning.

In [105], InfoNCE is used in a very similar way as we proposed in Paper III but is paired with an adversarial loss to generate representations in an unpaired I2I fashion for the first time, resulting in a network called CUT, which demonstrated superior performance over cycle consistency-based frameworks. Similarly, [45] introduce Dual Contrastive Learning GAN (DCLGAN) to further exploit contrastive learning for I2I and avoid the restrictions resulting from the cycle consistency. Both [105] and [45] use contrastive learning in their generator only, which can result in mode collapse due to overfitting of the discriminator [41], an issue relieved by using another contrastive loss on the discriminator output layer in [41]. Also, work done in [20] builds upon [45, 105] and constrains the semantic consistency of multi-scale pairwise features between the encoder and decoder of the generator using contrastive learning.

There is a lot of potential and ongoing research to combine the power of contrastive learning and GANs to improve image translation tasks, which in turn can be used for multimodal representation learning for downstream tasks such as registration and retrieval. It would be interesting to harvest upon our observations in Papers III-V and explore the topic of equivariant contrastive learning in the setting of unpaired I2I, which has the potential to address one main limitation of current CoMIR generation: the need for aligned training images.

After their initial publication in Paper III, CoMIRs themselves have been evaluated on a variety of different datasets in [82] and [96]: remote sensing images (RGB & NIR), QPI & FM images, and magnetic resonance images (MRI T1 & T2). The work by Nordling et al. [93, 94] extends the equivariance of the representations to more general transformations – arbitrary rotations, affine and deformable transformations – and successfully employs the monomodal, deformable registration framework INSPIRE [97] demonstrating the CoMIRs’ usefulness for deformable registration.

Currently, ongoing work extends CoMIRs to 3D and will evaluate their applicability to medical images. Furthermore, the study aims to compare the representations originating from slice-wise and volume-based training.

Another aspect of CoMIRs that would be exciting to explore is their potential for unsupervised segmentation. Contrastive learning can be used for self-supervised pre-training [36] of segmentation networks. Similarly to downstream tasks like registration, segmentation requires 2D (or 3D) representations that consider spatial relationships between pixels (or voxels), i.e., these representations should be rotation equivariant, an aspect that is largely ignored in current networks. If we can show that CoMIRs can serve as reasonable segmentation maps, we can study their usefulness in other partitioning tasks, such as replacing K -means clustering in multimodal registration based on the cross-mutual information function (CMIF) presented in [96].

Sammanfattning på Svenska

Modern artificiell intelligens (AI) och maskininlärningsmetoder har under det senaste årtiondet stått för stora framsteg inom många discipliner, inte minst datorseende, där de visat sig väldigt användbara för hitta och känna igen objekt i bilder. Trots detta finns det en betydande lucka när det gäller att utnyttja metodernas fulla potential för biomedicinska tillämpningar. Den främsta orsaken är den begränsade tillgången till träningsdata, då biomedicinska tillämpningar ofta kräver visuell och manuell bedömning, så kallad annotering, av träningsdata från personer med expertkunskap, som t.ex. cytologer, patologer och annan medicinsk expertis. Detta kan jämföras med tillämpningar kopplade till självkörande bilar, där det är tillräckligt att den som skapar träningsdata har den kunskap som krävs för att bedöma en trafiksituation. Maskininlärningsmodeller lär sig funktioner som är relevanta för en uppgift genom att tränas med stora mängder relevant data, man kan säga att datadrivna maskininlärningsmetoder är extremt 'datahungliga'.

Trots denna begränsning finns en stor potential i att använda maskininlärningsmetoder inom biovetenskap och medicin. Ett sätt att övervinna problemet med begränsad tillgång till annoterade bilder och maskininlärnings datahunger är att använda ytterligare information om den tillgängliga datan för att 'guida' AI-metoderna och göra inlärningsprocessen lättare. Denna 'guidning' kan t.ex. bestå av att använda expertkunskap för att få maskininlärningsmetoden att fokusera på specifika egenskaper i bilderna som är relevanta för uppgiften, eller att utnyttja rumsliga relationer i bilderna och begränsa sökrymden för att förenkla problemet.

Den här avhandlingen syftar till att etablera metoder som kombinerar styrkan hos klassisk bildbehandling med kraften i maskininläring genom att utnyttja kunskap kring tillämpningen och dess sammanhang. De utvecklade metoderna appliceras här i första hand på bilder från digital patologi, men metoderna är användbara för ett mycket bredare spektrum av tillämpningar. Två av studierna i avhandlingen syftar främst till att mäta texturegenskaper i mikroskopibilder, antingen i form av manuellt utvalda funktioner som är modifierade så att måtten fungerar som indata till neurala faltningsnätverk, eller genom att bygga in specifika lager som lär sig textur, direkt i faltningsnätverket under nätverkets inlärningsfas. De övriga studierna utvecklar metoder för multimodala bilder, det vill säga bilder som tagits med olika typer av sensorer som kan bidra med kompletterande information om ett biologiskt prov. För

att kombinera bilder från olika sensorer behöver de placeras i samma koordinatsystem med hjälp av bildregistrering. I avhandlingen presenteras metoder för representationsinlärning, vilket betyder att man tränar ett nätverk att hitta gemensamma nämnare för bilder från olika sensorer, som är oberoende av bildens rotation, och gör det möjligt att effektivt registrera alla bilder.

Avhandlingen innehåller följande artiklar:

Artikeln I presenterar en metod för automatiserad bildinsamling avsedd för bedömning av njurens funktioner (nefropatologi) på nano-nivå, med hjälp av ett transmissionselektronmikroskop. Områden som är relevanta för att ställa diagnos identifieras med hjälp av a priori kunskap om deras textur.

I artikel II används a priori kunskap om textur för bildklassificering med målet att upptäcka munhålecancer utifrån celler i borstprover från insidan av kinden. Det övergripande syftet är att möjliggöra ett effektivt och robust AI-baserat cancerscreeningsprogram.

Artikeln III presenterar maskininlärning för att generera bildrepresentationer så att informationen från par av bilder av samma objekt avbildade med olika typer av sensorer (modaliteter) kan kombineras. I studien används representationerna för att registrera bildpar och kombinera deras information för att på så sätt kunna utvinna ytterligare information som inte är tillgänglig från de individuella bilderna var för sig. Registeringssteget är nödvändigt för vidare biomedicinsk tolkning, t.ex. medicinsk diagnos eller cancergradering.

Artikeln IV bygger vidare på inlärd representationer för multimodala bilder och presenterar en pipeline som kombinerar dem med kraftfulla bilddeskriptorer från klassisk bildbehandling för att erhålla en metod för multimodal bildmatchning, d.v.s. givet en bild tagen i en modalitet kan metoden hitta motsvarande bild i en annan. Denna multimodala bildhämtningsmetod kan användas som ett första steg i en registreringspipeline för att hitta matchande par av multimodala bilder som visar samma område i ett avbildat prov.

Artikeln V bygger på de teoretiska maskininlärningsaspekterna som används i artikel III, och syftar till att analysera och förbättra de multimodala bildrepresentationer som används där och i artikel IV för efterföljande analysuppgifter.

Bibliography

- [1] Cancer i munhålan ska upptäckas tidigare med ny metod. <https://www.svt.se/nyheter/lokalt/blekinge/munhalecancer-ska-upptackas-tidigt-med-ny-metod>. Accessed: 2023-03-26.
- [2] De vill kontrollera cancer – hos tandläkaren. <https://www.svt.se/nyheter/inrikes/expertter-infor-screening-for-munhalecancer>. Accessed: 2023-03-26.
- [3] Ny metod för att upptäcka muncancer testas i dalarna. <https://www.svt.se/nyheter/lokalt/dalarna/ny-metod-for-att-upptacka-muncancer-testas-i-dalarna>. Accessed: 2023-03-26.
- [4] Submissions of the Eleventh International Conference on Learning Representations (ICLR). <https://openreview.net/group?id=ICLR.cc/2023/Conference#all-submissions>. Accessed: 2023-03-26.
- [5] Nora Al-Garaawi, Raja Ebsim, Abbas F.H. Alharan, and Moi Hoon Yap. Diabetic foot ulcer classification using mapped binary patterns and convolutional neural networks. *Computers in Biology and Medicine*, 140:105055, 2022.
- [6] Justinas Antanavicius, Roberto Leiras, and Raghavendra Selvan. Identifying partial mouse brain microscopy images from the allen reference atlas using a contrastively learned semantic space. In *Biomedical Image Registration*, pages 166–176. Springer Intl. Publishing, 2022.
- [7] Rao Muhammad Anwer, Fahad Shahbaz Khan, Jorma Laaksonen, and Nazar Zaki. Multi-stream convolutional networks for indoor scene recognition. In *Computer Analysis of Images and Patterns*, pages 196–208, Cham, 2019. Springer International Publishing.
- [8] Rao Muhammad Anwer, Fahad Shahbaz Khan, Joost Van De Weijer, Matthieu Molinier, and Jorma Laaksonen. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J. Photogramm. Remote Sens.*, 138:74–85, 2018.
- [9] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32*, pages 15535–15545. Curran Associates, Inc., 2019.
- [10] Buda Bajić, Tomáš Majtner, Joakim Lindblad, and Nataša Sladoje. Generalised deep learning framework for HEP-2 cell recognition using local binary pattern maps. *IET Image Processing*, 14(6):1201–1208, 2020.
- [11] Laura Barisoni, Kyle J Lafata, Stephen M Hewitt, Anant Madabhushi, and Ulysses GJ Balis. Digital pathology and computational image analysis in nephropathology. *Nature Reviews Nephrology*, 16(11):669–685, 2020.

- [12] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Lecture notes in computer science*, 3951:404–417, 2006.
- [13] R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957.
- [14] Karl Bengtsson Bernander, Joakim Lindblad, Robin Strand, and Ingela Nyström. Replacing data augmentation with rotation-equivariant CNNs in image-based classification of oral cancer. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 25th Iberoamerican Congress, CIARP 2021, Porto, Portugal, May 10–13, 2021, Revised Selected Papers 25*, pages 24–33. Springer, 2021.
- [15] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.
- [16] Ingwer Borg and Patrick JF Groenen. *Modern Multidimensional Scaling – Theory and Applications*. Springer New York, 2005.
- [17] Ali Borji. Pros and Cons of GAN evaluation measures: New developments. *CoRR*, abs/2103.09396, 2021.
- [18] Hamid Reza Boveiri, Raouf Khayami, Reza Javidan, and Alireza Mehdizadeh. Medical image registration using deep neural networks: A comprehensive review. *Computers & Electrical Engineering*, 87:106767, 2020.
- [19] Paul D Brown. Transmission electron microscopy—a textbook for materials science, by David B. Williams and C. Barry Carter. *Microscopy and Microanalysis*, 5(6):452–453, 1999.
- [20] Xiuding Cai, Yaoyao Zhu, Dong Miao, Linjie Fu, and Yu Yao. Constraining multi-scale pairwise features between encoder and decoder using contrastive learning for unpaired image-to-image translation, 2022.
- [21] Juan C. Caicedo, Angel Cruz, and Fabio A. Gonzalez. Histopathology image classification using bag of features and kernel functions. In *Artificial Intelligence in Medicine*, pages 126–135, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
- [22] Paul Campagnola. Second harmonic generation imaging microscopy: applications to diseases diagnostics. *Analytical chemistry*, 83(9):3224–3231, 2011.
- [23] Davide Castelvocchi. Abstracts written by ChatGPT fool scientists. <https://www.nature.com/articles/d41586-023-00056-7>, 2023. Accessed: 2023-01-20.
- [24] Dongdong Chen, Mike Davies, Matthias J Ehrhardt, Carola-Bibiane Schönlieb, Ferdia Sherry, and Julián Tachella. Imaging with equivariant deep learning: From unrolled network design to fully unsupervised learning. *IEEE Signal Processing Magazine*, 40(1):134–147, 2023.
- [25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [26] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8185–8194, 2020.

- [27] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proc. Intl. Conf. on Machine Learning*, volume 48 of *Proc. of Machine Learning Research*, pages 2990–2999. PMLR, 2016.
- [28] Taco Cohen and Max Welling. Steerable CNNs. In *5th Intl. Conf. on Learning Representations, ICLR 2017*. OpenReview.net, 2017.
- [29] Alberto Diaspro. *Optical fluorescence microscopy: From the spectral to the nano dimension*. Springer Science & Business Media, 2010.
- [30] Albino Eccher and Ilaria Girolami. Current state of whole slide imaging use in cytopathology: Pros and pitfalls. *Cytopathology*, 31(5):372–378, 2020.
- [31] Kevin Eliceiri, Bin Li, and Adib Keikhosravi. Multimodal Biomedical Dataset for Evaluating Registration Methods (patches from TMA Cores). Zenodo, June 2020.
- [32] Kevin Eliceiri, Bin Li, and Adib Keikhosravi. Multimodal biomedical dataset for evaluating registration methods (full-size TMA cores). Zenodo, Feb 2021.
- [33] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.
- [34] Pasquale Foggia, Gennaro Percannella, Paolo Soda, and Mario Vento. Benchmarking hep-2 cells classification methods. *IEEE transactions on medical imaging*, 32(10):1878–1889, 2013.
- [35] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [36] Sebastian Gerard and Josephine Sullivan. Contrastive pretraining for semantic segmentation is robust to noisy positive pairs, 2023.
- [37] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [38] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., USA, 2006.
- [39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [40] A. Ardeshir Goshtasby. *Image Registration; Principles, Tools and Methods*. Advances in Computer Vision and Pattern Recognition. Springer-Verlag London, 1 edition, 2012.
- [41] Yao Gou, Min Li, Yu Song, Yujie He, and Litao Wang. Multi-feature contrastive learning for unpaired image-to-image translation. *Complex & Intelligent Systems*, pages 1–12, 2022.
- [42] Simon Graham, David Epstein, and Nasir Rajpoot. Dense steerable filter CNNs for exploiting rotational symmetry in histology images. *IEEE Transactions on Medical Imaging*, 39(12):4124–4136, 2020.
- [43] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the*

- thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.
- [44] Neda Haj-Hosseini, Joakim Lindblad, Bengt Hasséus, Vinay Vijaya Kumar, Narayana Subramaniam, and Jan-Michaél Hirsch. Early detection of oral potentially malignant disorders: A review on prospective screening methods with regard to global challenges. *Journal of Maxillofacial and Oral Surgery*, pages 1–10, 2022.
- [45] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 746–755, 2021.
- [46] Robert M. Haralick, K. Shanmugam, and Its’Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.
- [47] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [48] Dong-Chen He and Li Wang. Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):509–512, 1990.
- [49] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [51] Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. Similar image search for histopathology: SMILY. *NPJ digital medicine*, 2(1):56, 2019.
- [52] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *CoRR*, abs/1905.09272, 2019.
- [53] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [54] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *Intl. Conf. on Learning Representations*, 2019.
- [55] Peter Hobson, Brian C Lovell, Gennaro Percannella, Mario Vento, and Arnold Wiliem. Benchmarking human epithelial type 2 interphase cells classification methods on a very large dataset. *Artificial intelligence in medicine*, 65(3):239–250, 2015.
- [56] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros.

- Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [57] James Jabalee, Anita Carraro, Tony Ng, Eitan Prisman, Cathie Garnis, and Martial Guillaud. Identification of malignancy-associated changes in histologically normal tumor-adjacent epithelium of patients with HPV-positive oropharyngeal cancer. *Anal Cell Pathol (Amst)*, 2018.
- [58] Stephan W Jahn, Markus Plass, and Farid Moinfar. Digital pathology: advantages, limitations and emerging perspectives. *Journal of Clinical Medicine*, 9(11):3697, 2020.
- [59] Simon Jégou, Michal Drozdal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 11–19, 2017.
- [60] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local binary convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19–28, 2017.
- [61] Aakash Kaku, Sahana Upadhyaya, and Narges Razavian. Intermediate layers matter in momentum contrastive self supervised learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 24063–24074. Curran Associates, Inc., 2021.
- [62] Rajiv Kapoor, Deepak Sharma, and Tarun Gulati. State of the art content based image retrieval techniques using deep learning: a survey. *Multimed Tools Appl*, 80:29561–29583, 2021.
- [63] Adib Keikhosravi, Jeremy S. Bredfeldt, Abdul Kader Sagar, and Kevin W. Eliceiri. Chapter 28 - second-harmonic generation imaging of cancer. In Jennifer C. Waters and Torsten Wittman, editors, *Quantitative Imaging in Cell Biology*, volume 123 of *Methods in Cell Biology*, pages 531–546. Academic Press, 2014.
- [64] Adib Keikhosravi, Bin Li, Yuming Liu, and Kevin W. Eliceiri. Intensity-based registration of bright-field and second-harmonic generation images of histopathology tissue sections. *Biomed. Opt. Express*, 11(1):160–173, Jan 2020.
- [65] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Roberto Togneri, and Ferdous A Sohel. A discriminative representation of convolutional features for indoor scene recognition. *IEEE Transactions on Image Processing*, 25(7):3372–3383, 2016.
- [66] Daisuke Komura, Keisuke Fukuta, Ken Tominaga, Akihiro Kawabe, Hirotomo Koda, Ryohei Suzuki, Hiroki Konishi, Toshikazu Umezaki, Tatsuya Harada, and Shumpei Ishikawa. Luigi: Large-scale histopathological image retrieval system using deep texture representations. *bioRxiv*, 2018.
- [67] Nadezhda Koriakina, Nataša Sladoje, Vladimir Bašić, and Joakim Lindblad. Oral cancer detection and interpretation: Deep multiple instance learning versus conventional deep single instance learning, 2022.
- [68] Joseph B Kruskal and Myron Wish. Multidimensional scaling. *Quantitative Applications in the Social Sciences*, 11:234–778, 1978.
- [69] Gustaf Kylberg, Mats Uppström, and Ida-Maria Sintorn. Virus texture analysis using local binary patterns and radial density profiles. In *Progress in Pattern*

Recognition, Image Analysis, Computer Vision, and Applications: 16th Iberoamerican Congress, CIARP 2011, Pucón, Chile, November 15-18, 2011. Proceedings 16, pages 573–580. Springer, 2011.

- [70] Maxime W. Lafarge, Erik J. Bekkers, Josien P.W. Pluim, Remco Duits, and Mitko Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis*, 68:101849, 2021.
- [71] Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147, 1998.
- [72] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.
- [73] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018.
- [74] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: Diverse image-to-image translation via disentangled representations. *Int J Comput Vision*, 128:2402–2417, 2020.
- [75] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proc. of ACM Int. Conf. on Multimodal Interaction*, pages 503–510. ACM, 2015.
- [76] Lei Li, Xiaoyi Feng, Zhaoqiang Xia, Xiaoyue Jiang, and Abdenour Hadid. Face spoofing detection with local binary pattern network. *J. Visual Commun. Image Represent.*, 54:182–192, 2018.
- [77] Wenyi Lin, Kyle Hasenstab, Guilherme Moura Cunha, and Armin Schwartzman. Comparison of handcrafted features and convolutional neural networks for liver mr image adequacy assessment. *Scientific Reports*, 10(1):1–11, 2020.
- [78] Joakim Lindblad and Nataša Sladoje. Linear time distances between fuzzy sets with applications to pattern matching and classification. *IEEE Transactions on Image Processing*, 23(1):126–136, 2014.
- [79] Li Liu, Paul Fieguth, Yulan Guo, Xiaogang Wang, and Matti Pietikäinen. Local binary features for texture classification: Taxonomy and experimental study. *Pattern Recognition*, 62:135–160, 2017.
- [80] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou. An improved evaluation framework for generative adversarial networks, 2018.
- [81] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.
- [82] Jiahao Lu, Johan Öfverstedt, Joakim Lindblad, and Nataša Sladoje. Is image-to-image translation the panacea for multimodal image registration? a comparative study. *PLOS ONE*, 17(11):1–33, 11 2022.
- [83] Jiahao Lu, Nataša Sladoje, Christina Runow Stark, Eva Darai Ramqvist, Jan-Michaél Hirsch, and Joakim Lindblad. A deep learning based pipeline for efficient oral cancer screening on whole slide images. In *Image Analysis and*

- Recognition*, pages 249–261. Springer International Publishing, 2020.
- [84] Jiahao Lu, Johan Öfverstedt, Joakim Lindblad, and Nataša Sladoje. Datasets for Evaluation of Multimodal Image Registration. Zenodo, April 2021.
- [85] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129:23–79, 2021.
- [86] Tomáš Majtner, Buda Bajić, and Jürgen Herp. Texture-based image transformations for improved deep learning classification. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 25th Iberoamerican Congress, CIARP 2021, Porto, Portugal, May 10–13, 2021, Revised Selected Papers 25*, pages 207–216. Springer, 2021.
- [87] P Mallikarjuna, Alireza Tavakoli Targhi, Mario Fritz, Eric Hayman, Barbara Caputo, and Jan-Olof Eklundh. The KTH-TIPS2 database. *Computational Vision and Active Perception Laboratory, Stockholm, Sweden*, 11:12, 2006.
- [88] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057, 2017.
- [89] Vivien Marx. It’s free imaging—label-free, that is. *Nature methods*, 16(12):1209–1212, 2019.
- [90] David Mattes, David R. Haynor, Hubert Vesselle, Thomas K. Lewellyn, and William Eubank. Nonrigid multimodality image registration. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 1609 – 1620. International Society for Optics and Photonics, SPIE, 2001.
- [91] Stanislav Morozov, Andrey Voynov, and Artem Babenko. On self-supervised image representations for GAN evaluation. In *Intl. Conf. on Learning Representations*, 2021.
- [92] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *Intl. Journal of Medical Informatics*, 73(1):1 – 23, 2004.
- [93] Love Nordling. Contrastive multimodal image representations for deformable image registration, 2022.
- [94] Love Nordling, Johan Öfverstedt, Joakim Lindblad, and Nataša Sladoje. Contrastive learning of equivariant image representations for multimodal deformable registration. International Symposium on Biomedical Imaging, April 2023.
- [95] Johan Öfverstedt, Joakim Lindblad, and Nataša Sladoje. Fast and robust symmetric image registration based on distances combining intensity and spatial information. *IEEE Transactions on Image Processing*, 28(7):3584–3597, 2019.
- [96] Johan Öfverstedt, Joakim Lindblad, and Nataša Sladoje. Fast computation of mutual information in the frequency domain with applications to global multimodal image alignment. *Pattern Recognition Letters*, 159:196–203, 2022.
- [97] Johan Öfverstedt, Joakim Lindblad, and Nataša Sladoje. INSPIRE: Intensity and spatial information-based deformable image registration. *Plos one*, 18(3):e0282432, 2023.
- [98] Catherine M Oikonomou and Grant J Jensen. The atlas of bacterial & archaeal

- cell structure: an interactive open-access microbiology textbook. *Journal of Microbiology & Biology Education*, 22(2):e00128–21, 2021.
- [99] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.
- [100] Philip Oltermann. European politicians duped into deepfake video calls with mayor of kyiv. <https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko>, 2022. Accessed: 2023-01-20.
- [101] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [102] Sebastian Otálora, Roger Schaer, Oscar Jimenez-del Toro, Manfredo Atzori, and Henning Müller. Deep learning based retrieval system for gigapixel histopathology cases and open access literature. *bioRxiv*, 2018.
- [103] Zhibin Pan, Shiqi Hu, Xiuquan Wu, and Ping Wang. Adaptive center pixel selection strategy in local binary pattern for texture classification. *Expert Systems with Applications*, 180:115123, 2021.
- [104] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *CoRR*, abs/2101.08629, 2021.
- [105] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 319–345, Cham, 2020. Springer International Publishing.
- [106] YongKeun Park, Christian Depeursinge, and Gabriel Popescu. Quantitative phase imaging in biomedicine. *Nature photonics*, 12(10):578–589, 2018.
- [107] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
- [108] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. *Computer vision using local binary patterns*, volume 40. Springer Science & Business Media, 2011.
- [109] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, 2003.
- [110] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR, 09–15 Jun 2019.
- [111] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009.
- [112] Torsten Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE transactions on medical imaging*, 31(2):153–163, 2011.
- [113] David W. Romero and Suhas Lohit. Learning partial equivariances from data.

- In *Advances in Neural Information Processing Systems*, 2022.
- [114] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [115] John W. Sammon. A nonlinear mapping for data structure analysis. *Transactions on Computers*, C-18(5):401–409, 5 1969.
- [116] EKW Schulte. Standardization of biological dyes and stains: pitfalls and possibilities. *Histochemistry*, 95:319–328, 1991.
- [117] Christopher G Schwarz, Walter K Kremers, Terry M Therneau, Richard R Sharp, Jeffrey L Gunter, Prashanthi Vemuri, Arvin Arani, Anthony J Spychalla, Kejal Kantarci, David S Knopman, et al. Identification of anonymous MRI research participants with face-recognition software. *New England Journal of Medicine*, 381(17):1684–1686, 2019.
- [118] Y.R. Shen. Nonlinear optical susceptibilities. In *Encyclopedia of Materials: Science and Technology*, pages 6249–6255. Elsevier, Oxford, 2001.
- [119] Olivier Simon, Rabi Yacoub, Sanjay Jain, John E Tomaszewski, and Pinaki Sarder. Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. *Sci. Rep.*, 8(1):2032, 2018.
- [120] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [121] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, 2009.
- [122] Roman Stoklasa, Tomáš Majtner, and David Svoboda. Efficient k-NN based HEp-2 cells classifier. *Pattern Recognition*, 47(7):2409–2418, 2014.
- [123] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *CoRR*, abs/1906.05743, 2019.
- [124] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [125] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- [126] TinEye. <https://tineye.com/faq#what>, 2008. Accessed: 2021-12-13.
- [127] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018 Proceedings, Part II 11*, pages 210–218. Springer, 2018.
- [128] Tomas Vicar, Martina Raudenska, Jaromir Gumulec, Michal Masarik, and Jan Balvan. Quantitative phase microscopy timelapse dataset of PNT1A, DU-145 and LNCaP cells with annotated caspase 3,7-dependent and independent cell death. Zenodo, March 2019.
- [129] Tomas Vicar, Martina Raudenska, Jaromir Gumulec, Michal Masarik, and Jan Balvan. Fluorescence microscopy timelapse dataset of PNT1A, DU-145 and

- LNCaP cells with annotated caspase 3,7-dependent and independent cell death. Zenodo, February 2021.
- [130] Max A Viergever, JB Antoine Maintz, Stefan Klein, Keelin Murphy, Marius Staring, and Josien PW Pluim. A survey of medical image registration—under review, 2016.
- [131] Michele Volpi and Vittorio Ferrari. Semantic segmentation of urban scenes by learning local class interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.
- [132] Gufeng Wang and Ning Fang. Chapter four - detecting and tracking nonfluorescent nanoparticle probes in live cells. In *Imaging and Spectroscopic Analysis of Living Cells*, volume 504 of *Methods in Enzymology*, pages 83–108. Academic Press, 2012.
- [133] Li Wang and Dong-Chen He. Texture classification using texture spectrum. *Pattern recognition*, 23(8):905–910, 1990.
- [134] Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23078–23091. PMLR, 17–23 Jul 2022.
- [135] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [136] Jennifer Waters and Torsten Wittmann. *Quantitative imaging in cell biology*. Academic Press, 2014.
- [137] Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [138] Maurice Weiler, Fred A. Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant CNNs. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [139] Håkan Wieslander, Gustav Forslid, Ewert Bengtsson, Carolina Wählby, Jan-Michael Hirsch, Christina Runow Stark, and Sajith Kecheril Sadanandan. Deep convolutional neural networks for detecting cellular changes due to malignancy. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 82–89, 2017.
- [140] Arnold Wiliem, Yongkang Wong, Conrad Sanderson, Peter Hobson, Shaokang Chen, and Brian C Lovell. Classification of human epithelial type 2 cell indirect immunofluorescence images via codebook based descriptors. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 95–102. IEEE, 2013.
- [141] Hanwei Wu, Ather Gattami, and Markus Flierl. Conditional mutual information-based contrastive loss for financial time series forecasting. *arXiv preprint arXiv:2002.07638*, 2020.
- [142] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.
- [143] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio

- Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.
- [144] K Yogesan, T Jørgensen, F Albrechtsen, KJ Tveter, and HE Danielsen. Entropy-based texture analysis of chromatin structure in advanced prostate cancer. *Cytometry*, 24(3):268–276, 1996.
- [145] Xiaoping Zhou, Xiangyu Han, Haoran Li, Jia Wang, and Xun Liang. Cross-domain image retrieval: methods and applications. *J Multimed Info Retr*, 11:199–218, 2022.
- [146] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017.
- [147] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2266*

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-500386



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023