

# Deep Learning Techniques for Multi-Dimensional Medical Image Analysis

***Citation for published version (APA):***

Ghazvinian Zanjani, F. (2023). *Deep Learning Techniques for Multi-Dimensional Medical Image Analysis*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Electrical Engineering]. Eindhoven University of Technology.

***Document status and date:***

Published: 27/09/2023

***Document Version:***

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

***Please check the document version of this publication:***

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

***General rights***

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

***Take down policy***

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

# Deep Learning Techniques for Multi-Dimensional Medical Image Analysis

Farhad Ghazvinian Zanjani



# Deep Learning Techniques for Multi-Dimensional Medical Image Analysis

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de  
Technische Universiteit Eindhoven, op gezag van de  
rector magnificus, prof.dr. S.K. Lenaerts, voor een  
commissie aangewezen door het College voor Promoties,  
in het openbaar te verdedigen  
op woensdag 27 september 2023 om 13.30 uur

door

Farhad Ghazvinian Zanjani

geboren te Teheran, Iran



Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	Prof.dr.ir. S.M Heemstra-de Groot.
promotor:	prof.dr.ir. P.H.N. de With
copromotor:	dr. S. Zinger
leden:	prof.dr.ir. A. Smeulders (University of Amsterdam)
	prof.dr. J. Pluim
	prof.dr. P. Schelkens (Vrije Universiteit Brussel)
	prof.dr.ir. M. Misch
	prof.dr. J. van der Laak (Radboud Universiteit of Nijmegen)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

*To my parents, Marjan and Ali*

---

Deep Learning Techniques for Multi-Dimensional Medical Image Analysis

Farhad Ghazvinian Zanjani

Front cover and chapter title photos: Text-to-Image Generative AI

Cover design: Farhad Ghazvinian Zanjani

Printed by: Gildeprint B.V.

A catalogue record is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-5817-9

NUR-code: 984

---

Copyright © 2023 by Farhad Ghazvinian Zanjani

All Rights Reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic, mechanical, including photocopying, recording or by any information storage and retrieval system, without the prior permission of the copyright owner.

# Summary

## Deep Learning Techniques for Multi-Dimensional Medical Image Analysis

Medical imaging plays a vital role in various aspects of healthcare, such as screening, diagnostics, prognosis, treatment, and monitoring therapies. As images become larger and more complex, automated image analysis becomes increasingly important in aiding healthcare professionals. Furthermore, the integration of computerized imaging systems for data processing and assisting care providers presents innovative opportunities to enhance healthcare by providing richer information. This thesis aims to advance this rapidly expanding field by creating novel computational models for medical image analysis. These models are designed to assist clinicians in making informed decisions, performing image-guided interventions, and developing treatment plans, ultimately striving for safer and improved patient care.

This thesis introduces a range of advanced computational models, designed for image analysis across four distinct medical imaging domains. By leveraging recent advancements in deep learning models, these proposed computational models should address crucial data processing challenges specific to each domain with the overall purpose to enhance patient care. The research conducted in this thesis focuses on four imaging systems: 1D mass spectrometry imaging, 2D pathology bright-field microscopy, 3D volumetric ultrasound imaging, and 3D point cloud scanning. The contributing chapters of this dissertation are structured to investigate pertinent clinical issues, associated with each imaging system and present deep learning-based solutions to alleviate these challenges.

Chapter 2 serves as the initial technical case study within this thesis, focusing on the exploration of automated cancer detection and classification, using a novel emerging technology known as mass spectrometry imaging (MSI). The MSI data provides high-dimensional molecular profiles of cells at the surfaces of the tissue. The interpretation of the obtained mass spectrum information is a challenging task for the physician, since it involves an enormous amount of raw data that corresponds with a disease pattern. Due to the inherent complexity and high-dimensionality nature of data, two deep learning models are developed. It is shown that both a 1D convolutional network utilizing dilated convolution kernels, and a recurrent neural network can learn the local and global dependencies

across mass spectra to perform reliable cancer classification of MSI data. The proposed models can attain state-of-the-art performance, achieving an 86.5%  $F_1$  score for lung cancer diagnosis and an 84.0%  $F_1$  score for bladder cancer diagnosis in clinical data.

Chapter 3 studies histopathology imaging, which is widely employed in clinical practice for cancer diagnosis. The development of a precise, efficient, and generalizable CAD system for histopathology image analysis necessitates the exploration of various factors, including: (1) Addressing the inconsistencies in histology slide preparation across different histopathology laboratories, as well as the resulting color variations in image acquisition from slides; (2) Developing learning-based models to process ultra-large whole-slide images to perform cancer detection, segmentation, and grading capabilities at human-level performance; (3) The possible incorporation of compression techniques for storage and transmission of large-scale histopathology data, while assessing the impact of the most commonly used compression method on the performance of learning algorithms.

For the first aspect, two distinct unsupervised deep learning approaches are introduced that normalize the stain color of histopathology images obtained from different labs, thereby ensuring color consistency and reducing domain shift when the training data includes multiple datasets. For the second aspect, the patch-level representations of whole-slide images (WSIs) are processed jointly to perform inference on broader context than what is presented in each individual patch. To model the dependencies across patches, the learned features from individual patches are used to form a fully-connected graph with message-passing mechanism to propagate the information across nodes (i.e. patches). Empirical evaluations show that incorporating a Conditional Random Fields (CRFs) model for the dependencies between all patches inside a WSI and imposing a joint labeling strategy improves the tumor region detection. This can be observed by achieving a high kappa score equal to 0.876 in patient-level metastasis classification on breast lymph data. Regarding the compression, the influence of the JPEG 2000 compression technique is evaluated up to a compression ratio of 48:1. The results indicate that the impact is negligibly small for compression factor lower than 32, thereby allowing reliable automated detection even after compression up to this factor.

Chapter 4 focuses on deep learning solutions for robust needle detection and visualization in 3D ultrasound (US) image-guided interventions. US imaging is widely employed to guide minimally invasive interventions involving the percutaneous advancement of a needle or catheter to a specific target within the patient's body. These interventions are highly prone to failure and complications due to the challenging requirement of precise hand-eye coordination between the instrument and the US transducer. Continuous visibility of the needle and its tip throughout the intervention is crucial but challenging to achieve in practice. The continuous visibility is only attained when the transducer is manipulated to search for the optimal capturing position in the volumetric space, allowing for the identification of the most suitable needle-viewing plane during the inter-

vention. This chapter proposes two deep learning-based CAD algorithms that automatically detect and localize the needle in the 3D US volume without requiring hand-eye coordination from the physician. The localization performance in Chicken and Porcine data exceeds 80% precision with nearly 90% recall. With this automated support, the physician can give full attention to the needle and target tissue, maintaining intervention quality and patient comfort.

Chapter 5 focuses on a fundamental challenge within computational dentistry, namely, the development of a CAD system capable of accurately performing instance segmentation of teeth from imaging data. In the fields of implantology and orthodontics, an automated clinical workflow necessitates precise segmentation of teeth and gums in intra-oral scans, obtained from advanced optical measurement devices, known as intra-oral scanners (IOS). These scanners capture 3D surface profiles of anatomical structures within the oral cavity. Notably, this research represents the first investigation into teeth instance segmentation within 3D point clouds, derived from IOS data. The study introduces a novel deep learning-based solution, named Mask-MCNet, which processes the IOS data in its original resolution. By using Monte Carlo convolution operators, the model is capable of transferring the point representations to a regular 3D grid without using a quantization step. The model predicts instance labels for each individual point in the IOS scan and exhibits a remarkable accuracy in segmentation, achieving 98% mIOU (mean Intersection over Union) and mAP (mean Average Precision) scores when evaluated on clinical IOS data.

The thesis concludes with the overview of the posed research questions and answers, enhanced by the contributions of each chapter.

Upon reflecting on the four examined imaging problems, it becomes evident that their characteristics and the numbers of involved imaging dimensions vary significantly across different research directions. Despite this considerable variability, a promising observation is that leveraging a flexible neural architecture like Convolutional Neural Networks (ConvNets), capable of accommodating diverse input sizes and modalities, can yield high performance levels of detection and segmentation with an accuracy and quality that potentially assist clinicians and physicians. However, the substantial disparities in medical image analysis problems and the presence of task-specific specialized imaging systems currently impede the development of a generalized deep learning solution. However, in scenarios where multi-modal imaging is feasible, learning solutions can be integrated to fulfill complementary roles. An instance of such multi-modal analysis could involve joint processing of Mass Spectrometry Imaging (MSI) and histopathology images, or the fusion of Intra-oral 3D point cloud scans with dental Cone-beam Computed Tomography (CBCT) for enhanced analysis.



# Samenvatting

## Diepgaande leertechnieken voor multidimensionale medische beeldanalyse

Medische beeldvorming speelt een essentiële rol in verschillende aspecten van de gezondheidszorg, zoals screening, diagnostiek, prognose, behandeling en het monitoren van therapieën. Naarmate de beelden groter en complexer worden, groeit het belang van geautomatiseerde beeldanalyse om zorgprofessionals te ondersteunen. Bovendien biedt de integratie van geautomatiseerde beeldvormingssystemen voor gegevensverwerking en het assisteren van zorgverleners innovatieve mogelijkheden om de gezondheidszorg te verbeteren door het verstrekken van betere informatie. Dit proefschrift heeft als doel dit snelgroeiende vakgebied verder te ontwikkelen door nieuwe berekeningsmodellen te creëren voor de analyse van medische beelden. Deze nieuwe modellen zijn ontworpen om klinici te ondersteunen bij het nemen van geïnformeerde beslissingen, het uitvoeren van beeldgestuurde ingrepen en het ontwikkelen van behandelplannen, met als uiteindelijk doel een veiligere en verbeterde patiëntenzorg.

Dit proefschrift introduceert een reeks geavanceerde berekeningsmodellen die zijn ontworpen voor beeldanalyse in vier verschillende medische beeldvormingsdomeinen. Door gebruik te maken van recente ontwikkelingen in diepe leermethoden, zouden deze voorgestelde rekenmodellen fundamentele uitdagingen op het gebied van gegevensverwerking moeten oplossen of verbeteren, die specifiek zijn voor elk domein en met als algemeen doel de patiëntenzorg te verbeteren. Het onderzoek dat in dit proefschrift is uitgevoerd, richt zich op vier beeldvormingssystemen: 1D-beeldvorming met massaspectrometrie, 2D-microscopie voor pathologie, 3D-volumetrische echografische beelden en 3D-scanning met een puntenwolk. De bijdragen in de hoofdstukken van dit proefschrift zijn gestructureerd om relevante klinische vraagstukken te onderzoeken die verband houden met elk beeldvormingssysteem en om oplossingen op basis van diep lerende algoritmen te presenteren.

Hoofdstuk 2 dient als de initiële technische studie in dit proefschrift en richt zich op de verkenning van geautomatiseerde detectie en classificatie van kanker, met behulp van een nieuwe opkomende technologie genaamd massaspectrometrie-beeldvorming (MSI). De MSI-gegevens bieden hoog-dimensionale moleculaire profielen van cellen op het oppervlak van het



weefsel. Het interpreteren van de informatie in de verkregen massa-spectra is een uitdagende taak voor de arts, omdat het grote hoeveelheden ruwe gegevens omvat, die overeenkomen met een ziektepatroon. Vanwege de inherente complexiteit en de hoge dimensionaliteit van de gegevens worden twee diepe leermodellen ontwikkeld. Het wordt aangetoond dat zowel een 1D-convolutioneel netwerk met gedilateerde convolutiekernen als een recurrent neurale netwerk de lokale en globale afhankelijkheden tussen massaspectra kunnen leren om betrouwbare kankerclassificatie van MSI-gegevens uit te voeren. De voorgestelde modellen kunnen state-of-the-art prestaties behalen, met een  $F_1$  score van 86,5% voor de diagnose van longkanker en een  $F_1$  score van 84,0% voor de diagnose van blaaskanker in klinische gegevens.

Hoofdstuk 3 onderzoekt histopathologische beeldvorming, dat veel wordt gebruikt in de klinische praktijk voor de diagnose van kanker. De ontwikkeling van een nauwkeurig, efficiënt en generaliseerbaar CAD-systeem voor de analyse van dergelijke beelden vereist het onderzoeken van verschillende factoren, met name: (1) aanpakken van de inconsistenties in de voorbereiding van histologische preparaten in verschillende histopathologische laboratoria, evenals de kleurvariaties die resulteren uit het maken van beelden van preparaten; (2) ontwikkelen van op *deep learning* gebaseerde modellen om ultra-grote *whole-slide*-beelden te verwerken, die kankerdetectie, segmentatie en gradatie te kunnen uitvoeren met prestaties vergelijkbaar met die van een mens; (3) mogelijke toepassing van compressietechnieken voor de opslag en verzending van grootschalige histopathologiedata, terwijl de impact van de meest gebruikte compressiemethode op de prestaties van lerende algoritmen wordt beoordeeld.

Voor het eerste aspect worden twee verschillende, niet gesuperviseerde dieperende benaderingen geïntroduceerd die de kleur van histopathologiebeelden normaliseren die afkomstig zijn van verschillende laboratoria. Door deze aanpak wordt kleurconsistentie gegarandeerd en domeinverschuiving verminderd, wanneer de trainingsgegevens meerdere datasets bevatten. Voor het tweede aspect worden de op *patch*-niveau (deelbeeld) gemaakte representaties van volledige beeldplaten gezamenlijk verwerkt om inferentie uit te voeren op een bredere context dan wat er in elke individuele *patch* wordt gepresenteerd. Om de afhankelijkheden tussen patches te modelleren, worden de geleerde kenmerken van individuele patches gebruikt om een volledig verbonden netwerk te maken met een mechanisme om informatie over knooppunten (d.w.z. *patches*) te verspreiden. Uit empirische evaluaties blijkt dat het opnemen van een *Conditional Random Fields* (CRFs) model dat de afhankelijkheden tussen alle patches binnen een volledige beeldplaat beschrijft en het opleggen van een gezamenlijke labelstrategie, de detectie van tumorgebieden duidelijk verbetert. Dit is te zien aan een hoge kappascore van 0,876 bij patiëntgebaseerde classificatie van metastasen met borstlymfekliergegevens. Wat betreft compressie, wordt de invloed van de JPEG 2000 compressietechniek geëvalueerd tot aan een compressieverhouding van 48:1. De resultaten geven aan dat de impact verwaarloosbaar klein is voor compressiefactoren lager dan 32, waardoor betrouwbare geautomatiseerde detectie mogelijk blijft, zelfs na compressie van de beelden tot deze factor.

Hoofdstuk 4 richt zich op diepgaande leermethoden voor robuuste naald-detectie en visualisatie bij 3D-echografische (*ultrasound* of US) beeldgestuurde ingrepen. US-beeldvorming wordt veelvuldig gebruikt om minimaal invasieve ingrepen te begeleiden waarbij een naald of katheter percutaan naar een specifiek doel binnen het lichaam van de patiënt wordt geleid. Deze ingrepen zijn gevoelig voor complicaties en mislukkingen vanwege de uitdagende eis van precieze hand-oogcoördinatie tussen het instrument en de US-transducer. Continue zichtbaarheid van de naald en de punt ervan gedurende de ingreep is cruciaal maar moeilijk te realiseren in de praktijk. Continue zichtbaarheid wordt alleen bereikt wanneer de omvormer (*transducer*) manueel wordt gemanipuleerd om de optimale positionering in de volumetrische ruimte te zoeken, waardoor het meest geschikte naaldzichtvlak tijdens de ingreep kan worden geïdentificeerd. Dit hoofdstuk stelt twee CAD-algoritmen voor op basis van diepgaand leren, die automatisch de naald detecteren en lokaliseren in de 3D-US-volumedata zonder dat hand-oogcoördinatie van de arts vereist is. De lokaliseringsprestaties bij de Chicken en Porcine data overschrijden 80% precisie met bijna 90% recall. Met deze geautomatiseerde ondersteuning kan de arts zich volledig richten op de naald en het doelweefsel, wat de kwaliteit van de ingreep en het comfort van de patiënt ten goede komt.

Hoofdstuk 5 richt zich op een fundamentele uitdaging binnen de numerieke tandheelkunde, namelijk de ontwikkeling van een CAD-systeem dat in staat is om nauwkeurig de instantiesegmentatie van tanden uit beeldgegevens uit te voeren. De vakgebieden implantologie en orthodontie vereisen een geautomatiseerde klinische workflow een nauwkeurige segmentatie van tanden en tandvlees in intra-orale scans, verkregen met geavanceerde optische meetapparaten, bekend als intra-orale scanners (IOS). Deze scanners leggen 3D-oppervlakteprofielen vast van anatomische structuren in de mondholte. Opmerkelijk genoeg vertegenwoordigt dit onderzoek de eerste studie naar instantiesegmentatie van tanden binnen 3D-puntenwolken afgeleid van IOS-gegevens. De studie introduceert een nieuw, op *deep learning* gebaseerde oplossing, genaamd Mask-MCNet, die de IOS-gegevens verwerkt in hun oorspronkelijke resolutie. Met behulp van Monte Carlo-convolutieoperators kan het model de puntrepresentatiedata converteren naar een regelmatig 3D-rooster zonder gebruik te maken van een kwantisatiestap. Het model voorspelt instantielabels voor elke individueel punt in de IOS-scan en vertoont opmerkelijke nauwkeurigheid in segmentatie, met een mIOU (*mean Intersection over Union*) score van 98% en mAP (*mean Average Precision*) score bij evaluatie op klinische IOS-gegevens.

De dissertatie wordt afgesloten met een overzicht van de gestelde onderzoeksvragen en antwoorden, versterkt door de bijdragen van elk hoofdstuk.

Bij het reflecteren op de vier onderzochte beeldvormingsproblemen wordt het duidelijk dat hun kenmerken en het aantal betrokken beeldvormingsdimensies aanzienlijk variëren tussen verschillende onderzoeksrichtingen. Ondanks deze aanzienlijke variabiliteit is een veelbelovende constatering dat het benutten van

een flexibele neurale architectuur zoals gebruikt bij *Convolutional Neural Networks* (ConvNets), in staat is om diverse invoergroottes en modaliteiten te accommoderen, hetgeen kan leiden tot hoge prestatieniveaus bij detectie en segmentatie met een nauwkeurigheid en kwaliteit die mogelijk klinici en artsen kunnen ondersteunen. Echter, de aanzienlijke verschillen in problemen bij de analyse van medische beelden en de aanwezigheid van taakspecifieke gespecialiseerde beeldvormingssystemen belemmeren momenteel de ontwikkeling van een gegeneraliseerde *deep learning*-oplossing. Desalniettemin kunnen oplossingen bij scenario's waarin multimodale beeldvorming mogelijk is, geïntegreerd worden om complementaire rollen te vervullen. Een voorbeeld van een dergelijke multimodale analyse zou het gezamenlijk verwerken van *Mass Spectrometry Imaging* (MSI) data en histopathologiebeelden kunnen omvatten, of de fusie van intrasorale 3D-puntenwolksens met dentale *Cone-beam Computed Tomography* (CBCT) voor verbeterde analyse.

# Contents

<i>Summary</i>	<i>i</i>
<i>Samenvatting</i>	<i>v</i>
<b>1 Introduction</b>	<b>1</b>
1.1 Role of medical imaging in modern medicine . . . . .	1
1.2 Computer-aided diagnosis in medical imaging . . . . .	2
1.2.1 Data dimensionality in imaging . . . . .	3
1.3 Recent advancement in medical image analysis . . . . .	4
1.3.1 CADe and CADx systems . . . . .	5
1.4 Research scope and challenges . . . . .	6
1.4.1 Mass Spectrometry Imaging (MSI) . . . . .	6
1.4.2 Computational histopathology . . . . .	7
1.4.3 Instrument localization in 3D ultrasound scans . . . . .	8
1.4.4 Tooth instance segmentation in 3D point cloud data . . . . .	9
1.5 Problem statement and research questions . . . . .	10
1.6 Contributions . . . . .	13
1.6.1 Contributions to molecular imaging . . . . .	14
1.6.2 Contributions to computational pathology . . . . .	14
1.6.3 Contributions to image-guided interventions . . . . .	15
1.6.4 Contributions to computerized dentistry . . . . .	15
1.7 Thesis outline and scientific background . . . . .	16
<b>2 Mass spectrometry analysis</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.2 Related work . . . . .	22
2.2.1 Conventional machine learning approach . . . . .	22
2.2.2 Deep learning approach . . . . .	23
2.2.3 Contributions based on the related work . . . . .	25
2.3 Methodology . . . . .	26
2.3.1 MS analysis using 1D ConvNets . . . . .	26
2.3.2 MS analysis using RNN . . . . .	28
2.4 Experimental results . . . . .	29
2.4.1 Dataset . . . . .	29
2.4.2 Evaluation metrics . . . . .	30
2.4.3 Results . . . . .	30

2.5	Discussion and Conclusions . . . . .	38
<b>3</b>	<b>Two-dimensional Histopathological Image Analysis</b>	<b>41</b>
3.1	Introduction . . . . .	41
3.1.1	Histopathology . . . . .	41
3.1.2	Computational pathology . . . . .	41
3.1.3	Challenges in computational pathology . . . . .	42
3.2	Stain-color normalization . . . . .	44
3.2.1	Related work . . . . .	44
3.2.2	GAN-based method . . . . .	46
3.2.3	Empirical evaluation of GAN-based method . . . . .	49
3.2.4	Neural-GMM method . . . . .	53
3.2.5	Empirical Evaluation of Neural-GMM method . . . . .	58
3.2.6	Discussion . . . . .	60
3.3	Cancer detection and classification . . . . .	62
3.3.1	Metastases detection and grading using ConvNets . . . . .	62
3.3.2	Cancer detection using CRFs on deep embedded spaces . . . . .	63
3.3.3	Empirical Evaluation . . . . .	69
3.3.4	Discussion and Conclusion . . . . .	71
3.4	Impact of JPEG compression on deep neural networks . . . . .	72
3.4.1	Introduction and related work . . . . .	72
3.4.2	Materials and Methods . . . . .	75
3.4.3	Automated Tumor Detection . . . . .	78
3.4.4	Experiments and Evaluation . . . . .	79
3.4.5	Results . . . . .	81
3.4.6	Discussion . . . . .	84
3.5	Conclusions . . . . .	85
<b>4</b>	<b>Needle Localization in Volumetric 3D Ultrasound Imaging</b>	<b>89</b>
4.1	Introduction and related work . . . . .	89
4.2	Deep learning methods for needle localization . . . . .	91
4.2.1	Proposed frameworks . . . . .	91
4.2.2	Patch classification - primary technique . . . . .	92
4.2.3	Semantic segmentation - secondary technique . . . . .	94
4.2.4	Needle axis estimation and visualization - post-processing . . . . .	96
4.2.5	Implementation details . . . . .	97
4.3	Experimental results . . . . .	97
4.3.1	Patch classification . . . . .	98
4.3.2	Semantic segmentation . . . . .	100
4.3.3	Axis estimation accuracy . . . . .	104
4.4	Discussion . . . . .	105
4.5	Conclusion . . . . .	106
<b>5</b>	<b>Teeth Instance Segmentation in 3D Point Cloud Data</b>	<b>109</b>
5.1	Introduction . . . . .	109
5.2	Related work . . . . .	111

5.2.1	Conventional IOS segmentation approaches . . . . .	112
5.2.2	Deep learning approaches . . . . .	112
5.3	Model 1: Semantic segmentation in 3D point cloud scans . . . . .	113
5.3.1	Technical aspects and contributions of method 1 . . . . .	113
5.3.2	Non-uniform re-sampling . . . . .	115
5.3.3	Model architecture . . . . .	116
5.3.4	Experiments and evaluation of Model 1 . . . . .	120
5.3.5	Discussion . . . . .	122
5.4	Model 2: Instance segmentation in 3D point cloud scans . . . . .	122
5.4.1	Related work on instance segmentation in 3D point cloud . . . . .	125
5.4.2	Concepts and novelty of the proposed approach . . . . .	126
5.4.3	Mask-MCNet . . . . .	128
5.4.4	Experimental results . . . . .	138
5.4.5	Discussion . . . . .	145
5.5	Conclusions . . . . .	145
<b>6</b>	<b>Conclusions</b>	<b>147</b>
6.1	Conclusion of the individual chapters . . . . .	147
6.2	Discussion on the research questions . . . . .	150
6.3	Utilization and outlook . . . . .	154
	<i>Bibliography</i>	161
	<i>Acknowledgment</i>	177
	<i>Publication list</i>	179
	<i>Curriculum vitae</i>	183



# CHAPTER 1

## Introduction

### 1.1 Role of medical imaging in modern medicine

Medical imaging is an essential part of modern medicine that helps in the diagnosis, treatment, and monitoring of various diseases. It enables healthcare professionals to visualize and evaluate the internal structure and function of the human body. Medical imaging has multiple significant impacts on medicine, including the following list.

(1) *Early Detection and diagnosis* of diseases, conditions, and injuries, which can lead to earlier treatment and better outcomes for patients. (2) *Accurate diagnosis* by providing detailed images of the internal anatomic structures, which helps healthcare professionals to make an accurate diagnosis. This can lead to better treatment planning and more effective treatment. (3) *Performing minimally invasive procedures* that improve the patient safety and can be used to guide biopsies and local surgeries, which reduce the risk of complications and results in a faster recovery. (4) Supporting the healthcare professionals to provide *personalized treatment* plans based on the individual patient's needs and conditions. (5) Facilitating and playing a vital role in *research and development* in the medical domain, since it allows researchers to study the effects of diseases and treatments on the human body.

Considering the above impacts and influences, it is not surprising to state that more than 50% of all medical data has become imaging data or image-based in one form or the other. This percentage is growing every year, not only because image data are increasing stored and used for further analysis, but also because of the amount of modalities and types of imaging systems are increasing as well. Therefore, it can be safely claim that medical image data have become an essential tool in medical diagnosis and treatment planning.



## 1. INTRODUCTION

---

Understanding the different types of imaging systems can help healthcare professionals to select the most appropriate imaging modality for measuring a particular medical condition or performing a diagnosis. While employing an appropriate imaging modality is important, the analysis and interpretation of the acquired imaging data needs specific expertise and experience. This explains why at present physicians are well educated and trained in the usage of specific imaging modalities, like X-ray, Magnetic Resonance Imaging (MRI), Ultrasound (US) imaging, (digital) video endoscopy, etc.

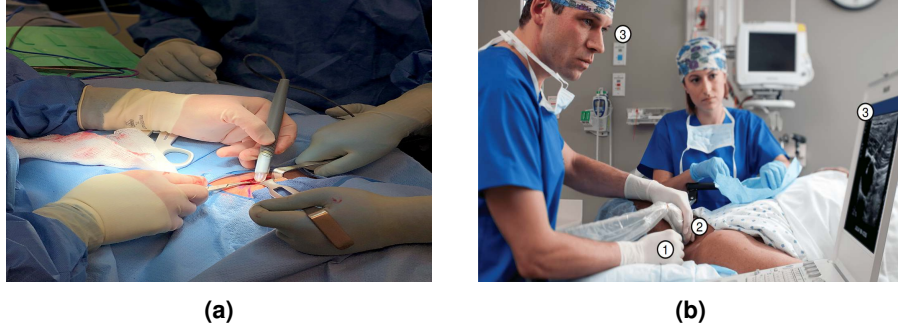
### 1.2 Computer-aided diagnosis in medical imaging

With the advancements in computer vision and machine learning algorithms, Computer-aided diagnosis (CAD) has become a valuable tool in the field of medical imaging. CAD has the ability to analyze large amounts of medical data to provide accurate and timely diagnoses. One of the key benefits of CAD is its ability to detect small abnormalities in medical images that may not be visible to the human eye. This can be particularly useful in the detection and diagnosis of diseases such as cancer, where early detection can significantly improve treatment outcomes. CAD can also help radiologists to more accurately interpret complex medical images, reducing the chances of misdiagnosis and improving patient outcomes.

Another significant benefit of CAD is its ability to reduce the time and cost associated with medical imaging. By automating the analysis of medical images, CAD can significantly reduce the time and resources required for diagnosis. This can be particularly beneficial in situations where rapid diagnosis is critical, such as emergency medical situations. Furthermore, by reducing the need for manual interpretation of medical images, CAD can help to reduce the likelihood of mistakes and subtleties in diagnosis, leading to better patient outcomes. In this sense, CAD can facilitate the average-skilled physician in assisting the diagnosis and interpretation and even improve its accuracy.

Another aspect of employing CAD that has been less developed is real-time interventional diagnosis and surgery that requires direct/online analysis of acquired data that can be processed by using an efficient and accurate CAD model. Two examples of CAD systems that recently have been developed are MasSpec Pen [1] for real time, in-vivo cancer diagnosis during surgery using the mass spectrometry modality, and in-vivo needle detection and localization in 3D ultrasound imaging. Overall, the usage of computer-aided diagnosis in medical imaging has a proven capability to improve patient outcomes and reduce the burden on healthcare providers.

Conventional CAD methods in medical imaging rely on expert-defined rules and algorithms to detect specific features or patterns in the images. These methods are based on traditional statistical and mathematical techniques and require extensive domain knowledge to design and optimize such techniques. Conventional methods are often effective in specific scenarios where the diagnostic fea-



**Figure 1.1** Two examples of online CAD systems. (a) MasSpec Pen (Image source: wikipedia.org) is a mass spectrometry-based cancer detection and diagnosis system that can be used for ex-vivo and in-vivo tissue sample analysis. (b) Interventional US-based needle detection and localization in 3D scans (Courtesy of Philips Ultrasound).

tures are well-defined and can be easily extracted from the images. However, they may not be able to handle the complexity of large datasets and can be limited by the assumptions and biases of the experts who design them.

Alternatively, new machine learning and deep learning models have emerged as powerful tools for CAD in medical imaging that revolutionized the conventional methods. These models can learn to automatically extract relevant features and patterns from large datasets of medical images, without fully relying on expert-defined rules or assumptions. Machine learning and deep learning models are particularly effective in scenarios where the diagnostic features are complex and difficult to define explicitly. These models can also improve over time when they are exposed to more data, leading to more accurate and reliable diagnoses.

### 1.2.1 Data dimensionality in imaging

One factor that has an important role in the design of a deep learning architecture is the intrinsic dimensionality of the input data. The architecture of a model can be adjusted to best suit the dimensionality of the input data to improve its efficiency and accuracy. In medical imaging, systems produce data in variable amounts of dimensions that sometimes require specific network architectures for achieving the highest performance.

*One-dimensional (1D)* data, such as time-series data or mass spectrometry, require the neural architecture to be capable of learning the patterns over time and can effectively process time-sequential or spectrally-ordered data. Convolutional neural networks (ConvNets) with 1D convolution kernels and Recurrent neural networks (RNNs) are two different families of networks that are commonly being used for analyzing data that have a long-term temporal structure like time-series data.

*Two-dimensional (2D)* images mostly have high local dependencies within their 2D space. The type of employed neural architecture should be capable of learning spatial patterns and can effectively process image-based data. For processing 2D images, mostly the ConvNets with 2D kernels or vision transformers [2] are adjusted to improve the efficiency and accuracy of the model.

*Three-dimensional (3D)* data such as volumetric 3D scans usually process by the ConvNets with 3D convolution kernels. This type of network is capable of learning both spatial and temporal patterns and can effectively process 3D volumes. However, depending on the size of the volume and the computational capacity of the CAD system, real-time deployment of such models is challenging.

*Unstructured 3D* data such as point cloud scans are composed of sparse data points in 3D space. For processing such unstructured data, usually the Multilayer Perceptron (MLP) architecture is typically used. Compared to the 2D/3D image data, processing a point cloud requires dedicated design in the architecture to address some inherent properties of the data, such as variable size and permutation of sampled points.

In conclusion, the design of a deep learning architecture should be based on the dimensionality of the input data. By choosing the appropriate type of neural network and adjusting its architecture, the efficiency and accuracy of the model can be improved.

### 1.3 Recent advancement in medical image analysis

Medical image analysis has been an active area of research and development, where significant advancements have been made in recent years. Some of the notable advancements are listed below.

*Deep Learning* techniques, such as convolutional neural networks (ConvNets), have shown great promise in improving the accuracy of medical image analysis. These techniques have been used in various applications, such as segmentation, detection, and classification, and have contributed to improved diagnosis and treatment planning.

*Multi-Modal Imaging*, which combines data from multiple imaging modalities, such as MRI, CT, PET, and ultrasound, has been shown to improve the accuracy of medical image analysis. Each modality provides a part of the description of the patient, where an alternative modality serves as complementary analysis of the patient's condition, leading to better diagnosis and treatment planning.

*3D Imaging* technologies, such as CT, MRI, and US 3D scans provide volumetric data, which can be processed by an advanced image analysis technique for 3D modeling and producing more detailed information about the patient's condition. This has led to improved accuracy and reduced false positives.

*Personalized Medicine* is an emerging field that uses patient-specific data, such as medical history, genetic information, and imaging data, to develop customized treatment plans. Medical image analysis plays a crucial role in personalized

medicine by providing accurate and precise information about the patient's condition.

*Explainable AI* helps clinicians to understand how the learning-based AI system makes decisions. Recent advancements in explainable AI techniques enable better understanding of the reasoning behind the diagnosis, leading to an increased trust and adoption of these automated CAD systems.

Overall, these advancements in CAD systems have resulted into significant improvements in the accuracy and efficiency of diagnosis and treatment planning, ultimately leading to better patient outcomes.

#### 1.3.1 CAdE and CADx systems

CAD for medical image analysis includes CAdE (Computer-Aided Detection) and CADx (Computer-Aided Diagnosis), two related but distinct technologies.

In radiology, CAdE systems are designed to assist radiologists in identifying suspicious regions in medical images such as mammograms, CT scans, and MRIs. These systems analyze the images and provide alerts when they detect potential abnormalities in anatomy, such as tumors or lesions. CAdE systems do not provide a definitive diagnosis but serve as a second opinion to assist radiologists in making accurate diagnoses.

Alternatively, CADx systems are designed to provide a diagnosis based on medical images. These systems use algorithms and machine learning techniques to analyze medical images and provide a diagnosis or a likelihood of a certain disease or condition.

Both CAdE and CADx systems rely on advanced image processing and machine learning algorithms to analyze both 2D or 3D medical images. Additionally, both systems can potentially improve the accuracy of diagnosis by detecting subtle abnormalities that may be missed by human interpretation. This advantage originates from the capability of a CAD system to memorize the relevant features, learned through multiple training examples for enhancing the decision-making. Another aspect where CAdE can facilitate physicians is its operational speed in comparison with human: a subtle abnormality can be detected in a fraction of a second.

However, there are also some distinctions between the two technologies. CAdE systems are primarily designed to assist radiologists in detecting potential abnormalities, while CADx systems provide a more definitive diagnosis. Additionally, CADx systems are often more complex than CAdE systems and require a larger training dataset to achieve the desired high levels of accuracy.

This thesis presents several deep learning-based methods for building CAdE and CADx systems in four different medical fields, including (1) cancer diagnosis by classification mass spectrometry data, (2) metastases detection and segmentation in histopathological slides, (3) medical instrument detection and localization in US volumes, and (4) teeth detection and segmentation in digital dentistry.

### 1.4 Research scope and challenges

Advanced deep learning-based solutions for processing medical imaging data in each of the four aforementioned imaging modalities provides substantial improvement in automating a clinical procedure by designing better CAD systems. This thesis investigates the core problems in each of these four medical imaging domains and provides learning algorithms for each of those problems. The following subsections introduce each core problem in detail.

#### 1.4.1 Mass Spectrometry Imaging (MSI)

Mass Spectrometry Imaging (MSI) is a powerful technique that allows for the direct visualization and spatial mapping of molecules in tissue samples. Recent advances in MSI have led to the development of high-resolution imaging techniques. These techniques enable the visualization of smaller molecules and provide more detailed information about their distribution in tissue samples. In contrast to other imaging systems like CT, MRI, or bright-field microscopy that provide information about the structural features of the examined tissue, MSI provides molecular features. Integration of MSI data with other clinical and "omics" data sources, such as genomics and proteomics, can provide a more comprehensive understanding of disease mechanisms and pathways. This approach can lead to the identification of new biomarkers and targets for therapy.

Moreover, MSI has the potential to be used for early cancer detection. MSI allows for the visualization and quantification of the presence and the amount of biomolecules, such as proteins, lipids, and metabolites, within tissues. In cancer, there are often changes in the types and amounts of biomolecules occurring, while MSI can detect these changes with high sensitivity and specificity. For example, MSI has been used to detect changes in lipid metabolism in breast cancer, and alterations in protein expression in lung cancer. MSI can be used also to analyze biopsy samples, which can provide information on the molecular makeup of the tissue and thereby potentially aid in the early detection of cancer. However, further research is needed to fully validate the use of MSI for early cancer detection, as well as to optimize the technique and develop standardized protocols for its use in clinical settings.

The most important challenge in developing a deep learning solution for many healthcare domains in general and in particular for MSI analysis is the limitation of gathering large-scale clinical datasets. This limitation is mainly due to the lower availability of some advanced medical imaging systems and their lower throughput, compared to the vision sensors. A further complication is that the development of a new data-driven technique needs acquiring clinical data in an uncommon and unmaturing procedure that is typically prohibited by the patient safety considerations and medical approval boards.

Besides the limitation of relatively low availability of sufficient clinical data, the manual labeling of the MSI samples is challenging, since the acquired spectral MSI data is hard to be interpreted or discerned by the human eye. There-

fore, usually a complementary imaging system is applied in a multi-modal data-acquisition setup. Thus, the labels collected from the other imaging system are transferred across modalities and are assigned to the MSI samples. Such a multi-modal data acquisition introduces other challenges such as image alignment across modalities which is prone to errors in transferring label data into the MSI data.

Furthermore, each MS sample presents a molecular profile of the examined sample in the form of a histogram of intensities over several thousands mass-to-charge bins. For cancer detection and classification, the processing technique needs to be capable of learning relevant features across those bins. This can be a complex problem for the learning model, since there is a high redundancy in the data and only a subset of bins convey relevant information for the defined classification task. Therefore, the employed neural network should be capable to learn and identify informative patterns both locally and globally across data elements (i.e. bins). It is evident that this can be a complex task, especially within a limited data problem.

Since automated analysis of MSI data is less frequently explored, it is not clear how the association of bins can result in finding patterns for cancer detection and classification. However, this is important for imposing an inductive bias<sup>1</sup> in the architecture design of a neural network for achieving a higher performance in relevant clinical diagnosis tasks.

The above-mentioned points have hampered the progress of deep learning techniques for MSI analysis. Advances in this domain requires developing a learning model that can handle the existing challenges to achieve the highest performance for cancer diagnostics.

### 1.4.2 Computational histopathology

Computational pathology is an interdisciplinary field that combines digital pathology, machine learning, and other computational techniques to analyze and interpret histological images of tissues. This field aims to develop automated algorithms and tools to assist pathologists in the diagnosis and treatment of diseases.

Traditionally, pathologists have relied on manual examination of tissue samples to diagnose diseases such as cancer. With the advent of digital pathology, histological images can now be digitized and analyzed using computer algorithms. This approach allows for more objective and quantitative analysis of tissue samples and can help to identify subtle patterns and features that may be difficult to discern through manual examination alone.

Nowadays, CAD systems based on deep learning techniques are commonly used in computational pathology to develop algorithms that can recognize and classify various tissue types and disease states. These algorithms can be trained

---

<sup>1</sup>The inductive bias refers to incorporating domain knowledge in the design of the neural network architecture.

on large datasets of histological images, allowing them to learn patterns and features that are indicative of specific diseases.

Computational pathology has the potential to significantly improve the accuracy and diagnosis speed of diseases and associated treatment decisions. Furthermore, it may provide new insights into the underlying mechanisms of a disease. It is an exciting and rapidly evolving field that holds great promise for the future of pathology and medicine. However, there are several challenges in designing a highly accurate CAD system for analyzing the histopathological slides. Some of the key challenges in this field are briefly discussed below.

*Data variability:* Tissue samples can exhibit significant variation in their histological appearance like their colors, even when originating from the same anatomical location. This variability can pose a challenge for developing automated algorithms for analyzing histopathological images. There are several factors that are related to the introduction of such visual variations, which are discussed later (more detail in Chapter 3). The presence of such uninformative visual variations among images increase the generalization error of the learning model. Therefore, developing a robust CAD system requires an adaptation or normalization mechanism to reduce such variations across datasets.

*Large-scale histopathological data:* Histopathological images can be very large in size, where the amount of data generated during image acquisition and analysis can be enormous. For example, a single histopathological slide can include more than  $100k \times 100k$  pixels. Therefore, sampling the Whole Slide Images (WSIs) by small patches and individual analysis of each patch, is a common approach. However, this affects the learning model, since the prediction per patch is limited to the amount of visual context in each patch. Therefore, decision-making at the WSI level requires a fusion mechanism across the patch-level predictions.

*Histopathological image compression:* As mentioned above, the histopathological WSI size is enormous, which can pose challenges for data storage and transmission. Therefore, employing image compression techniques is essential. In case of using a lossy compression technique like JPEG compression, the impact of the introduced image degradation or artifacts should be accounted for when designing a data-driven CAD system. Although there are some research works that investigate the impact of image compression on the human observer for histopathological image analysis, there is no evidence how it affects the training and inference of a deep learning model.

### 1.4.3 Instrument localization in 3D ultrasound scans

Automating tracking of interventional instruments like an injection needle under ultrasound guidance provides a significant improvement and ease of use to the quality of minimally-invasive procedures. Due to a thin planar beamforming field in conventional 2D US imaging, an extensive bi-manual manipulation of the US probe is required in exploring the fine structures in the 3D tissue volume. Therefore, by manual manipulation of the US probe, it is less likely to completely align the US image plane with the sheaf of planes that include the needle axis. Al-

ternatively, available 3D US transducers can provide the 3D US volume, instead of a single US plane. The obtained US volume can be processed for automated detection and visualization of the sheaf of needle planes to the medical specialist during intervention. Later in this thesis (in Chapter 4) the focus is on designing a robust and efficient deep learning model for needle localization in 3D US volumes. This approach facilitates simplified manual coordination of the imaging equipment.

For designing and developing a reliable and accurate CAD system to localize a needle in US volume, there are several challenges that need to be addressed adequately. Physical properties of US imaging introduce inherent quality degradation into the image data such as anisotropy, speckle noise, clutter, and imaging artifacts. Hence, the appearance of a needle in US images is distorted, which makes the detection harder. Moreover, the signal-to-noise ratio (SNR) of US data is very low, which causes ambiguity for the object detection and localization algorithm. Reflection of US waves at interfaces between fluid, fat, muscle and connective tissues create speckled echogenic regions and reduce the visual contrast. The appearance of such echogenic structures in the US images locally resemble the needle shape that to some extent clearly complicate needle discrimination. Lastly, the intensity of US reflection from the surface of the needle, depending on its angle and type can be much lower than other anatomic structure such as bone, tendon, fascia, and muscle fibers.

To address the above-mentioned challenges for robust and accurate needle localization in 3D US volume, Chapter 4 presents two different deep learning-based techniques to process 3D volumetric data for detection and segmentation of needle voxels.

#### **1.4.4 Tooth instance segmentation in 3D point cloud data**

Intra-oral scanners are advanced imaging devices for optical measurement of the surface profiles of anatomical structures inside the oral cavity of the patient. Similar to other 3D scanners, Intra-oral Scanners (IoS) project a light source (laser, or structured light) on the surface of objects to be scanned, in this particular case, the dental arches. Based on the underlying technique, the time-of-flight of the laser or the deformation of the projected pattern on the subject's surface is measured by the imaging sensors and processed by the scanning software. The obtained point cloud can be further processed and converted into a 3D surface model (i.e. mesh) by using triangulation techniques. Such a precise 3D model is widely used for implant treatment and orthodontic planning. A scan of one dental arch consists of a large set of points (e.g. hundreds of thousands) in the 3D Cartesian coordinate system.

The obtained intra-oral scans can have several use cases in dentistry. Segmenting the teeth instances in the scans supports different dental treatments. For example, it advances the orthodontic planning in several ways as follows.



## 1. INTRODUCTION

---

- *Fabrication of orthodontic appliances*: segmentation of the teeth instances can be used to create precise 3D models of a patient's teeth, which can then be used to fabricate custom orthodontic appliances such as braces or aligners.
- *Virtual simulations*: Orthodontists can use segmented scans to create virtual simulations of how a patient's teeth will move during treatment. This allows the orthodontist to plan and adjust treatment more precisely, and can also help patients to understand what to expect during the course of the treatment.
- *Monitoring treatment progress*: segmentation of scans allows to monitor the patient's progress during treatment, comparing scans taken at different stages, to assess how each individual tooth is moving and treatment is adjusted as needed.

Overall, automating the process and analysis of IoS provides orthodontists with a powerful tool for diagnostics, planning, and monitoring, which can lead to more precise and effective outcomes for patients. Such automation requires developing multiple processing modules which lead to accurate and reliable tooth instance segmentation as a solution to the core problem.

Tooth instance segmentation refers to assigning a unique label (e.g. index) to all points belonging to each individual tooth. Relevant problems for an accurate segmentation are (1) the existing variations in the patients dentition, (2) missing data due to non-reflected areas of the mouth cavity, (3) presence of outliers such as implants on some teeth crowns of the patient.

In addition to the complexity of the data, precise segmentation requires processing the point cloud at its highest available resolution. While the average size of IoS scans includes several ten thousands of points, this can be challenging for learning systems with limited memory capacity for an efficient training. Although using a common down-sampling technique can resolve this issue, it is apparent that preserving the geometrical details such as the curvatures at the borders of neighboring teeth are important in the overall performance of the segmentation method.

To advance the CAD systems in the dentistry, Chapter 5 of this thesis focuses on the tooth instance segmentation as the core problem in this field. Two different methods for IoS segmentation are presented and the aforementioned challenges are addressed.

### 1.5 Problem statement and research questions

To address the mentioned technical and clinical challenges and to design deep learning-based solutions for various medical imaging systems with different intrinsic data dimensionality, this thesis defines the following problem statement.

**Problem statement** *The key problem of this thesis is to design and develop deep learning-based methods to analyze and extract important clinical information from the*

*data that is collected in four different medical imaging modalities. Each of the data modalities has its own intrinsic dimensionality that should be considered to achieve the highest performance of the aimed clinical diagnosis task.*

### Research questions

Based upon the above problem statement, specific research questions (RQ) can be derived, which are formulated below.

**RQ1. Metastases detection in 1D mass spectrometry data** MSI can detect a wide range of molecular features that are associated with cancer. For example, lipids and proteins are two important classes of molecules that, based on some clinical evidence, are related to cancer diagnosis. The patterns of such molecules spread over the entire mass spectrum. Therefore, the molecular features in MSI are scattered locally and globally over the mass-to-charge bins of MS data. This property requires specific processing by the machine learning algorithm, in order to make it capable of learning the association with the desired high-level clinical label, such as a binary label as “normal” or “cancer” tissue. Therefore, the following research questions are formed.

**RQ1a.** *What is the influence of the receptive fields of convolution kernels in ConvNets to enhance the expressiveness of the learned features and add to the overall performance of cancer detection?*

**RQ1b.** *For learning the existing dependencies across the mass spectrum, which architecture of Recurrent Neural Networks (RNNs) operates best and can the RNNs outperform the ConvNets with 1D kernels for cancer detection?*

**RQ2. Deep learning-based 2D histopathological image analysis** Histopathology image analysis is the most common clinical method for cancer diagnosis. Designing a robust, reliable, and efficient CAD algorithm for the analysis of 2D histopathological images can involve several modules, each requiring different computational techniques. This thesis studies three important challenges, which have been introduced in Section 1.4.2, including stain-color variability, cancer detection in large-scale histopathology images, and the impact of image compression on the training and inference of a deep learning-based CAD system.

Due to the clinical procedure involved in stain colorization of tissue slides prior to imaging, and different characteristics of the imaging system, the acquired pathological images may have significant differences in their chromatic information. Since the color information is an important clue for the diagnosis of cancerous cells, the deep learning CAD model should be robust against such color variations. Although some techniques such as color augmentation reduce the generalization error, the CAD model still suffers from the unseen color variations across data, which are provided by different pathology laboratories.

## 1. INTRODUCTION

---

Another important problem in histopathology image analysis is the capability of doing inference on the whole-slide images, as the sizes of WSIs are enormous. Therefore, the common approach is partitioning each WSI into small image patches and processing them individually. As a result of this, integrating the learned features for doing a holistic inference and learning the transformation between patch-level features and the target clinical label becomes an important part of the learning model.

Last but not least, due to the large size of WSIs, image compression techniques and particularly the JPEG2000 algorithm are used for compression and storage of the histopathological imaging data. Since a lossy compression degrades the quality of original images, several research works investigated the impact of such degradation on the diagnostics performance of a professional observer. Unfortunately, it has not been broadly investigated what is the influence of compression artifacts and degradation are on the CAD system for cancer detection. This thesis elaborates on several research questions on the above issues as follows.

**RQ2a.** *Is it possible to model the process of stain-colorization in pathology with a generative neural model? And how can we make this framework a generic solution for various types of histopathological examinations?*

**RQ2b.** *How can we create a hybrid algorithm that combines a probabilistic model and a neural network to ensure a high level of fidelity in color conversion process?*

**RQ2c.** *How can we effectively combine and integrate data across features that have been learned through individual patch-level analysis? Is it feasible to achieve this fusion by employing a probabilistic graphical model?*

**RQ2d.** *How does the utilization of a standardized data compression technique impact the training and inference performance of a deep learning model? What compression ratio is recommended to achieve optimal performance for histopathology image analysis?*

**RQ3. Deep learning-based needle localization in 3D US scans** The ConvNets can achieve a substantial improvement in the detection accuracy of needle voxels in 3D US data, by better learning and extracting discriminating features for this task. However, processing of 3D US volumes by common 3D convolutional kernels is time-consuming. Furthermore, needle detection in US volumes can be formulated either as patch classification problem or an image classification procedure. Each approach has its own advantages and shortcomings in terms of efficiency, latency, and overall accuracy. The dominant issue is the elegant exploitation of 3D US volumetric data for needle detection and its localization.

**RQ3a.** *How can we use 2D orthogonal convolutional kernels as a proxy for 3D kernels in ConvNets and develop a shared branch architecture for efficiency and low number of learning parameters?*

**RQ3b.** *Can the segmentation of needles in a 3D ultrasound (US) volume be formulated as an image-slice segmentation task? Does this approach yield better results compared to the patch classification technique, in terms of relevant evaluation metrics?*

**RQ4. Deep learning-based tooth instance segmentation in 3D point cloud** As mentioned earlier, automating the analysis of IoS provides implantologists and orthodontists with a powerful tool for diagnosing, planning, and monitoring, which can lead to more precise and effective outcomes for patients. Accurate and reliable tooth instance segmentation in IoS scans is the core problem for enabling such automation.

In semantic segmentation of teeth, for assigning a unique clinical label to the points of each tooth on the dental arch, the shape and relative position of each tooth with respect to other teeth needs to be taken into account by the deep learning model. Therefore, the input to the network should contain a global structure of dentition. Alternatively for accurate segmentation, the preservation of geometrical details of scans is important. Since processing the whole IoS with a huge number of points is not feasible for common computational hardware, a re-sampling technique is required that preserves the important information of data.

As an alternative solution for processing the large-size point cloud, instead of quantizing points by applying a conventional technique such as occupancy grid, a neural network can be employed for learning implicit features and transferring local information from the points to the closest node on the regular 3D grid. This facilitates processing large-scale point clouds while the fine detail information is processed and abstracted, such that the structure of the point cloud is presented on a regular 3D grid.

**RQ4a.** *Can we create a non-uniform re-sampling method that effectively preserves both the local intricate details and overall structure of dentition in intraoral scanner (IOS) point clouds? Additionally, how can we utilize a neural adversarial framework to impose the prior distribution of teeth positions on the dental arch?*

**RQ4b.** *Can we devise a novel tooth instance segmentation method that utilizes a regular grid representation for the 3D point cloud data obtained from intraoral scanner (IOS) while addressing the degradation issues commonly associated with conventional discretization methods such as occupancy volume representation?*

## 1.6 Contributions

The scientific and technical contributions of this thesis can be partitioned into four groups, each linked to a particular medical imaging system, which are summarized below.

### 1.6.1 Contributions to molecular imaging

For improving mass spectrum classification, the use of ConvNets based on 1D convolutional kernels with various dilation factors, is proposed. The conducted experiments on cancer classification of MS data show that such an architecture outperforms state-of-the-art results with 1-3% in balanced accuracy by learning better discriminating features across short-distance and long-distance mass spectra.

Furthermore, as an alternative approach, recurrent neural networks are exploited for learning local and global dependencies in the mass spectrum. An architecture search on several design choices of the RNN is performed and the architecture that operates best on the cancer classification is introduced. The number of LSTM modules and layers, employing a bidirectional architecture, different pooling mechanisms, and using a batch normalization operator, are some of the important factors that are explored in the architecture search. The proposed RNN-based network shows superior performance in comparison with the ConvNet solution by achieving an accuracy growth of 1.87% and 1.45% on two clinical datasets.

### 1.6.2 Contributions to computational pathology

**A. Stain-color normalization:** An important aspect of a CAD system is its generalization power on unseen data. We formulate the stain-color variation in histopathology image analysis as a generative process, modeled by adversarial generative networks. The proposed framework is developed by making minimal assumptions about the histopathological image contents and does not require clinical labels for learning stain-color normalization. These properties make the proposed method more generic, so that it is applicable across various histopathological image analysis tasks for examination of different tissue types.

Moreover, as an alternative stain-color normalization method, called neural Gaussian Mixture Model (NeuralGMM), is introduced to ensure the model *faithfulness* to the structure of tissues in the process of stain-color conversion. We show that the proposed NeuralGMM outperforms state-the-art models in histopathological color normalization. This improvement is measured across five laboratories with a clearly lowered standard deviation and smaller coefficient of variation of NMI (normalized median intensity), being as low as 0.017 and 0.036, respectively.

**B. Metastases classification in Histopathology WSIs:** Due to the enormous size of WSIs, the patch-level analysis is usually employed as a common practice. For improving the overall performance, we have proposed to incorporate the patch-level features and predictions in a graph. Each node of such a fully-connected graph represents the neural features, learned from each patch of WSIs. We have found that applying conditional random fields (CRFs) on the graph and performing data fusion across patches, the overall performance increases signif-

icantly. This result was reported as state-of-the-art achievement (top-2) on an international leader board for metastases segmentation and classification.

**C. Impact of image compression on DL model:** Since image compression is an essential module for storing and transferring histopathological images, we have investigated the effect of various image compression ratios on the training and inference of a deep ConvNet. By introducing various scenarios of training and inference on low and highly compressed images, we have studied the impact of image degradation on the classification performance of neural network.

The results show that when the network is trained on uncompressed images, its performance on the compressed images in term of  $F_1$  score and AUC values up to a factor of 24 is maintained. However, for a ratio of 32:1 or higher, the  $F_1$  scores start to drop significantly. Additionally, we have found that when the compressed images are used for training the network, the improvement is significant and the model performs reliably by achieving an  $F_1$  score of 0.93 with a compression factor of 164:1, while it is 0.58 when the model is trained on uncompressed images.

### 1.6.3 Contributions to image-guided interventions

For efficient processing of volumetric US data, we propose 2D orthogonal convolutional kernels as a proxy for 3D kernels in ConvNets. It is shown that the proposed ConvNet architecture can detect and localize the needle in US data by processing the orthogonal multiple views of US patches. This approach improves the needle detection by 25% in  $F_1$  score, compared to conventional Gabor-based feature classification.

As an alternative method, instead of US patch classification, we propose a semantic needle segmentation technique in 3D US volumes. The proposed method is based on decomposing the 3D volume into 2D cross-sections for labeling the needle parts and reconstructing the 3D needle labels from multiple views. Therefore, in the proposed approach, the number of parameters in the convolution kernels decreases exponentially compared to full 3D kernels, so that the network requires fewer training samples and executes faster.

The semantic segmentation approach based on 2.5D US information achieves a 84%  $F_1$ -score in the porcine leg datasets that are acquired with a lower-resolution phased-array transducer. These results show strong semantic modeling of the needle context in challenging situations, where the intensity of the needle is inconsistent and even partly invisible.

### 1.6.4 Contributions to computerized dentistry

An end-to-end deep learning model is proposed for semantic tooth segmentation from point clouds derived from IOS data. For analysis of point clouds in their original spatial resolution (resulting in predictions for all points), we have introduced a non-uniform re-sampling technique and a compatible loss weighting,

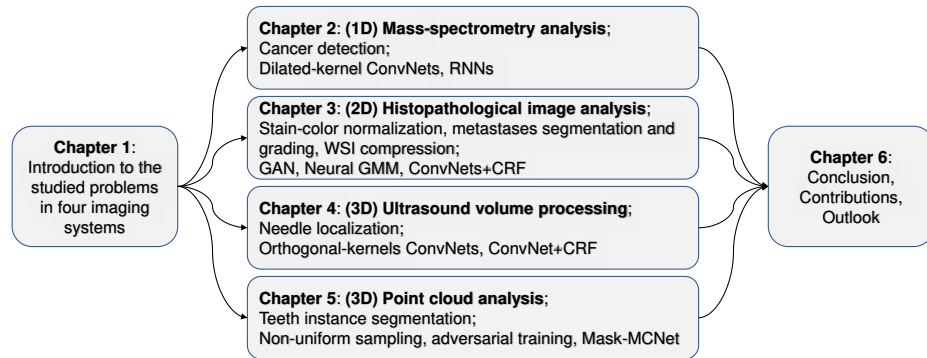
based on foveation and Monte Carlo sampling. This re-sampling approach includes both local, fine-detail information and the sparse global structure of data, which is essential for an accurate prediction of each individual point.

Additionally, it is shown that semantic segmentation of teeth can be improved when the distribution of labels is used in the form of adversarial training. This is mainly due to the high dependency of the semantic label of each tooth to its relative position with respect to other teeth on the dental arch.

In an alternative approach, we present Mask-MCNet, a new end-to-end deep learning framework for tooth instance segmentation in a 3D point cloud of an intra-oral scan. In contrast to existing deep learning models, the Mask-MCNet does not employ a voxelization or down-sampling step for processing the large point cloud. The IoS scan is processed at its native high resolution, thereby preserving the finely detailed geometrical information, which is crucial for accurate teeth segmentation. The conducted experiments have shown that the proposed model achieves a 98% mIoU on the test data, thereby outperforming the state-of-the-art networks in tooth instance segmentation. This level of performance is close to the human level and obtained in only a few seconds of processing time, whereas for a human it would form a lengthy and labor-intensive task.

### 1.7 Thesis outline and scientific background

In this section, an overview is provided of the chapters presented in this thesis with the related publication background. The structure of thesis is visualized in Figure 1.2 and is briefly summarized as follows.



**Figure 1.2** Layout of the thesis

As shown in Figure 1.2, the layout of this thesis is simple and covers four contributing chapters. These chapters address individually different imaging

systems, where deep learning is applied to improve information processing and extraction. The data dimensionality is one of the most important properties of every learning model, which steadily grows across these four imaging systems, linked to Chapters 2 to 5.

**Chapter 2** studies deep learning methods for cancer classification in mass spectrometry imaging. Two different deep learning architectures including ConvNets with dilated 1D convolutional kernels and RNN architecture are investigated. Both proposed techniques improve the cancer classification of MS data, by learning the local and global patterns. The findings of this chapter were published in IEEE ISBI Conf. 2019 [C-3] and SPIE Medical Conf. 2019 [C-4].

**Chapter 3** studies three important problems in 2D histopathological image analysis including (a) stain-color normalization, (b) metastases segmentation and classification, and (c) impact of image compression on training and inference of a deep ConvNet. Different solutions and in-depth investigations for each problem are presented. The contributions of this chapter for stain-color normalization have been published in the MICCAI Conf. 2018 [C-5], MIDL Conf. 2018 [C-6], and IEEE ISBI Conf. 2018 [C-7]. The contributions on metastases segmentation and classification have been published in SPIE Medical Conf. 2018 [C-9] and in a journal paper IEEE Trans. Med. Imaging (TMI) 2018 [J-5]. The investigation on impact of image compression on training and inference of a deep ConvNet was published in the Journal of Medical Imaging 2019 [J-3].

**Chapter 4** presents two deep learning solutions for needle detection and classification in US volumes. Patch-level analysis for detecting the needle voxel and binary segmentation of orthogonal multi-view images are investigated, where the patches are sampled from the 3D US volumes. The contributions of this chapter have been published in the IJCARS Journal 2018 [J-4], the IEEE IUS Conf. 2018 [C-8], the SPIE Medical Conf. 2018 [C-10], and the MICCAI Conf. 2018 [C-11].

**Chapter 5** focuses on 3D point cloud analysis of intra-oral scans. Teeth instance segmentation is studied as core problem in orthodontic and implantology CAD systems. The chapter presents two first deep learning-based solutions for accurate tooth instance segmentation in IoS data. The contributions of this chapter have been published in the Neurocomputing Journal 2021 [J-1], the MICCAI Conf. 2019 [C-1], and the MIDL Conf. 2019 [C-2].

**Chapter 6** summarizes the achieved results and addresses the research questions by discussing the associated contributions. The chapter ends with a future outlook.





## Mass spectrometry analysis

### 2.1 Introduction

A mass spectrum represents the number of ionized particles as a function of their mass-to-charge ratio in a material. Mass Spectrometry Imaging (MSI) as a novel molecular imaging technique renders a chemical profile for an examined material. MS is actually a single point measurement. By repeating the measurements on a two-dimensional (2D) grid, we create imaging of the surface of the considered material. Therefore, the term imaging in this context refers to 2D scanning of a thin section of tissue by measuring the mass spectrum at several locations which are structured on a 2D grid. Clinical diagnostics of cancer requires not only histological and immunohistochemical investigations, but also involves molecular pathological techniques. Advances in the characterization of cancer proteomes improve tumor classification and facilitate the assessment of prognosis and prediction to therapy. MSI is a promising candidate to address the current challenges in the clinical diagnostics of cancer [3].

A mass spectrometer consists of at least three main components: an ionization source, a mass analyzer, and an ion-detection system. The ionization source converts the molecules into gas-phase ions, so that they can be manipulated by external electric and magnetic fields. The moving ions are separated and sorted by a mass analyzer with respect to their mass-to-charge ( $m/z$ ) ratios. The separated ions are measured and visualized by a histogram according to their  $m/z$  ratios. Regarding the instrumentation in MSI, there are different types of mass-spectrometry analyzers depending on the ion sources in use. One important ionization technique is matrix-assisted laser desorption/ionization (MALDI). By this technique, the time-of-flight (ToF) values of ions from the surface of the tissue to the ion detector are measured, which can be different for molecules with different mass-to-charge values.

## 2. MASS SPECTROMETRY ANALYSIS

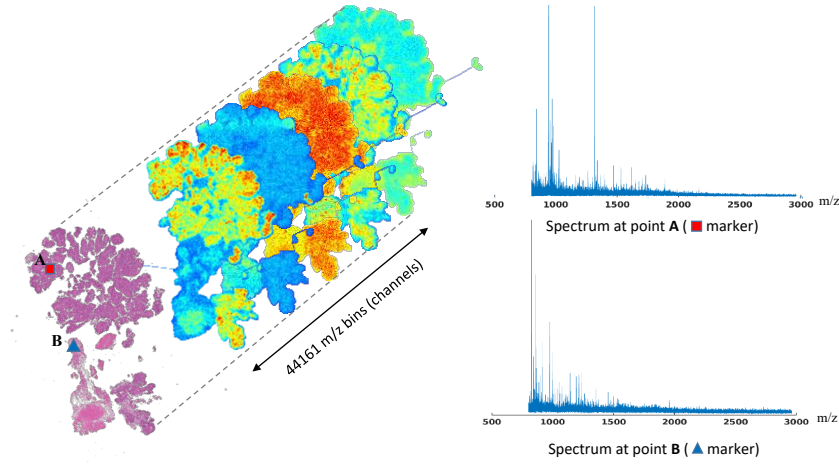
---

The MALDI-MSI method is able to create ions from large biomolecules like proteins and peptides. It has been introduced as an ideal method to combine classical tissue proteomics and histology [3]. MALDI-MSI has the potential to complement histopathological evaluation for confirming diagnosis and aiding in therapeutic management [3]. MALDI-MSI can be performed on fresh or frozen tissues, as well as on formalin-fixed paraffin-embedded (FFPE) tissues [4]. Due to its applicability to FFPE tissue samples, it has a high potential for applications in the histopathological domain [5] and in particular for supporting the diagnosis of tumor typing and sub-typing. The MALDI-MSI method creates a 3D tensor that can be represented as a 2D image with several thousands of channels. These channels contain the mass spectrum of the scanned tissue at each pixel location. Such rich and complex data has a high potential for addressing different diagnostic and prognostic clinical studies, like detecting tumorous and healthy regions, classifying cancer sub-types, grading metastases, etc. However, the analysis and interpretation of MS data require specialized methods.

Bright-field microscopy is a common practice in clinical pathology. However, MSI as new emerging technology has several advantages over the conventional histopathology by bright-field microscopy [6]. Unfortunately, an important hurdle in utilizing MSI in clinical practice is the difficulty of the visual interpretation of its multi-dimensional data. Whereas the diagnosis of diseases by visual inspection of stained histopathological slides (e.g. by pathologists) is straightforward, this is not the case for the MSI modality. A one-dimensional plot of MS data for a single point/pixel can be comprehensible for an expert who is knowledgeable about the association of each  $m/z$  bins and a chemical structure, but diagnosis through a 3D tensor of data is problematic, because it requires considering the spatial dependencies between local regions, which is not a straightforward task for a human. This has encouraged to develop new CAD systems for MSI data. Figure 2.1 shows an example of both a histopathological H&E slide and the corresponding MSI from a slice of bladder tissue. While the RGB color, shape, and appearance of cells in H&E slides provide a good visual indication of existing cancerous cells in each region, analysis of the MS data is more complicated.

Mass spectrometry has opened a new window in the clinical laboratory with its remarkable sensitivity and specificity, which is typically not produced by other analytical techniques [7]. However, MS is still underutilized in various clinical settings, mainly because of its analysis complexity and human interpretation. Continuous advancements in making MS more robust with various types of automation and robotics also warrant its high-throughput performance with an affordable cost-to-benefit balance. One major use case of this technique is cancer diagnosis at early stages.

To employ the MSI modality in clinical practice, several challenges have to be addressed in the interpretation of the measured signal by a computational model. The complexity and high dimensionality of MSI data impose some constraints on the choice of the computation technique. Recent advances in machine learning with high-dimensional data such as images have achieved promising results in



**Figure 2.1** Comparing the amount of data and complexity between a histopathology H&E slide and MSI data (left). An example of two spectra at two pixels on the bladder tissue slice is shown (right). In this example, Point A marks urothelial cell carcinoma and Point B marks the healthy detrusor muscle tissue.

high-level tasks such as classification and regression. In this chapter, we discuss and investigate the capability of deep learning models for analysis of MSI data for cancer detection and classification of histopathology data. More specifically, the research of this chapter addresses the following issues.

- *Architecture search*: Because of high complexity and dimensionality of MSI data, the choice of the network architecture is crucial for finding an efficient and elegant solution.
- *Data size*: Since the acquisition of MSI data takes considerable time and involves a high data amount, the amount of obtained samples is rather limited. This makes the generalization of the deep learning model difficult because of the limited amount of data samples.
- *Limited comparison methods*: Deep learning with MSI data is rather novel, so that comparison with other solutions is hardly possible.
- *Annotated data*: As mentioned earlier, the MSI data cannot be interpreted by pathologists. For annotating the MSI data, a second modality (e.g. bright-field microscopy) should be used for annotating the region-of-interest (ROI). This introduces noise in the spatial accuracy of the labels, since both tissue slices are not ideally aligned in the spatial domain.

This chapter is discussing above points in the following way. The work commences with choosing a baseline architecture that is extended based on the characteristics of the MSI data. This leads to new improved deep learning

architectures. Due to the small data size and the over-fitting problem, the architecture is constrained to be efficient in terms of the number of learning parameters. The last two points cannot be effectively addressed in this study because they are more fundamental and a consequence of being the one of the first to explore such a complex data.

This chapter is organized as follows. Section 2.2 discusses related work on MSI data and its application to cancer detection and classification. A second part of related work dedicated to the deep learning and associated network architectures. Section 2.3 presents two developed methods for cancer detection in pathology. Section 2.4 presents the results for various architecture choices and compares the improved architecture with the existing methods. Finally, Section 2.5 discusses the results and concludes the chapter.

### 2.2 Related work

The existing methods for the analysis of MS data can be partitioned roughly into two groups: (1) conventional digital signal processing and pattern recognition techniques that may be leveraged by using machine learning classification or regression methods, (2) advanced deep learning models that provide an end-to-end learning framework with potentially a higher performance. In the following, we briefly look into these two approaches.

#### 2.2.1 Conventional machine learning approach

Considering the large size of a typical MALDI-MSI dataset, extracting meaningful and interpretable information for a given clinical task is challenging [8]. A straightforward approach for analysis of mass spectrum data is peak detection [9]. Based on selecting a set of peaks that are statistically related to the target attributes, like the respective tissue characteristics or a set of relevant biomarkers that are hypothesized, the MS data can be analyzed. Such a feature selection approach requires prior knowledge to limit the set of features to a biologically meaningful subset [10]. For automating the feature extraction from the MSI data, several conventional signal processing techniques have been studied. Example techniques include exploring statistical correlation by non-negative matrix factorization [11, 10], tensor rank decomposition [12], k-means clustering [13], t-SNE dimensionality reduction [14], probabilistic mixture models [15], hierarchical clustering combined with PCA [16, 17], and using conventional machine learning classifiers like random forest [18]. Since the aforementioned techniques based on extracting a set of handcrafted features differ in their descriptive properties, employing each of them obtains a different performance in MS analysis. Usually, the use of the *explicit* features (supported by a mathematical model) through these methods limits their performance.

### 2.2.2 Deep learning approach

As an alternative, deep learning models are able to extract a set of *implicit* (lack of direct interpretation) and task-related features, based on the objective function of the method. By providing a sufficient amount of data with an acceptable rate of noise in data labels and considering diversity in training samples, the learning models have shown to outperform the conventional methods in many different domains and particularly in computer vision. It is worth mentioning that the impact of data availability and diversity on the performance can be influenced by the design and configuration of a learning model.

For addressing the shortcomings of conventional feature extraction and selection methods, a few studies recently investigated deep learning models for MS data analysis [19, 20, 21, 22]. For example, in the work of Spencer et al. [19], auto-encoders were used to reduce the dimensionality of the MSI data for identifying only the most indicative variables. Inglese et al. [20] used a parametric t-SNE model [23] for clustering of the MSI data. The parametric t-SNE is a non-linear dimensionality reduction method, which utilizes the flexibility of deep neural networks and the capability of similarity measure of the t-SNE distribution. The authors showed that this model can retrieve the local structure of high-dimensional MS data for visualization. Deep learning techniques for cancer detection and classification in MS data have not been explored properly and have been reported in a few studies only [22, 21, 24]. In the work of Behrmann et al. [22], a ConvNet was used with one-dimensional convolutional kernels which gained state-of-the-art performance on tumor classification. Zhang et al. [21] reported using a Recurrent Neural Network (RNN) for MS analysis. Tran et al. [24] used a combination of ConvNet and RNN for protein characterization in proteomics research. In line with these research works, we first explain briefly the one-dimensional ConvNets and the RNN and then we propose the contributions by using these models for cancer detection and classification in MS data.

#### A. One-dimensional ConvNets

The ConvNet was originally introduced with 2D convolutional kernels in the seminal work of Yann LeCun in 1990 [25] who formulated the backpropagation algorithm to train the first ConvNet, the so-called *LeNet*. By exploiting the graphical processing units (GPU) and training the 2D ConvNets on very large datasets, the popularity of the deep ConvNets peaked and eventually they became the *de-facto* standard for various ML and computer vision applications over the years. However, the conventional ML techniques were still unchallenged and performed more efficiently. Although the 2D ConvNets have been frequently used in processing sequential data including natural language processing (NLP), speech recognition, and biomedical time-series signals, the direct utilization of deep 2D ConvNets for a 1D signal processing application naturally needs a proper 1D-to-2D conversion of the input domain. Such a conversion is usually computationally expensive, like computing a 2D spectrogram from wave sounds. Apart from the conversion operations, the difference in computa-

tional complexity between 1D and 2D convolutions is significant. For example, the 2D convolution for a 2D matrix of  $N \times N$  dimensions and a kernel of size  $K \times K$  is in the order of  $O(N^2 \cdot K^2)$ , while in the corresponding 1D convolution on a vector of size  $N$  is  $O(N \cdot K)$ . This means that converting a 1D signal into a 2D matrix for employing a 2D ConvNet is rather costly.

To address such drawbacks, in 2015 the first compact and adaptive 1D ConvNet was presented for early arithmetic detection in 1D ECG signals [26]. Afterwards, 1D ConvNets have gradually become popular with state-of-the-art performance in many signal processing applications, like speech recognition [27], etc. In the aforementioned recent studies, compact 1D ConvNets have demonstrated superior performance for these applications, which have limited labeled data and high signal variations [28].

For the first time in the work of Thomas et al. [19], deep learning was introduced for MSI analysis, but mainly focused on unsupervised dimensionality reduction by using an auto-encoder. Later, dimensionality reduction by using a neural network was employed for finding the metabolic regions within tumors [20]. The first supervised deep learning model for tumor classification in MS data was reported in the work of Behrmann et al. [22]. This work has inspired us to study in depth because it at the time the only work based on ConvNet that has been applied to MS data analysis. Therefore, it is discussed more in depth in the following text.

The authors presented a 1D ConvNet called *IsotopeNet*, that consisted of four residual layers, one locally-connected layer and a fully-connected layer at the end. The performance of the *IsotopeNet* was reported on two datasets. The first task included a binary classification of two sub-types of lung tumor, namely adenocarcinoma versus squamous cell carcinoma (called ADSQ task). The second task was a binary classification of primary tumor types of lung versus pancreas tumor (called LP task). The *IsotopeNet* with around 14,000 parameters and 10-layer depth achieved 84.5% and 95.0% on LP and ADSQ tasks, respectively. This is about 3-4% higher accuracy than a ResNet-34 architecture with more than 2M parameters. Apart from the fact that a smaller network is less prone to overfitting, the *IsotopeNet* was specialized for MS data, by considering the kernel size, the locally-connected layer, and incorporating domain knowledge about the MS data. The author estimated the number of m/z bins of large measurable isotope patterns of peptides, based on the average amino acid. This prior knowledge about the length of the pattern in input space was imposed by adjusting the size of the receptive field of the 1D convolution kernel to be roughly equal to the size of the large isotope patterns, such that one variable can encode such a local feature. Furthermore, the author argued that using a locally-connected layer after the last convolutional layer enables the network to handle each local region differently, due to unshared weights and this led to focusing only on important peptides for the given classification task. Later, we will consider this model as a baseline and specifically study the impact of these two previous adjustments to the model architecture on the tumor classification task in MS data.

### B. Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are a family of ANNs, initially proposed for learning temporal dynamics in sequential data, containing unknown dependencies between the elements of a sequence. The fundamental research on RNN took place in the 1980s. Hopfield in 1982 introduced a family of RNNs in which all connections across layers were equally sized. In the Hopfield network [29], a pattern is placed in the network by setting the values of the nodes. Afterwards, by executing the network on a given input according to its update rules (which is called Hebbian learning), eventually, another pattern is read out. The Hopfield network is useful for recovering a stored pattern from a corrupted version. Thus it does not process sequences of patterns and is not a general RNN. However, the convergence of the network is guaranteed.

An early architecture for supervised learning on sequences was the *Jordan* network that was introduced by Jordan in 1997 [30]. The Jordan network consisted of a three-layer (i.e. one hidden) feed-forward network with the addition of a set of context units. Each context unit was connected to the output and hidden layer through backward and forward connections, respectively. The fixed backward connections saved a copy of the previous values of the hidden units in the context units. Thus the network could maintain a sort of state, allowing it to perform sequence prediction that was not possible with a standard MLP network. Later in 1990, a similar architecture was introduced by Elman, called *Elman* network [31], that was simpler than the Jordan network. In the Elman network, each context unit is only connected to a hidden node. Each such unit took a copy of hidden values from the previous temporal sample with unity weight. This value then was fed back into the same hidden node along with a learnable weight. This architecture is equivalent to a simple RNN in which each hidden node has a single self-connected recurrent edge. The idea of fixed-weight recurrent weight that makes hidden nodes self-connected, was adopted in subsequent work on long-short term memory (LSTM) networks [32].

The LSTM networks are a subset of RNNs and address the problem of exploding and vanishing gradients in conventional RNNs [33]. LSTM networks are especially capable of handling the learning of long-term dependencies. The recurrent properties of RNNs have proven to be useful in multi-dimensional sequence processing and irregular pattern extraction and are able to learn spatial dependencies in signals such as MSI data [34]. The use of LSTM networks can further increase the performance of present gold-standard cancer-identification methods [35, 24].

#### 2.2.3 Contributions based on the related work

To obtain a higher performance, based on some evaluation metrics that will be introduced later, we investigate two types of deep learning models: 1D ConvNets and RNNs for their competitive performance on MS data analysis. More specifically, our contributions concentrate on the following aspects.



- It will be shown that the receptive field of a 1D ConvNet has an important impact on the processing of the MS data. We consider that MS data are different from natural images, so that the low-level features can spread over the entire spectrum. For computational efficiency, instead of increasing the size of kernels, we propose using dilated convolutional kernels in the layers of a 1D ConvNet. We show that setting a different rate of dilation in network layers can lead to extracting more relevant features and thereby result in a higher performance.
- Several architectures of the RNN model with LSTM modules are explored. We report the performance on cancer classification with respect to the variations in architecture and eventually, we propose a network that performs state-of-the-art classification on two different cancer datasets.

### 2.3 Methodology

#### 2.3.1 MS analysis using 1D ConvNets

The computational workflow of a deep ConvNet on high-dimensional data such as images is mainly based on two mechanisms. The first mechanism is learning the parameters of a set of convolutional filters as basis functions. The second mechanism is the hierarchical architecture of the network that imposes different levels of abstraction to the input data. Although these two mechanisms in a ConvNet have been known to be very effective for many different data modalities, including medical images, their applicability to MS data can be justified by considering two properties of the MS data. First, the neighboring  $m/z$  bins in a spectrum are correlated based on their time-of-flight (ToF), where the spectrum peaks are spread over several  $m/z$  bins. Since high correlations between neighboring data points exist, the convolutional operation can be an effective mechanism for exploiting such correlations. Second, the combination of the appearing peaks in the spectrum, which indicate a group of ions, can represent isotope patterns. At a higher level, the combination of several isotopes represent tryptic-digested proteins that may contribute to various measured peptides, resulting in specific patterns across the entire mass range [22]. Exploiting existing hierarchical context in data is well-suited for a deep ConvNet, consisting of several hierarchical layers that can learn different levels of abstraction from the input data. Therefore, we apply a ConvNet to the analysis of MS data because of its architectural fit to the structure of the data. Besides this choice, we know that a ConvNet has high performance in other domains as well.

Although the aforementioned characteristics of MS data encourage the use of ConvNet, the high dimensionality of the signal and its sparsity should be handled carefully for achieving higher performance in a CAD system. The number of mass bins in a typical mass spectrum is in order of tens of thousands and the chemical patterns may spread over the entire range of the spectrum. Learning useful kernel parameters require the network to be either very deep, or have a set of large kernels resulting in a large number of parameters and therefore a

high computational complexity. This can be problematic because in pathology, providing labeled MS data at a large scale is cumbersome and training a deep ConvNet under a small-scale data regime is prone to a higher risk of over-fitting.

As mentioned earlier, the IsotopeNet as a state-of-the-art model in cancer classification has a relatively low number of training parameters and its convolutional filter size has been tuned with respect to the size of the largest isotope in data. Although designing an efficient network with a low number of parameters is appreciated, adjusting the size of the convolutional kernel imposes a bias on the model which would be different for data with a different resolution of  $m/z$  bins or for a different target task. This requires prior knowledge about the size of relevant features in data that may hinder the generalization of the model for a different type of MS data or task. Furthermore, in MS data both the mid-level and high-level contexts/patterns may spread over the entire spectral range, so that a standard ConvNet with *shared* convolutional filters with limited receptive fields cannot learn them. To overcome these issues, we study the use of *dilated* convolutional kernels in a 1D ConvNet for MS analysis, by ignoring the type of task and existence of some specific biochemical structure (i.e. isotope) in given input data.

### Dilated convolutional kernels

Dilated convolutions [36] offer a method of increasing the receptive field without losing input and mid-level feature resolution. The concept of a dilated ConvNet was shown to be successful in several domains, such as 2D image analysis [36, 37, 38] and natural language processing [39]. The discrete convolution operator  $\otimes$  and the dilated convolution between feature vector  $F$  and kernel  $k$  are defined as

$$\begin{aligned} (F \otimes k)(p) &= \sum_{s+t=p} F(s)k(t), \\ (F_l \otimes k)(p) &= \sum_{s+lt=p} F(s)k(t). \end{aligned} \quad (2.1)$$

The standard discrete convolution ( $\otimes$ ) is simply the dilated convolution with dilation factor equal to  $l = 1$ . The dilated convolution in prior works is referred to as convolution with a dilated filter that has been used in wavelet decomposition [36]. The dilated convolution with the same number of parameters as a standard convolution can be applied to the input domain with a different receptive field by employing different dilation factors. Therefore, the receptive field of each dilated convolution kernel can be adjusted to learn patterns that spread over a larger space in the input domain to each layer of the ConvNet. Furthermore, using several dilated convolution filters with various dilation factors in each layer of a network can perform multi-scale context aggregation, since the receptive field of each kernel can learn a filter that is efficient for a different scale of the data context.

Here, we investigate the use of dilated filters in the 1D ConvNet for MS data classification. The IsotopeNet is used as the baseline architecture by replacing its convolutional filters with dilated ones. For learning features at different contextual scales in the first few layers of the network, we gradually increase the

dilation factors and in the deeper layers the dilation factors are decreased. By applying such a variation in the receptive fields of layers, in the first few layers of the network the hierarchical contexts of data are learned and in the second half of layers, aggregations of learned features are performed. This configuration is very effective in the processing of satellite imagery where segmentation of small objects in a very large image is important [37]. In the experiment, we show the effectiveness of employing dilated filters in cancer detection and classification.

### 2.3.2 MS analysis using RNN

An alternative for learning the patterns that are spread over a long range of the mass spectrum is by considering the long-range dependencies among mass bins. Modeling such dependencies can be implemented by using RNNs. For studying this hypothesis, a variation of the LSTM architecture proposed by Zhang et al. [21] is considered as the baseline LSTM model. This model consists of a single LSTM layer with 500 LSTM units, followed by a dense layer with two output nodes and the rectified linear unit (ReLU) activation function. The input to the network is all mass spectra of a sample (e.g. a vector with 27,286 elements) and the output is the binary label representing the input class. For achieving better performance on MS data classification, several variants to the baseline model are investigated and their impacts on the classification performance are evaluated as follows.

- The number of units in the LSTM layer is varied from 10 up to 1,000. The addition of more LSTM units increases the dimensionality of the latent space and leads to an increased learning capacity of the network.
- The number of layers (depth) of the model is varied between 1-4 layers deep. Exploiting more layers helps the network learn higher-order dependencies and consequently capture more complex patterns in a hierarchical structure.
- The network can be with and without bidirectional architecture. For merging the forward and backward networks, averaging, concatenation, and summation operators are used.
- Batch normalization and dropout, on the weights of the input layer, are introduced. Applying dropout can decrease over-fitting whilst batch normalization can speed up the training time [40, 41, 42].
- The L1 regularization, or Lasso regression, is used for dimensionality selection due to the sparsity of the MS sequences, in order to further increase the performance of the model [43].
- Data augmentation is proposed for increasing the variation in the data and improving the generalization power of the model. However, applying it to the MSI data has no established precedent and is not straightforward. A

variation of the PCA data augmentation method called *Fancy-PCA* is employed, which is adjusted from the work of Krizhevsky et al. [44]. To do so, first, the covariance matrix of training data is computed. Afterwards, by applying the Singular Value Decomposition, Eigenvalues are augmented by a random factor in the range of  $[1.0, 1.7]$  and by backward transformation the new samples are generated.

Combinations of all above considerations are proposed for improving the baseline model and achieving a higher performance for MSI data classification. Moreover, the performance of the LSTM approach is compared with the PCA-LDA and IsotopeNet models. With experiments, it is shown that LSTM networks can learn the local and non-local patterns in MS data and outperform the two state-of-the-art methods in mass spectra classification.

## 2.4 Experimental results

### 2.4.1 Dataset

Two datasets are used to evaluate the performance of the proposed models for MS data analysis. The first publicly available dataset is used in the work of Boskamp et al. [10] and Behrmann et al. [22]. This dataset includes the MS data from two types of lung cancer: carcinoma and squamous cell carcinoma. For better comparison of the acquired results with prior works, the same data partitioning for training and testing is used. The data is divided into 8 sets and are used with a fourfold cross-validation scheme. The MS data consists of a total of 4,672 samples from 8 patients.

The second dataset consists of the MS of bladder tissue from 9 patients. The tissues were serially sectioned by the Academic Medical Center (AMC) in Amsterdam. Half of the provided sections (e.g. all odd-numbered sections) are used for hematoxylin and eosin (H&E) staining and bright-field microscopy. The histological annotations were provided by the uropathologist and were used as the ground truth for the other half of the sections (all even-numbered sections). Alternatively, the last is used for MS imaging at Maastricht University (Maastricht, The Netherlands). The MS data were obtained at a spatial resolution of  $50\ \mu\text{m}$  with a RapifleX Time-of-Flight mass spectrometer across the mass-to-charge-ratio ( $m/z$ ) range of 800-3,000 Da (see Figure 2.2). The resolution of MS data is 0.0498 Da, resulting in 44,161 bins over the whole range. A subset of 13,000 samples are used, all representing either urothelial cell carcinoma or healthy urothelium, and healthy detrusor muscle tissue. The samples are distributed evenly over those two classes. The dataset is divided into a threefold cross-validation scheme at the patient level. Here, the task is the classification of mass spectra belonging to tumorous and muscular tissue.



**Figure 2.2** RapifleX time-of-flight MALDI-MSI equipment. The tower of ion detector/reflector with a 300-cm effective flight path (left photo). 3D laser-beam generator (right photo).

### 2.4.2 Evaluation metrics

The performance of the binary classification is measured with the balanced accuracy (BA) metric, calculated by  $\frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$ , where  $P$  and  $N$ ,  $TP$ , and  $TN$  are the numbers of positive, negative, true positives, and true negatives predictions, respectively. The BA measure combines sensitivity and specificity and takes class proportions into account that is more indicative than accuracy, since it considers the possible effect of an imbalanced dataset [45]. We also report the  $F_1$  score, which takes recall and precision into account. Additionally, the interquartile ranges (IR) are calculated to measure the robustness of the classifier with respect to different datasets and random elements in the case of neural networks. Lastly, the area under the Receiver Operating Curve (AUC) is reported for completeness of the evaluation.

### 2.4.3 Results

#### A. Performance Evaluation of 1D ConvNet

In a preliminary experiment, the effect of the locally-connected layer in the IsotopeNet architecture is tested. Since a locally-connected layer is a convolutional layer with unshared weights, the number of parameters grows with the input spectra length. At the input of the locally-connected layer in the IsotopeNet architecture, the signal has a length of 1,820 bins. Using a kernel size of five and a bias vector, the number of trainable parameters in the locally-connected layer becomes equal to 10,920. This is about 70% of the total 15,806 trainable parameters in the model. The original IsotopeNet performance is compared to a slightly changed version, in which the locally-connected layer is replaced by a regular convolutional layer with a kernel size of five. This results in a network with only

**Table 2.1** Influence on the performance of IsotopeNet when using either a locally-connected layer or a convolutional layer.

	Locally-connected layer	Convolutional layer
No. of parameters	15,806	4,891
Training time [min.]	22.85	19.55
Execution time [sec.]	2.73	0.11
Balanced accuracy	0.83	0.82

**Table 2.2** Training configurations for two evaluated datasets

	Lung data	Bladder data
Dropout probability	0.30	0.40
Weight decay	0.05	0.09
Learning rate*	0.05	0.0005

\* Adam optimization[46]

4,891 trainable parameters. This preliminary test has been performed only on the lung dataset with the cross-validation scheme as mentioned above. Three performance indicators are measured: training time for 300 epochs, the execution time for processing one sample of data, and the balanced accuracy. The results of the preliminary test to investigate the effect of the locally-connected layer are shown in Table 2.1.

To investigate the effect of the receptive fields, we use IsotopeNet as a baseline. All convolutional kernels in the residual blocks of IsotopeNet were replaced by dilated kernels, leaving the rest of the architecture unchanged. In our experiments, the training configurations of IsotopeNet were changed according to Table 2.2.

To find the optimal receptive field size, different dilation rates are tested on a datafold of the lung dataset. Ten percent of the training data is separated as a validation set. The performance is measured after 500 epochs of training. As mentioned earlier, the dilation rates are set by first increasing and then decreasing the dilation rate in deeper layers. A step-wise increase of the dilation rate results in a larger receptive field, necessary to capture large patterns. Moreover, a step-wise decrease of the dilation rate in the later layers of the network ensures that mid-level feature resolution, which is lost due to the spaced kernels, is recovered. To reduce the effect of randomness, all networks use the same weight initialization. Weights are initialized according to the method of He et al. [47] for this test and all other training instances. The two best-performing networks from this preliminary test are employed for extensive testing on all datafolds of the lung dataset. For the bladder data, only the best performing dilated CNN is trained and its performance is compared to the two previously mentioned benchmarks. Table 2.3 shows the results of the preliminary test for finding the best dilation rate.

## 2. MASS SPECTROMETRY ANALYSIS

**Table 2.3** Balanced accuracy (BA) for varying dilation rates and their corresponding effective receptive fields (ERFs). Best scores are printed in bold.

ERF	200	273	447	765	1039	2275	6115
BA	0.928	<b>0.932</b>	0.928	0.928	0.923	<b>0.943</b>	0.931

**Table 2.4** Architectural specification of several layers of the dilated networks

Layer	Kernel	Stride	Feature maps	Dilation rate (Net. 1)	Dilation rate (Net. 2)
Residual block 1	3	1	8	1/3	1/3
Residual block 2	3	5	8	9/27	9/27
Residual block 3	3	1	8	16/5	16/48
Residual block 4	3	3	1	2/1	144/48
Locally-connected layer	5	1	1	1	1
Dropout layer	-	-	-	-	-
Fully-connected layer	-	-	1	-	-

The two best-performing networks have an ERF size of 273 m/z bins (in the following referred to as DilationNet1) with a balanced accuracy of 0.932 and an ERF size of 2,275 m/z bins (in the following referred to as DilationNet2) with a balanced accuracy of 0.943. It is worth mentioning that the reported accuracy is calculated on the validation set. Since samples from the same patient are highly correlated and such samples are present in both training and validation partitions, the performance of the validation set will be significantly better than the performance of the test set. However, here we only consider the relative model performance for finding the best ERF. In comparison, IsotopeNet has an ERF of 81 m/z bins.

The general structure of the models DilationNet1 and DilationNet2 can be observed in Table 2.4. It should be noted that the indicated dilation rates in the Table are applied to the first and second convolutional layers within the actual residual blocks.

For the second dataset, several adjustments to the network structure have been made. The bladder dataset has roughly a three times higher resolution. With increasing resolution, the m/z bin size decreases and the receptive field captures smaller parts of the mass spectrum. IsotopeNet was designed with an ERF size of 13.02 Da [22]. To preserve the receptive field size for the bladder dataset, the signal is down-sampled in the third residual block with a factor of three using a strided convolution. This results in an ERF size of 11.11 Da. Furthermore, the hyperparameters have to be adapted to the bladder dataset. The weight decay factor and the dropout probability are found with a grid search on one datafold by evaluating balanced accuracy of the validation set. The learning rate is determined following the method of Smith [48], in which the learning rate is gradually increased after each batch until the loss no longer converges. The network is then reset and trained with the found learning rate.

**Table 2.5** Performance of the proposed DilationNets in comparison with the two baseline methods (BA=Balanced Accuracy, IR=interquartile Range). The best values are printed in bold.

Classifier	Lung dataset		Bladder dataset				
	Median BA	IR	Median BA	BA IR	Median $F_1$	$F_1$ IR	AUC
PCA-LDA	0.8176	0.0631	0.7734	<b>0.0718</b>	0.7157	<b>0.0975</b>	0.8890
IsotopeNet	0.8275	<b>0.0408</b>	0.8115	0.1634	0.8339	0.1948	0.8928
DilationNet1	<b>0.8374</b>	0.0423	<b>0.8328</b>	0.1349	<b>0.8666</b>	0.1838	<b>0.9235</b>
DilationNet2	0.8173	0.0425	-	-	-	-	-

The training is aborted if the average loss of the last 25 epochs drops below a predefined threshold ( $5 \times 10^{-2}$ ). Although regularization techniques are exploited, over-fitting cannot be avoided for the bladder dataset. Instead, every 10 epochs the balanced accuracy of the validation set is calculated. If the balanced accuracy reaches a local minimum and starts growing again for the next three validation set evaluations –i.e. 30 epochs– the weights of the minimum are loaded and the test set performance is evaluated.

The classification performance for the lung dataset is evaluated with a four-fold cross-validation scheme, where every datafold is evaluated four times, resulting in a total of 16 evaluations. For the bladder dataset, classifier performance is evaluated with a twofold cross-validation scheme, where every datafold is evaluated eight times to reduce the impact of the random elements, like the dropout layer and weight initialization.

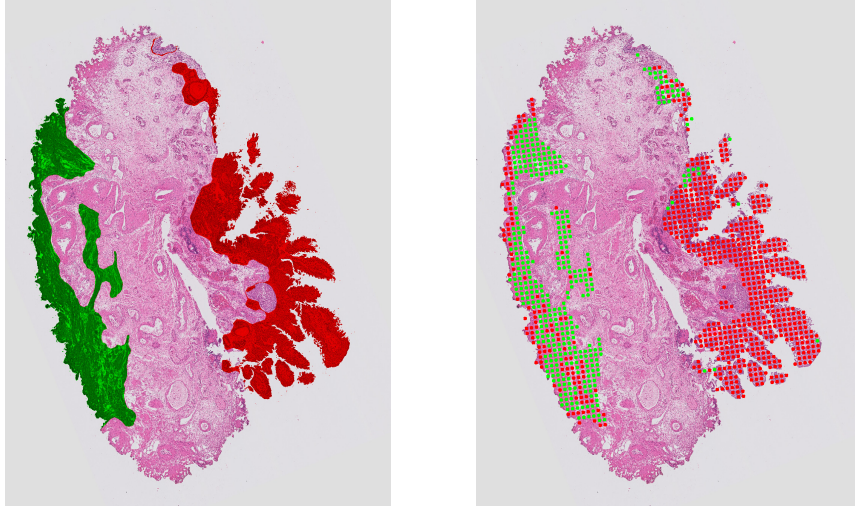
The results for both datasets are summarized in Table 2.5. For the lung dataset, the highest median balanced accuracy of 0.8374 is reached by the DilationNet1. The interquartile ranges (IRs) of all networks are very similar with the lowest IR of 0.0408, achieved by IsotopeNet. The PCA-LDA method shows the lowest performance on both datasets.

For the bladder dataset, the DilationNet achieves the highest median BA (0.833),  $F_1$  score (0.866) and AUC (0.923). The lowest IR is achieved by the PCA-LDA classifier, followed by DilationNet1. The ROC curves of all three classifiers on the bladder dataset are depicted in Figure 2.4. Figure 2.3 visualizes an example of the predicted carcinoma by DilationNet1 on a bladder tissue slice.

## B. Performance Evaluation for the RNN model

The best RNN architecture is explored such that the maximum classification performance of the MS data is obtained. Several configurations are examined such as increasing complexity (e.g. by adding hidden layers), implementing a bidirectional architecture, applying Lasso regression (or  $L_1$  regularization) and augmenting the MS data. For finding the best network architecture, all experiments are performed on the lung cancer dataset only. Thereafter, the best model is selected and its performance is evaluated on the Bladder dataset. This data split is adopted to ensure that the architectural search is not biased towards the obtained performance on the training data.





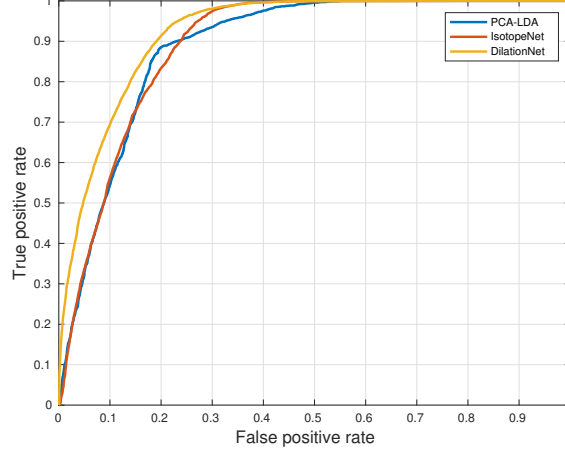
**Figure 2.3** Visualized bladder slice example of carcinoma detection. Left: slice with urothelial cell carcinoma (red) and healthy detrusor muscle tissue (green). Right: model prediction with sampled (laser) spots at tissue surface.

Furthermore, a comparison is made between the LSTM approach and the two state-of-the-art methods (i.e. PCA-LDA and IsotopeNet). In addition to the balanced accuracy scores and as complementary metrics, we report the  $F_1$  score and AUC of the best architecture.

Table 2.6 shows the performance of the baseline RNN architecture along with several different architecture choices and their corresponding performance on the Lung cancer dataset. For training all variants of the model, the following constant configuration is applied: a batch size of 32, a learning rate of  $1 \times 10^{-4}$ , a maximum number of 100 epochs, the RMSprop optimizer, binary cross-entropy loss, and the use of HE parameter-initialization [47].

The following observations are made on the experimental results.

- *Increasing the number of LSTM units* in a single-layer network leads to a higher classification performance. However, using more than 50 units shows no noticeable improvement. For example, employing 1,000 LSTM units shows the best performance: the balanced accuracy is improved only about  $7 \times 10^{-3}$  in comparison with 500 units.
- *Increasing the depth of network* by adding more hidden LSTM layers, does not increase the accuracy further after 2 layers. It seems that two layers of LSTM can sufficiently model the dependencies across latent feature dimension.
- *The bidirectional architecture* by an ensemble of two networks (forward and backward) improves the performance as expected by 1% (for averaging merging).



**Figure 2.4** The ROC curves of PCA-LDA, IsotopeNet, and DilationNet.

- *Batch normalization* drops the accuracy by 1.2%.
- *Dropout* technique marginally improves the accuracy by 0.9%.
- *Data augmentation* does not enhance the performance.

Table 2.7 shows the best choice of the key aspects of the conducted architecture search that is summarized in Table 2.6. The proposed architecture consists of two bidirectional LSTM networks, with 100 LSTM units in its hidden layers, and the use of  $L_1$  regularization. The dense layer maps the input mass spectra into a 100-dimensional space on which the LSTM units are applied. The  $L_1$  regularization is applied to the weights of the dense layer, to enforce some weights towards zero. The bidirectional network uses average merging to combine the output of the LSTM layers in forward and backward directions. Finally, a dense output layer, followed by a softmax activation layer, maps this embedded space into the output labels. Dropout does not contribute to the final architecture because it does not improve the performance on the Lung dataset. However, when the model is trained on the Bladder dataset, the dropout is added. This is necessary to avoid over-fitting on the bladder mass spectra because of its higher dimensionality, which is equal to 44,161 (about 38.2% higher than the Lung dataset). Figure 2.5 illustrates a visual example of the performance of the RNN architecture on the Bladder dataset.

Table 2.9 presents a comparison of three methods in terms of balanced accuracy, AUC, and  $F_1$  score results. The PCA-LDA has been implemented according to the work of Boskamp et al. [10] for up to 100 principal components. The RNN and IsotopeNet methods use an inter-quartile range for the dispersion of the results. It can be observed that the RNN model achieves the best scores among the three methods.

**Table 2.6** RNN architecture sample search on the Lung cancer dataset. The optimal architecture is based on different factors from which a combination is chosen.

Exploration parameters deviating from the base architecture	Balanced accuracy
(Baseline) one layer, 100 LSTM units	0.8278
10 LSTM units	0.8044
20 LSTM units	0.8215
50 LSTM units	0.8296
200 LSTM units	0.8304
500 LSTM units	0.8308
1000 LSTM units	0.8381
Two layers	0.8295
Three layers	0.8281
Four layers	0.8287
Bidirectional, concatenated merging	0.8301
Bidirectional, average merging	0.8359
Bidirectional, summing merging	0.8289
Two Layers, Batch normalization	0.8180
Two Layers, Dropout (0.5 dropout rate)	0.8350
$L_1$ Regularization (before baseline)	0.8479
Data augmentation (x2 signal)	0.8282
Data augmentation (x3 signal)	0.8265
Data augmentation (x4 signal)	0.8278
Batch normalization and dropout	0.8302
$L_1$ Reg., batch norm. and dropout	0.8415
(2x) $L_1$ Reg., batch norm., dropout	0.8501

**Table 2.7** Specification of the proposed RNN LSTM architecture via architectural search

Layer	Shape
Input*	(-, 1, 27286)
Dense ( $L_1$ regularized)	(-, 1, 100)
Bidirectional (average merging)	(-, 1, 100)
Bidirectional (average merging)	(-, 100)
Dense (with Softmax activation)	(-, 2)
* with Bladder data, input shape	(-, 1, 44,161)

**Table 2.8** Comparison of ConvNet and RNN architectures by time and parameters

Method	Training time per epoch [s]	No. of param.
IsotopeNet	12.75	<b>13,935</b>
RNN	<b>2.00</b>	3,050,502

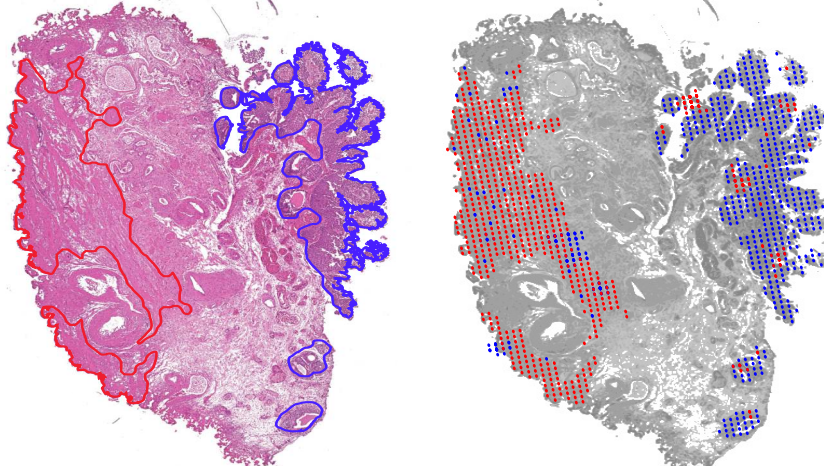
**Table 2.9** Performance comparison of three methods, in terms of balanced accuracy, AUC and  $F_1$  score results (best scores in bold).

Method	Lung cancer data	Bladder dataset
	Balanced Accuracy	
PCA-LDA	0.787	0.669
IsotopeNet	0.845 $\pm$ 0.007	0.812 $\pm$ 0.163
RNN	<b>0.864</b> $\pm$ 0.012	<b>0.826</b> $\pm$ 0.023
	$F_1$ score - AUC	
PCA-LDA	0.826 - 0.787	0.673 - 0.666
IsotopeNet	0.853 - 0.841	0.834 - <b>0.893</b>
RNN	<b>0.865 - 0.870</b>	<b>0.840</b> - 0.832

The RNN model obtains a 7.68% higher accuracy on the Lung dataset than the conventional PCA-LDA method [10]. It also outperforms the IsotopeNet, built upon the recent advances in ConvNets, by about 1.87% higher accuracy on the same dataset. On the Bladder dataset, the RNN performs 15.7% and 1.45% higher than the PCA-LDA and IsotopeNet, respectively.

Furthermore, Table 2.8 presents some implementation details and shows the measured training times of the RNN and the ConvNet architecture in TensorFlow on a GTX-1080 graphics card. Although the proposed RNN model has over 200 times more parameters (due to its large linear input layer) than the IsotopeNet, it trains 6 times faster. This is explained by the lack of convolution operators, as used by ConvNets, and by the simple forward and backward gradient propagation of the LSTM modules in the RNN model.

Adding more than 100 units to a single-layer RNN network does not further improve the results. This can be caused by saturated learning of the network, meaning that there are increasingly fewer dependencies to be learned. Similarly, increasing the depth without performance improvements can indicate that a single-layer network already has sufficient complexity for learning the task. The reason that data augmentation does not improve the performance may result from the aspect that the added variations to the data do not influence the critical input space close to the decision boundary between the two classes.



**Figure 2.5** Prediction of the proposed model on bladder tissue; (left) H&E slide with the overlaid ground truth; (right) model predictions at the laser spots of sampled MS data in MSI.

### 2.5 Discussion and Conclusions

A mass spectrum provides the molecular profile of an examined tissue, which represents a rich and informative measure that can be used for cancer diagnosis. In this chapter, we have studied two promising deep learning frameworks for cancer detection and classification by exploiting mass spectrometry data. The application of the deep learning frameworks to mass spectrometry imaging is a new emerging technology and has been insufficiently researched for employment in clinical practice. This new technology poses two important challenges. (1) It requires expensive and dedicated hardware, and (2) it is difficult to comprehend the measured signal by a pathologist. These two challenges hinder its application in clinical cancer diagnosis. Mass spectrometry imaging is different from well-established bright-field microscopy which presents each pixel of a pathological image with the RGB channels, while each pixel of an MS image presents a mass spectrum (i.e. histogram of masses). Understanding and interpreting a mass spectrum image with several thousands of channels is difficult for a pathologist and therefore requires developing appropriate computational algorithms for data abstraction or automated diagnosis.

*Dilated ConvNet:* Due to long-range dependencies between the elements of the mass spectrum and the existing hierarchical structure of semantics in data, we have based our study on two deep learning frameworks that can leverage such dependencies and hierarchical structuring for improving the cancer detection. Based on recent advances in deep learning models, we have shown that using a set of dilated convolution operators with various dilation factors learn long-range dependencies in the signal by capturing local and global features for

cancer detection. Stacking several neural layers in the network, each with a set of kernels with different dilation factors is able to learn the hierarchical structure within the data. Exploiting the dilated convolutions in the proposed DilationNet has resulted in capturing such patterns with different resolutions in the input space. DilationNet exploits the fact that the cancer mass signature spreads over large parts of the spectrum. The effective receptive field (ERF) of convolutional kernels can be increased by incorporating dilated convolutions, without growing the number of parameters and hence the computational complexity. The proposed network outperforms two state-of-the-art MSI cancer classifiers by 1-3% in balanced accuracy, thereby illustrating that ERF is an important hyperparameter, which needs careful adjustment.

*RNN framework:* As an alternative and second deep learning framework for cancer detection and classification on MSI data, the proposed RNN model consisting of 2 hidden layers in a bidirectional architecture, outperforms the recent advanced CNN approach in MS classification by a moderate 1.87% and 1.45% higher accuracy on two clinical datasets, but with an impressively 6 times faster training time. The efficiency of the CAD system is important when using such a novel technology. Similar to the first framework, employing a recurrent network is motivated by the need of incorporating both local and global dependencies between the elements of the mass spectrum. Capturing these dependencies implies the direction of the solution for learning cancerous fingerprints in data. This finding is in agreement with the presence of biological signatures for isotopes and proteins in mass spectra. The LSTM networks have proven to model well the long irregular dependencies in sequential data for learning the patterns captured by MS data. It is important to mention that the proposed model has similar performance results on both lung and bladder data (the latter was left out during the exploration phase), which indicates the proposed model has high generalization power and is not biased to a specific dataset.

*Limitations:* One of the limitations with our study that can be considered as future work is processing the MSI data in two dimensions by processing neighboring spectra located in a 2D region to improve the classification results. To this end, the combination of ConvNets and RNN can be an alternative approach to further increase the performance, since their combined properties can complement each other.

Another limitation of our work can be the manual architecture search for designing a high-performance RNN architecture. Although the proposed architecture search for different aspects of the RNN has been extensive in terms of experimental effort and the chosen properties are understandable, the best configuration is not necessarily optimal. This is mainly because of the limited computational budget, leading to the choice of the coordinate descent search for seeking the best configuration of each factor of freedom considered for the design of the proposed network. Recently, Neural Architecture Search (NAS) has been introduced as a technique for automating the design of neural networks, which has been shown to be comparable or even outperform the handcrafted

architectures[49].

This chapter has shown that deep learning-based CAD models can achieve state-of-the-art cancer detection and classification performance in mass spectrometry data. It is expected that by developing more accurate and reliable computation models, the clinical application of this new emerging technology can be facilitated. The next chapter concentrates on the existing technology of bright-field microscopy that is widely is being used in clinical practice for cancer diagnosis. In order to advance this technology, the study will address some existing challenges by using deep learning models.

## Two-dimensional Histopathological Image Analysis

### 3.1 Introduction

#### 3.1.1 Histopathology

Histopathology refers to examining the microscopic structure of tissue (histology) in order to investigate human diseases (pathology). The histopathology process starts with extracting a piece of human tissue via surgery, biopsy or autopsy. For preserving the morphology of cells for a long period of time, the tissue is stabilized by placing it in a fixative such as formalin, which is the most commonly applied fixation. Afterwards, the fixated tissue is cut into extremely thin slices, in order of a few micrometers, by using a microtome. The cells morphology in such a thin slice can be observed by transmitting light through the tissue. The thin tissue slice is stained and mounted on a glass slide prior to examining it under a microscope.

The staining procedure is used to highlight the morphological structure of the tissue. The most common staining procedure consists of two chemical components: hematoxylin and eosin (H&E). Hematoxylin stains cell nuclei in blue, while eosin stains the rest (cytoplasm, connective tissue, etc.) in multiple shades of pink. This staining is used as a reference for many critical examinations by the pathologists, like grading a tumor in breast cancer.

#### 3.1.2 Computational pathology

In conventional pathology, doctors examine tissue slices by placing a glass slide under a physical microscope. To investigate the entire surface of a specimen, the pathologist uses different levels of magnification between coarse, medium, and fine focus, to go from overview to fine details and acquire image data for



documentation. With recent advances in computer science and information technology in health care, the field of histopathology is undergoing a major digital revolution. New emerging pathology scanners digitize the process of imaging from the glass slide, so that the pathologists have access to the digital whole-slide images on computers in an automated process. Whole-slide imaging enables the field of pathology to employ recent advances in computer-assisted analysis to facilitate visual inspections, remote diagnosis, efficient archiving cases, and leveraging computer-aided diagnostics.

Computer-aided diagnostics (CAD) refer to a computerized procedure that assists doctors in the diagnosis and interpretation of medical images. The application of CAD in histopathology can be multifold. CAD can be employed for the detection and segmentation of important human tissue structures to reveal some abnormalities. CAD also can be used for grading the cancer such as Gleason scoring for prostate cancer or classification of cancer into different categories.

#### 3.1.3 Challenges in computational pathology

Designing a robust, reliable, and efficient CAD algorithm for the analysis of the histopathological images requires addressing several challenges. Some challenges are related to normalizing the data, whereas another challenge corresponds with automated analysis of the data. The third aspect of the challenges refer to the compression of large-size images. It is evident that these aspects are quite different from each other, although they are related to the same imagery.

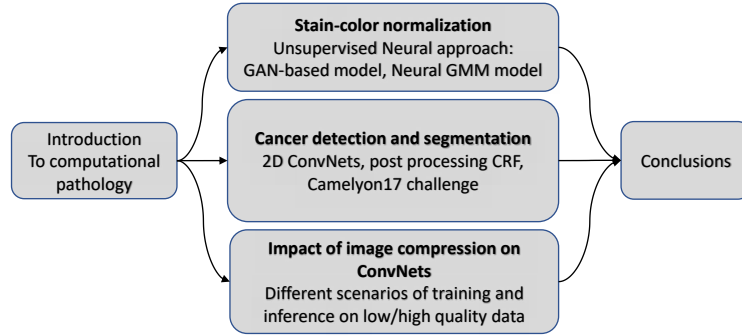
- **Stain-color normalization:** Stains of pathology slides are typically colored by using dyes for improved visualization to the pathologists. The difference of color between the regions of an examined tissue can also be used as an implicit feature in a learning-based CAD model. However, unseen color variations in image data degrades the performance of a data-driven algorithm. The causes of the color variations in stained slides can be very broad. For example, two different pathology labs may use slightly different amount of dyes, or the dyes are provided by a different manufacturer. Furthermore, scanning a tissue with two different scanners may obtain slightly different color images due to scanner-specific characteristics and the lack of a calibration process. Since the causes of the color variations are an imposed starting condition and cannot be simply controlled, the design of a robust CAD system to (un)seen variations in the color of images, stain color normalization is an important aspect for careful consideration and may need a specific preprocessing step. A straightforward approach would be stain color normalization, while such methods reduce uninformative variation in the color and intensity of pathological images. Hence, it is in our interest of research to explore a learning-based model for stain-color normalization for pathological images.
- **Automated analysis of histology images:** Detecting tumor cells in microscopic examination of stained histological slides can be a tedious and

time-consuming task for pathologists. This procedure is error-prone and depends on the qualification and skills of the individual person. As a core problem of computational pathology and to facilitate pathologists, this chapter aims at the design of a deep learning model for automated metastases detection and their segmentation, and possibly grading of cancer. Based on recent advances in deep learning computational models, we study this problem within this framework for developing a solution.

- **Compression of pathological images:** Last but not least, the histopathological slides are huge in size and can easily exceed the size of an ultra-high-definition (UHD) image. In order to store or communicate such large images in a system, the compression rate is an important parameters that also affects the automated analysis results. Designing a data-driven algorithm that processes the whole-slide images or efficiently transfers them via a communication channel, e.g. in a remote diagnosis system, requires the use of compression techniques. Using a lossy compression leads to a degraded analysis performance so that a careful trade-off should be made. Therefore, the impact of image compression on a detection and segmentation algorithm should be known and evaluated. This chapter will study the effect of image compression on both training and inference of the data when a deep learning model is used for histopathological image analysis.

The directions of the proposed solutions for the individual challenges are commonly based on a deep learning approach. As shown in Figure 3.1, for the stain-color normalization, we develop two deep learning models that outperform existing methods with the highest stability in image colors. Regarding the automated detection and segmentation of cancer, we exploit the ConvNet models that is combined with probabilistic graphical model to improve the segmentation of images. This model has shown superiority over all proposed models in an international competition for breast cancer metastases detection and segmentation. For the compression, for the first time, we study the impact of compression rate on the training and inference of a ConvNet on the histopathological images.

As shown in Figure 3.1, this chapter is organized to introduce and investigate each of these challenges separately because of their different nature and to present the proposed solutions in detail. To this end, each challenge is elaborated in a separate section of this chapter. In Section 3.2, different deep learning models will be introduced for specifically addressing and learning stain-color normalization. Section 3.3 deals with automated detection and segmentation of cancer, starting with state-of-the-art developments, after which a unified model for both the detection and segmentation tasks is designed. In Section 3.4, the impact of JPEG2000 compression on the performance of a deep learning model both in training and inference phases is studied. In each of these main sections, related work, methodology and empirical evaluations are addressed and discussed and completed with individual conclusions. In Section 3.5, the complete chapter is finalized and conclusions are discussed and generalized for building a CAD system for computational pathology.



**Figure 3.1** Layout overview of the chapter

## 3.2 Stain-color normalization

As mentioned earlier, histopathology involves a manual staining procedure for preparing tissues prior to microscopic imaging for cancer diagnosis. This non-quantified procedure may cause a considerable variation in the color characteristics of tissue samples. Such systematic color variations affect the CAD performance when they are very different from the image data that have been included in the training set. To compensate for these effects in a CAD system, stain color normalization is a common practice. Recent studies show that statistical normalization of data in general [50, 51] and additional color normalization of H&E stained histopathology images can increase the computational efficiency of CAD systems and lead to a higher performance [52, 53, 54].

### 3.2.1 Related work

Various stain normalization methods have been proposed for stain-color normalization in histopathology images [55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 52, 65, 66, 67]. Previously published studies often follow one of three main concepts: stain-color deconvolution, template matching, and multi-task learning. In the sequel, first, these three approaches are briefly explained, including their properties and drawbacks, and then the research contributions are described.

**A. Stain deconvolution:** Considering prior knowledge of the reference stain vector for every dye, which is present in whole-slide images (WSIs), the stain deconvolution methods split an input RGB image into three stain channels, each representing the actual color components of the stain used. This class of stain-normalization methods is used extensively in histopathology image analysis [61]. Ruifrok *et al.* [68] introduced this prior knowledge by manually selecting pixels that represent a specific stain class and then computing the color deconvolution vector. Because of some drawbacks of this semi-automatic procedure, several studies were done later for automatic extraction of stain color by using the sin-

gular value decomposition (SVD) technique [57], probabilistic Gaussian mixture model (GMM) [56] using a prior for stain-matrix estimation [58] and stain-color descriptions along with training a supervised relevance vector machine [61]. Although these solutions all aimed at better estimation of stain vectors, the estimation of the stain vectors was solely restricted to a limited analysis of color information in the image contents, while the spatial dependencies and underlying tissue structures were ignored [52]. Such ignorance causes some shortcomings for approaches based on stain deconvolution when severe stain-color variations occur in the data.

**B. Template color-matching** was proposed by *Reinhard et al.* [55] and relied on aligning the statistical color properties (e.g. mean and standard deviation) of the source image with a template image. The authors used a separate linear transform for each channel of the image in the Lab color space. Since each dye contained its own specific reaction pattern, each dye gave an independent contribution to the final color in the image. Therefore, a single transformation on each color channel underperformed to modern techniques. The drawbacks of such an approach also were mentioned in [61, 52]. For addressing this issue, applying separate transformations on stain classes [56, 52] or on tissue classes [60] were investigated. For avoiding artifacts at the border of different classes under different transformations, a weighted contribution of these transformations in the final color image was considered. Two proposed solutions were estimated weights of the GMM [56] and training of a naive Bayesian classifier [52]. Although the mentioned studies reduced the artifacts in normalized color images, the usage of a template-matching approach showed two inherent major drawbacks. Firstly, the color transformation between the source image and the template image was split into two transformation classes (one for background with no staining and one for all tissue structure in the foreground [56]), or three transformation classes (background, elliptical hematoxylin-stained cell nuclei and other eosin-stained tissue structures in H&E staining). The assumption of presenting a limited (e.g. up to three) number of classes for structures present in histopathology images can be easily violated if the tissue type changes, and defining more tissue classes needs prior shape knowledge and designing additional informative shape descriptors for classification. For example, in [52] elliptical shape of nuclei and the use of randomized Hough transformation on the Canny edge map was proposed for detecting nuclei, while other similar objects in shape like blood cells were removed by an empirical threshold. Secondly, template color-matching methods did inference on new samples with unknown chromatic characteristics by first estimating their statistical properties and then fitting a model to the data. This procedure can be prone to high error if the number of test samples is limited or does not represent well the population statistics.

**C. Multi-task learning:** Recently, the capacity of deep generative models has been explored for performing stain-color normalization [69, 70]. This involves the application of a GAN [71] in the framework of multi-task learning (also called *stain-style transfer* model). These methods use a GAN for learning color normalization. Nonetheless, for learning the color conversion, an additional *discrimina-*

*tive* task should be defined (e.g. supervised classification of tumors from normal tissue). The GAN tries to convert the image colors to maximize the performance of a classifier. These methods benefit from using a neural network for learning the transfer function of colors, but defining the problem in the principle of multi-task learning introduces some limitations. For example, this approach is not a generic solution because it is not able to address some histology studies, e.g. in the absence of labels for the data, or even when the slide-level labels are available only.

To overcome the mentioned shortcomings with existing methods, in the following, we present two different deep learning-based methods for stain-color normalization, namely, GAN-based and neural GMM-based models.

### 3.2.2 GAN-based method

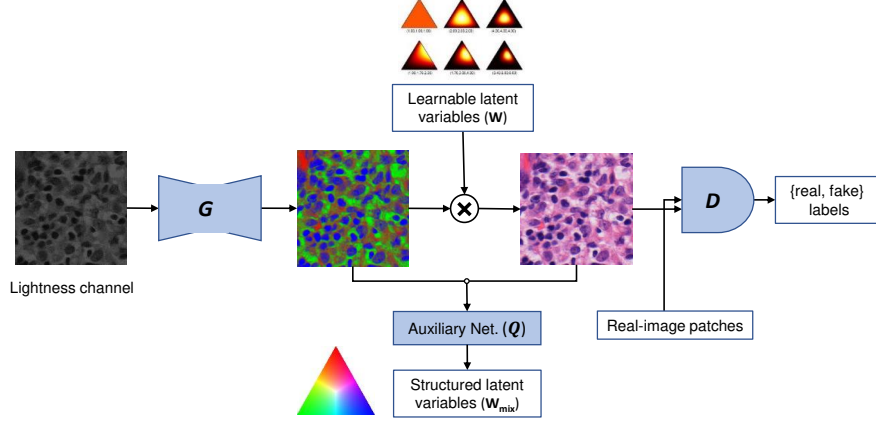
This subsection presents a new model for stain-color normalization based on the GAN framework. The contributions to the stain-color normalization problem are twofold.

- An unsupervised end-to-end learning model is presented that can decompose the pixels in the histopathology images into several clusters and then model the stain-color attributes of each cluster by using structured latent variables of the model. Unlike the previous studies, there are neither any hard constraints on the number of clusters for image structures nor any explicit assumptions about the present tissues or visual contents, such as defining a prior shape for cells or any assumed available labels for images.
- The proposed model benefits from (1) the learning capacity of deep ConvNet for decomposing the image contents, and (2) a better learning of a smooth mapping function between source and template images that is not only based on the first and second-order color statistics [55], [56] or the first Eigenvector of the chromatic plane [52].

The GANs [71] learn a generative distribution  $p_G(x)$  over the training sample set  $X$ , by minimizing its discrepancy with the real data distribution  $p_{\text{data}}(x)$ . The GANs include a generator network  $\mathcal{G}$  that generates samples  $\mathcal{G}(z)$ , given an uninformative (noise) variable  $z$ , drawn from a prior  $p_{\text{noise}}(z)$ . The generator is trained by playing against an adversarial discriminator network  $\mathcal{D}$  that tries to distinguish between samples from  $p_{\text{data}}(x)$  and  $p_G(x)$  [71]. The MinMax formulation is given by a value function  $V$ , specified by:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}) = E_{x \sim P_{\text{data}}}(\log \mathcal{D}(x)) + E_{z \sim \text{noise}}(\log(1 - \mathcal{D}(\mathcal{G}(z)))) \quad (3.1)$$

Since the standard GANs formulation does not have any restriction on the contribution of noise variables ( $z$ ) in generating the samples, the individual dimensions of  $z$  can become highly entangled and not in correspondence with the interpretable features of the data [72] (e.g. with a realistic colorizing of the H&E images in our case study). *InfoGAN* [72] has introduced a set of structured latent



**Figure 3.2** Diagram of GAN-based stain-color normalization, where the lightness channel of training image patches is colorized by the generator  $\mathcal{G}$ . The  $\mathcal{G}$  network includes a ConvNet and a set of learnable latent variables ( $W$ ). The colorized patches are evaluated by a discriminator network ( $D$ ) to be realistic. The  $Q$  network at the bottom retrieves a part of  $W$ , what is called structured latent variables ( $W_{\text{mix}}$ ).

variables  $\mathbf{c} = \{c_1, c_2, \dots, c_L\}$  that along with the uninformative noise variable  $\mathbf{z}$ , are pushed into the generator model  $\mathcal{G}(\mathbf{z}, \mathbf{c})$ . Hence, the generator learns a mapping between  $p(\mathbf{z}, \mathbf{c})$  and  $p_G(x)$ . Based on InfoGAN, discovering the latent factors  $\mathbf{c}$  in an unsupervised way is performed by adding an information-theoretic regularization term to the MinMax value function of the GANs [72], that the updated formula is specified by:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V_I(\mathcal{D}, \mathcal{G}) = V(\mathcal{D}, \mathcal{G}) - \lambda I(\mathbf{c}; \mathcal{G}(\mathbf{z}, \mathbf{c})). \quad (3.2)$$

Here, the function  $I(\cdot)$  represents mutual information (MI) and  $\lambda$  is a contribution weight (e.g. in our experiments,  $\lambda=1$ ). Because computing  $I(\mathbf{c}; \mathcal{G}(\mathbf{z}, \mathbf{c}))$  needs access to the posterior  $p(\mathbf{c}|x)$ , which is hard to maximize, the authors suggest using an auxiliary distribution  $Q(\mathbf{c}|x)$  as a lower bound to approximate  $p(\mathbf{c}|x)$ . Estimating  $Q(\mathbf{c}|x)$  is performed by training a neural network called *auxiliary* network ( $Q$ ). In simple words, after generating samples from pairs input  $(\mathbf{z}, \mathbf{c})$ , the  $Q$  network should retrieve the latent variables  $\mathbf{c}$  via the mutual information. The authors showed that these structured variables can be used for encoding some semantic information into the generated images, such as geometric transformations like controlling the rotation or scale of objects.

Inspired by InfoGAN, we construct a GAN for learning the chromatic distribution of H&E images through a finite set of latent variables and consequently generate different colorized versions of images by adjusting the distribution parameters. The proposed model is slightly modified from InfoGAN for two reasons. Firstly, the proposed model aims to operate as a color-normalizing model

which must be faithful to the visual contents of the input image by preserving the structures present in the source image. Therefore, instead of generating images from a noise distribution, our generator network  $\mathcal{G}$  performs a colorization task by receiving the monochrome *lightness* channel (of the CIEL\*a\*b color space) of the source image. The colorization of the lightness channel is performed by a ConvNet and using a set of structured latent variables. The CIEL\*a\*b color space has been selected because of its higher performance for reconstructing histopathological images [73]. Secondly, in the proposed model, the structured latent variables  $c$  have the role of color-system matrix for representing color for image structures. These variables control the color of learned clusters from the input images. Figure 3.2 depicts the diagram of the proposed model. The model consists of three ConvNets: a generative network ( $\mathcal{G}$ ), a discriminator network ( $\mathcal{D}$ ), and an auxiliary network ( $\mathcal{Q}$ ), which are all trained simultaneously. A detailed description of each network is provided below.

**Generator network ( $\mathcal{G}$ )** learns how to generate a colorized H&E image in its CIEL\*a\*b color space by giving the image lightness channel and a set of latent variables, drawn randomly from a prior distribution. The  $\mathcal{G}$  net consists of several convolutional, nonlinear Rectified Linear Unit (ReLU) functions, max-pooling, and batch normalization in its hidden layers. The generator maps the given lightness channel to a latent  $k$ -simplex probability subspace (e.g.  $k = 3$ ) through its softmax layer with  $k$  outputs. Each point in this probability simplex represents a  $k$ -dimensional probability vector  $\mathbf{p}_{G_{1 \times 3}}$  that is softly clustering the pixels of the input image into  $k$  clusters. Afterwards, the produced  $\mathbf{p}_{G_{1 \times 3}}$  vectors in this latent space are passed on to the second part of  $\mathcal{G}$  net, which is transformed linearly to the output for generating the full-color images. Analogous to the *color-system matrix* of Ruifrok *et al.* [68], we call this linear transformation, the color-system transform. This transformation consists of matrix multiplications, applied to the  $\mathbf{W}_{c_{3 \times 4}}$  and  $\mathbf{W}_{m_{4 \times 3}}$  matrices. We consider the elements of each  $i$ -th row of the  $\mathbf{W}_c$  matrix as the parameters of Dirichlet distributions ( $i = 1, 2, 3$ ) with random values ( $\alpha^i = [\alpha_1, \alpha_2, \dots, \alpha_4]$ ). This transformation can be considered as a stochastic process on the  $\mathbf{p}_{G_{1 \times 3}}$  vectors. To avoid swapping colors between image structures in each training iteration, we impose a constraint ( $\arg \max_i (\alpha^i)$ ) on randomly drawn  $\alpha^i$  parameters. This constraint forces the rows of the color-system matrix to be sampled from three isolated regions of the probability simplex, so it leads to assigning consistent colors to structures/clusters.

The Dirichlet distribution is used as a prior for the color-system matrix. If we consider the contribution of used staining dyes in colorizing each pixel as a multinomial distribution over dyes, then using a different amount of dye or any variation in the staining procedure alters the parameters of the multinomial distribution. The Dirichlet distribution prior allows for more flexible modeling of data, when considering it as a distribution over possible parameter vectors of the multinomial distribution. In fact, this concept can be seen as a *distribution over distribution* [74]. From the experiments, we observe that using such a prior for  $\mathbf{W}_c$  can lead to a good approximation of the posterior in a generative model. The  $\mathbf{W}_m$  transformation is used for shifting and scaling of network output to

produce CIEL\*a\*b space of images. The elements of the  $\mathbf{W}_m$  matrix and all parameters of the  $\mathcal{G}$  net are learned by gradient descent optimization. As a standard GANs framework, the  $\mathcal{G}$  net does not have its own loss function and its parameters are jointly optimized with the  $\mathcal{D}$  net, by back-propagated gradients from the discriminator loss.

**Auxiliary network ( $\mathcal{Q}$ )** has been defined in the context of the *InfoGAN* [72], for learning disentangled structured latent variables by the GANs. The  $\mathcal{Q}$  net has almost an inverse functionality compared with the  $\mathcal{G}$  net. It receives the colorized images along with their  $\mathbf{p}_G$  vectors from the  $\mathcal{G}$  net and estimates the elements of  $\mathbf{W}_c$  at its output. Its loss function has been defined to maximize the mutual information between the output and the elements of the  $\mathbf{W}_c$  matrix.

**Discriminator network ( $\mathcal{D}$ )** minimizes its loss function and tries to distinguish the colorized images from their original colorful version. The architecture of the  $\mathcal{D}$  net is very similar to the  $\mathcal{Q}$  net and consists of similar convolutional hidden layers. However, it is also different from  $\mathcal{Q}$  net, since instead of the last max-pooling and the fully-connected layers, it has a global average-pooling layer and a scalar output after applying a sigmoid function. The probability at the output of the  $\mathcal{D}$  net is learned to be maximal (e.g. equal to unity) when given a real colorful image and is minimal (e.g. zero) when supplied with a fake colorized image from the  $\mathcal{G}$  net. Table 3.1 specifies the detailed layering of the architectures of these three networks.

**Reconstruction loop:** For training of the three networks mentioned above and apart from the value function in Eq. (3.2), we perform a reconstruction loop as a supervisory signal through  $\mathcal{G}$  and  $\mathcal{D}$  networks. This introduces another loss function that measures the reconstruction quality of a given real image. The conducted experiments have empirically shown that this extra procedure facilitates in faster convergence of the networks. To this end, we first apply the  $\mathcal{G}$  net to the lightness channel of an input train sample. It produces  $\mathbf{p}_G$  vectors at its hidden layers. We supply the same real image along with the obtained  $\mathbf{p}_G$  vectors to the  $\mathcal{Q}$  net. Consequently, the  $\mathbf{W}_c$  matrix is replaced with the estimated values at the output of the  $\mathcal{Q}$  net. Now, the  $\mathcal{G}$  network should generate the same full-color copy of the given image. The  $L_2$  loss between input and the generated sample is used for measuring the reconstruction quality. This process is almost similar to what happens by an auto-encoder network.

For the inference, the  $\mathcal{Q}$  and  $\mathcal{G}$  networks are first applied to the template image, so the color-system matrix belonging to the template image is obtained. Afterwards, by using this color matrix, the  $\mathcal{G}$  net colorizes the lightness channel of any given source image to resemble the colors of the template image.

### 3.2.3 Empirical evaluation of GAN-based method

The architecture of the proposed GAN-based stain-color normalization model has been shown in Table 3.1. The model is trained on image patches of size  $299 \times 299$  pixels. For training the model, the ADAM optimizer with a fixed learning rate equal to  $1 \times 10^{-4}$  has been used.



### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS

**Table 3.1** Specification and architecture details of the individual layers of the GAN-based stain-color normalization model

Generator network ( $\mathcal{G}$ )					
Layer name	Layer type	Input size	Output size	Kernel size	Activation
c1	convolution	$299 \times 299 \times 1$	$299 \times 299 \times 16$	$3 \times 3$	ReLU
c2	convolution	$299 \times 299 \times 16$	$299 \times 299 \times 32$	$3 \times 3$	ReLU
p1	pooling	$299 \times 299 \times 32$	$150 \times 150 \times 32$	-	max
c3	convolution	$150 \times 150 \times 32$	$150 \times 150 \times 64$	$3 \times 3$	ReLU
p2	pooling	$150 \times 150 \times 64$	$75 \times 75 \times 64$	-	max
c4	convolution	$75 \times 75 \times 64$	$75 \times 75 \times 128$	$3 \times 3$	ReLU
c5	convolution	$75 \times 75 \times 128$	$75 \times 75 \times 256$	$3 \times 3$	ReLU
p3	up-pooling	$75 \times 75 \times 256$	$150 \times 150 \times 256$	-	max
c6	convolution	$150 \times 150 \times 256$	$150 \times 150 \times 256$	$3 \times 3$	ReLU
p4	up-pooling	$150 \times 150 \times 256$	$299 \times 299 \times 128$	-	max
c7	convolution	$299 \times 299 \times 128$	$299 \times 299 \times 32$	$3 \times 3$	ReLU
c8	convolution	$299 \times 299 \times 32$	$299 \times 299 \times 3$	$3 \times 3$	ReLU
output	softmax	$299 \times 299 \times 3$	$299 \times 299 \times 3$	-	-
Discriminator/Auxiliary networks ( $\mathcal{D}$ , $\mathcal{Q}$ )					
Layer name	Layer type	Input size	Output size	Kernel size	Activation
c1	convolution	$299 \times 299 \times 6$	$299 \times 299 \times 32$	$3 \times 3$	ReLU
c2	convolution	$299 \times 299 \times 32$	$299 \times 299 \times 32$	$3 \times 3$	ReLU
p1	pooling	$299 \times 299 \times 32$	$150 \times 150 \times 32$	-	max
c3	convolution	$150 \times 150 \times 32$	$150 \times 150 \times 32$	$3 \times 3$	ReLU
p2	pooling	$150 \times 150 \times 32$	$75 \times 75 \times 32$	-	max
c4	convolution	$75 \times 75 \times 32$	$75 \times 75 \times 16$	$3 \times 3$	ReLU
p3	pooling	$75 \times 75 \times 16$	$32 \times 32 \times 16$	-	max
c5	convolution	$32 \times 32 \times 16$	$32 \times 32 \times 16$	$3 \times 3$	ReLU
p4	pooling	$32 \times 32 \times 16$	$16 \times 16 \times 16$	-	max
c6	convolution	$16 \times 16 \times 16$	$16 \times 16 \times 8$	$3 \times 3$	ReLU
Auxiliary network ( $\mathcal{Q}$ )					
Layer name	Layer type	Input size	Output size	Kernel size	Activation
p5	pooling	$16 \times 16 \times 8$	$8 \times 8 \times 8$	-	max
f1	fully-connected	$1 \times 512$	$1 \times 32$	-	ReLU
f2	fully-connected	$1 \times 32$	$3 \times 8$	-	ReLU
output	linear	$3 \times 8$	$3 \times 4$	-	-
Discriminator network ( $\mathcal{D}$ )					
p5	global-pooling	$16 \times 16 \times 8$	$1 \times 1 \times 8$	-	average
output	fully-connected	$1 \times 8$	$1 \times 1$	-	sigmoid

#### A. Dataset

For evaluation, our focus is on inter-laboratory variations of the H&E staining in the lymph-node dataset, since this is a major concern in the large-scale application of CAD in pathology. For better comparison with recent studies, we have used a similar dataset, as introduced in [52]. The dataset contains 625 images (each  $1388 \times 1040$  pixels) from 125 digitized H&E-stained WSIs of lymph nodes from 3 patients, collected from five different Dutch pathology laboratories, each using their own routine staining protocols. More details about this dataset can be found in [52]. The proposed model is trained on  $299 \times 299$  randomly cropped patches and evaluated on full-size WSIs by using leave-one-out cross-validation, based on the laboratories where the samples were originally collected.

**Table 3.2** Statistics of the NMI measure for hematoxylin dye for five laboratories. The columns indicate the standard deviations (SD) and coefficients of variation (CV) of NMI.

Method	Lab 1		Lab 2		Lab 3		Lab 4		Lab 5		Average	
	SD	CV	SD	CV	SD	CV	SD	CV	SD	CV	SD	CV
Original	0.033	0.065	0.031	0.060	0.037	0.078	0.029	0.049	0.028	0.051	<b>0.032</b>	<b>0.060</b>
Macenko[57]	0.029	0.052	0.026	0.046	0.020	0.037	0.025	0.044	0.020	0.035	<b>0.024</b>	<b>0.043</b>
Reinhard[55]	0.032	0.058	0.025	0.044	0.020	0.035	0.030	0.052	0.029	0.049	<b>0.027</b>	<b>0.047</b>
Khan[61]	0.066	0.156	0.067	0.155	0.085	0.158	0.054	0.110	0.049	0.093	<b>0.064</b>	<b>0.135</b>
Bejnordi[52]	0.016	0.029	0.015	0.027	0.018	0.034	0.029	0.055	0.024	0.044	<b>0.021</b>	<b>0.038</b>
Proposed method	0.024	0.053	0.019	0.043	0.020	0.043	0.027	0.057	0.024	0.053	<b>0.022</b>	<b>0.050</b>

## B. Results of the GAN-based model

The performance of the proposed method is compared to that of four previously published algorithms: linear appearance normalization by Macenko *et al.* [57], statistical color properties alignment by Reinhard *et al.* [55], nonlinear mapping for stain normalization by Khan *et al.* [61] and whole-slide image color standardizer by Bejnordi *et al.* [52]. Similar to prior works [60], [52], the normalized median intensity (NMI) measure is used to evaluate the color constancy of the normalized images. The NMI measure is specified by:

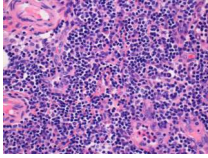
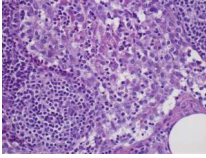
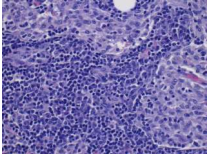
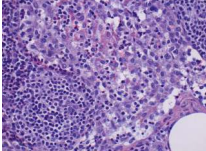
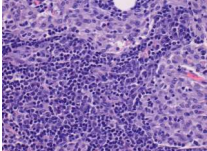
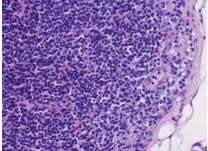
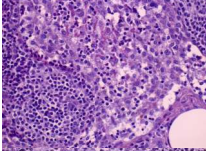
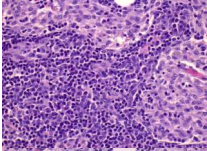
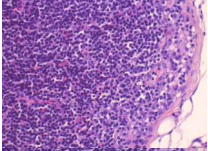
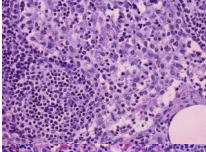
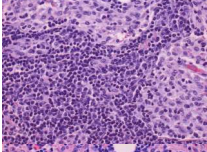
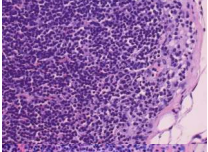
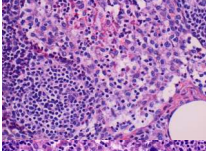
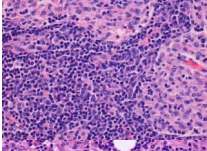
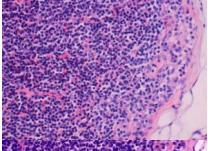
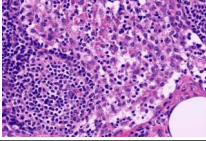
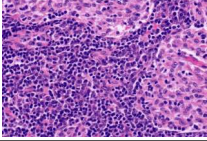
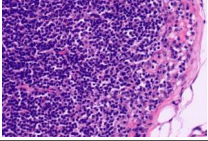
$$\text{NMI}(x) = \frac{\text{Median}(\{U(i)\}_{i \in x})}{P_{95}(\{U(i)\}_{i \in x})}, \quad (3.3)$$

where  $U(i)$  denotes the average of RGB values for the pixel  $i$  in the image  $x$  and  $P_{95}$  denotes the 95% percentage level. To increase the robustness of the NMI measure in Eq. (3.3) against noisy pixels in the image, the NMI measure is divided by the 95<sup>th</sup> percentile, instead of normalizing by the maximum value. The standard deviation (SD) of the NMI values and coefficient of the variation (CV) (i.e., standard deviation divided by mean) of the NMI values are computed for the images of different laboratories, prior to and after standardization, using different methods for comparison.

Similar to prior work, the quantitative analysis of results is based on the color constancy of nuclear staining and eosin staining, independently. To evaluate the color constancy of the nuclear staining, nuclei are first detected automatically. Similar to [52], the fast radial symmetry transform [75] and marker-controlled watershed [76] algorithms are employed for nuclei detection. The dataset includes the manually annotated eosin-stained regions for 25 images. The statistical results of the NMI metric for hematoxylin and eosin regions are listed in Table 3.2 and Table 3.3, respectively. The obtained results indicate that the proposed model outperforms many previous methods for stain-color normalization. Figure 3.3 illustrates the normalization of two example images from two distinct labs by using different methods. As can be observed, the proposed method has acceptable qualitative performance in the normalization of different image structures, including the white background to resemble the template image.

As shown by the experiments, the proposed GAN-based model can learn the chromatic space of H&E images and normalize them. The color-normalized image preserves the structures of the source image, while being forced to pertain a

### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS

	Lab I	Lab II	Lab III
Template	Source Images		
			
Method	Color-converted images		
Macenko <i>et al.</i> [57]			
Reinhard <i>et al.</i> [55]			
Khan <i>et al.</i> [61]			
Bejnordi <i>et al.</i> [52]			
Ours			

**Figure 3.3** Qualitative evaluation of stain-color normalization of different methods on three H&E images from three distinct labs.

high mutual chromatic information with a template. In contrast to most previous methods of which the a-priori computation of statistical properties from source and template images is essential for the inference stage, the proposed GAN-based model can be applied instantly to unseen given images. This can be crucial for methods when the number of test samples is small and consequently the limited amount of data would hamper the estimation of true statistics. Moreover, the proposed framework takes minimal assumptions about the number, shape, color

**Table 3.3** Statistics of the NMI measure for eosin dye for collection of 25 images from five labs. The columns indicates the standard deviation (SD) and coefficient of variation (CV) of NMI

Method	Original	Macenko[57]	Reinhard[55]	Khan[61]	Bejnordi[52]	Proposed method
NMI SD	0.0563	0.0362	0.0386	0.0434	<b>0.0191</b>	0.0195
NMI CV	0.0748	0.0439	0.0494	0.0555	0.0220	<b>0.0218</b>

and other image attributes of H&E images. This leads to more generic modeling that can be applied to histopathological images from different organs, containing different tissue structures and potentially to other staining modalities such as immunohistochemistry staining. The experiments have demonstrated that the proposed model can outperform many previous methods qualitatively and quantitatively.

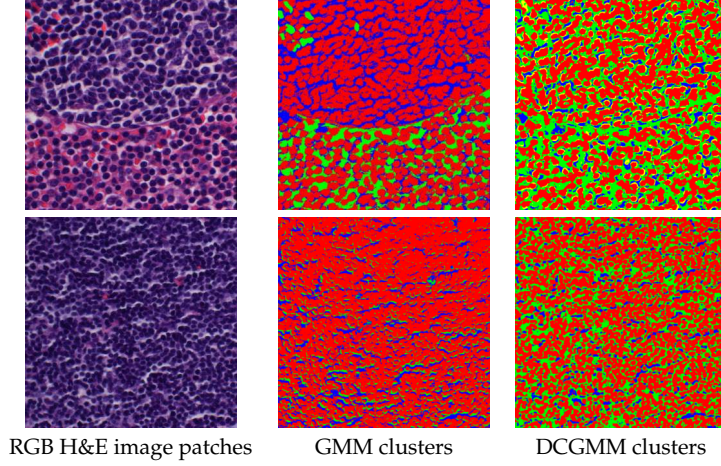
The next section studies a different, alternative unsupervised learning method for the stain-color normalization problem. The proposed method is based on hybrid modeling that combines the probabilistic GMM and ConvNets in an end-to-end learning framework.

### 3.2.4 Neural-GMM method

Gaussian mixture models (GMMs) [77] is a well-established probabilistic modeling that is typically deployed for unsupervised clustering in machine learning frameworks.

The GMM has been used in histopathology stain-color normalization modeling [56, 60]. Nevertheless, these methods measure the similarity of each individual pixel in the image with all other pixels, regardless of their underlying structures and visual contents of the images. The GMM still lies at the core of numerous modern applications, however, its similarity measure is limited to local relations in the data space (i.e. pixel-level relations) and is thus unable to capture hidden, hierarchical dependencies in latent spaces [78]. Fortunately, deep neural networks and in particular ConvNets can encode rich latent structures of visual input data. However, employing deep learning models in an unsupervised framework requires considering some inductive biases in their architecture or in the training procedure. For addressing the shortcomings of the standard GMM method for stain-color normalization, in this section, we propose a neural GMM-based model. Our contributions are threefold.

- *Deep Convolutional Gaussian Mixture Model (DCGMM)* is introduced, which is a neural-augmented GMM for unsupervised stain-color normalization of histopathological images. The DCGMM benefits from the capability of a ConvNet for performing soft clustering of tissue structures and is using the GMM likelihood for learning the multi-modal color distribution of the input images.
- *Context-aware distribution modeling*: The GMM-based prior work treats the color of each pixel independently of its underlying tissue structure and vi-



**Figure 3.4** Examples of pixel clustering, using the conventional GMM and the proposed DCGMM for histopathology images. Standard GMM treats the color of each pixel as an independent variable, ignoring the tissue structure and other visual context information. Unsupervised DCGMM optimizes the parameters of a ConvNet as a feature extraction model for maximizing the GMM likelihood.

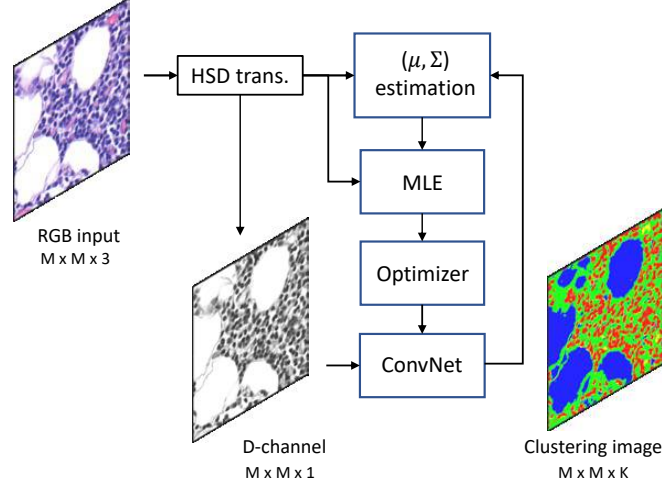
sual contents. In contrast, the proposed DCGMM fits a mixture of Gaussian distributions to the input images that is conditional to their visual contexts (including appearances and shapes of regions in the image intensity channel). The conducted experiments show that the DCGMM outperforms the previous methods specifically when strong staining variations appear in the data (see Figure 3.4). This is mainly due to the independency of the proposed method from the chromatic information in tissue clustering and consequent distribution modeling.

- *Joint optimization of the GMM and ConvNet’s parameters:* Instead of using a common expectation-maximization (EM) algorithm for maximizing the likelihood function, an end-to-end learning procedure is introduced for jointly optimizing the parameters of a ConvNet and the GMM. To our knowledge, this is the first neural-augmented GMM that may be employed for other applications than color normalization such as color distribution modeling. However, in the following, this study concentrates only on its application in stain-color normalization of histopathology images.

In the following, first a brief overview of the standard GMM method is provided, while introducing the notation used later. Second, the DCGMM method is described in detail.

**A. Gaussian Mixture Model** of data vector ( $\mathbf{x}$ ), can be presented as a linear superposition of  $K$  Gaussian distributions in terms of discrete *latent* variables ( $\mathbf{z}$ ), in the form of

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad (3.4)$$



**Figure 3.5** Block diagram of DCGMM during the training phase. The color RGB images are transformed into Hue-Saturation-Density (HSD) color space. A fully convolutional network computes a probability vector per pixel. So, the parameter of GMM can be updated and the maximum likelihood estimation (MLE) by maximizing Eq. (3.6) is performed to optimize the parameter of network using gradient-descent algorithm, iteratively.

The  $K$ -dimensional binary random variable  $\mathbf{z}$  has one-hot encoding ( $z_k \in \{0, 1\}$ ;  $\sum_{k=1}^K z_k = 1$ ), which represents the tissue class in the study. This variable is a hidden variable in the GMM. In Eq. (3.4), the mixing coefficients  $\pi_k$  should satisfy  $0 \leq \pi_k \leq 1$  together with  $\sum_{k=1}^K \pi_k = 1$ , in order to fulfill a valid probability definition over variable  $\mathbf{z}$  [77]. Here,  $\mathcal{N}$  stands for a multivariate normal distribution with mean  $\boldsymbol{\mu}_k$  and covariance matrix  $\boldsymbol{\Sigma}_k$ . If we consider  $\pi_k$  as prior probability of class  $z_k$ , its posterior probability called *responsibility* can be written [77] as:

$$\gamma(z_k) = p(z_k = 1 | \mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (3.5)$$

According to Eq. (3.4), the (natural) log-likelihood function for an image ( $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ) with the total number of pixels (observations) equal to  $N$  is

$$\ln(p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})) = \sum_{n=1}^N \ln\left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right). \quad (3.6)$$

Given the GMM, the objective is to maximize the likelihood function (Eq. (3.6)) with respect to the parameters  $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k)$ . The common approach for optimizing the parameters of the standard GMM is using the EM algorithm by iteratively evaluating the responsibilities (Eq. (3.5)) and re-estimating the parameters.

**B. Deep Convolutional GMM (DCGMM):** The recent development of deep generative models has invoked some extensions to the standard GMM [79, 80]. Two proposed approaches are: (1) constructing a stack of multiple GMM layers



on top of each other in a hierarchical architecture [79] that is optimized by an EM-based algorithm, and (2) using an auto-encoder neural network while imposing the GMM on its bottle-neck representations [80] that has been studied for unsupervised anomaly detection. This study presents a different extension to the standard GMM, by introducing the DCGMM that involves optimizing the parameters of a ConvNet to fit a GMM to the color distribution of histopathology images. For training the DCGMM, the standard gradient descent and the back-propagation algorithm are employed.

DCGMM aims to fit a GMM to the pixel-color distribution, conditioned on tissue classes. For processing the image and detecting the tissue classes, the high learning capability of the ConvNet has been exploited. To this end, the network estimates the responsibility coefficients (see Eq. (3.5)) that indicate the probability of each pixel in the image, belonging to a tissue cluster. For a better understanding, the E-step in an EM-based optimization is considered to be replaced by a neural network. However, all parameters of the GMM and ConvNet are jointly optimized by the gradient descent algorithm. In the DCGMM, the negative log-likelihood (maximizing Eq. (3.6)) is used as the loss function.

The proposed color normalization algorithm can be split into two phases: training the DCGMM and the color transformation (inference). In the training phase, a GMM is fitted to the data. At inference time, the template and source image are individually supplied to the model. Consequently, the parameters of the fitted Gaussian distributions and their mixture coefficients ( $\pi$ ) are computed in those two images. Afterwards, the multivariate Gaussian distributions of the source image are transformed (aligned) to have similar parameters of distributions as the template image, while  $\pi$  is kept unchanged. The remainder of this section explains these two phases in detail.

**B1. Fitting Gaussian distribution to the data:** First the RGB color values are transformed to the hue-saturation-density (HSD) color system [81]. In an HSD space of histopathological images, the density channel ( $D$ ) is linearly related to the amount of stain, while the other two chromatic channels are independent. This property suits well to the analysis of transmitted light microscopy, compared to the alternative color spaces [52]. We only use the normalized zero-mean (centered) density channel as the input to the network. Hence, we ignore the chromatic information and clustering of pixels is performed only based on tissue structures and their appearance (normalized density channel). This color ignorance alleviates the effect of strong staining variations in images. The DCGMM has a fully-convolutional architecture, consisting of several convolutional layers, ReLU non-linearity functions and (un)pooling operators. The reduced size of the feature map after applying two stages of max-pooling returns back to its original size by applying un-pooling operations. DCGMM uses a softmax layer at its output layer, to produce a valid probability vector for each pixel. Thus, it satisfies the constraint of  $\sum_{k=1}^K \gamma_k = 1$ . The network aims to predict the responsibility values (see Eq. (3.5)) for each pixel in the input image. Since the model is applied on H&E histopathological images, each pixel in the image mostly belongs to one out of three clusters ( $K = 3$ ): hematoxylin, eosin or background (i.e. not stained).

Because the biological composition of tissue related to each pixel in the image leads to a varying stain absorption ratio between pixels, the color of each pixel can be presented by a weighted sum of the different stains used. This property can be reflected in the responsibility coefficient ( $\gamma$ ) of a GMM, which is estimated in the softmax layer of the network in the proposed model. The calculation of the required partial derivatives of negative log-likelihood (i.e. loss function) with respect to its parameters ( $\pi$ ,  $\mu$  and  $\Sigma$ ), for performing a gradient descent algorithm can be found in [82, p. 45]. Therefore, the gradient of Eq. (3.6) with respect to the learning parameters ( $\pi$ ,  $\mu$ ,  $\Sigma$ ) of the GMM and the neural network can be computed. The pseudocode in Algorithm 1 shows the training procedure of the DCGMM in detail.

---

**Algorithm 1** - DCGMM training pseudocode

---

**Data:**  $\mathbf{X} \leftarrow$  training image

**Result:** Optimized parameters  $\theta_{net}$  of network ( $f_{net}$ )

$\theta_{net} \leftarrow$  Initialize network parameters

**repeat**

$\mathbf{X}_h, \mathbf{X}_s, \mathbf{X}_d \leftarrow \text{RGB2HSD}(\mathbf{X})$

$\gamma \leftarrow f_{net}(\bar{\mathbf{X}}_d, \theta_{net})$  ▷ (1)

$N_k \leftarrow \sum_{n=1}^N \gamma(z_{nk})$  ▷ (2)

$\mu_k \leftarrow 1/N_k \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$  ▷ (3)

$\Sigma_k \leftarrow 1/N_k \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$  ▷ (4)

$\pi_k \leftarrow N_k/N$  ▷ (5)

$\mathcal{L} \leftarrow -\sum_{n=1}^N \ln\{\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)\}$  ▷ loss

$\theta_{net} \leftarrow \theta_{net}^+ - \nabla_{\theta_{net}}(\mathcal{L})$  ▷ update

**until** convergence;

---



---

**Algorithm 2** - DCGMM inference pseudocode

---

**Data:**  $\mathbf{X}_{h,s,d}^t \leftarrow$  template image (in HSD color space)

$\gamma^t, \mu^t, \Sigma^t, \pi^t \leftarrow$  from Op.(1-5)

**Data:**  $\mathbf{X}_{h,s,d}^s \leftarrow$  source image (HSD)

$\gamma^s, \mu^s, \Sigma^s, \pi^s \leftarrow$  from Op.(1-5)

**for**  $k:=1$  to  $K$  *step 1* **do**

$\mathbf{Y} \leftarrow \mathbf{X}_k^s - \mu_k^s$  ▷ Centering

$\Phi^s \Lambda^s \Phi^{s-1} \leftarrow \Sigma_k^s$  ▷ SVD

$\mathbf{Z} \leftarrow \Lambda^{s-\frac{1}{2}} \Phi^{sT} \mathbf{Y}$  ▷ Whitening

$\Phi^t \Lambda^t \Phi^{t-1} \leftarrow \Sigma_k^t$  ▷ SVD

$\mathbf{X}_k^{new} \leftarrow \Phi^t \Lambda^{t\frac{1}{2}} \mathbf{Z} + \mu_k^t$

**end**

**Result:**  $\mathbf{X}^{new} \leftarrow \sum_{k=1}^K [\gamma_k^s \circ \mathbf{X}_k^{new}]$

---



### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS

**Table 3.4** Statistics of the NMI measure for hematoxylin and eosin dyes for five laboratories. The columns indicate the standard deviations (SD) and coefficients of variation (CV) of NMI.

Method	Hematoxylin												Eosin	
	Lab 1		Lab 2		Lab 3		Lab 4		Lab 5		Average		SD	CV
	SD	CV	SD	CV	SD	CV	SD	CV	SD	CV	SD	CV		
Original	0.033	0.065	0.031	0.060	0.037	0.078	0.029	0.049	0.028	0.051	0.032	0.060	0.0563	0.0748
Macenko[57]	0.029	0.052	0.026	0.046	0.020	0.037	0.025	0.044	0.020	0.035	0.024	0.043	0.0362	0.0439
Reinhard[55]	0.032	0.058	0.025	0.044	0.020	0.035	0.030	0.052	0.029	0.049	0.027	0.047	0.0386	0.0494
Khan[61]	0.066	0.156	0.067	0.155	0.085	0.158	0.054	0.110	0.049	0.093	0.064	0.135	0.0434	0.0555
Vahadane[83]	0.036	0.065	0.032	0.058	0.024	0.046	0.023	0.042	0.020	0.038	0.027	0.050	0.034	0.041
Bejnordi[52]	<b>0.016</b>	<b>0.029</b>	<b>0.015</b>	<b>0.027</b>	0.018	<b>0.034</b>	0.029	0.055	0.024	0.044	0.021	0.038	0.0191	0.0220
VAE-based[84]	0.029	0.052	0.022	0.043	0.021	0.064	0.028	0.050	0.025	0.055	0.025	0.0528	0.026	0.036
GAN-based	0.024	0.053	0.019	0.043	0.020	0.043	0.027	0.057	0.024	0.053	0.022	0.050	0.0195	0.0218
DCGMM	0.022	0.045	0.017	0.034	<b>0.017</b>	0.036	<b>0.014</b>	<b>0.030</b>	<b>0.017</b>	<b>0.035</b>	<b>0.017</b>	<b>0.036</b>	<b>0.0188</b>	<b>0.0209</b>

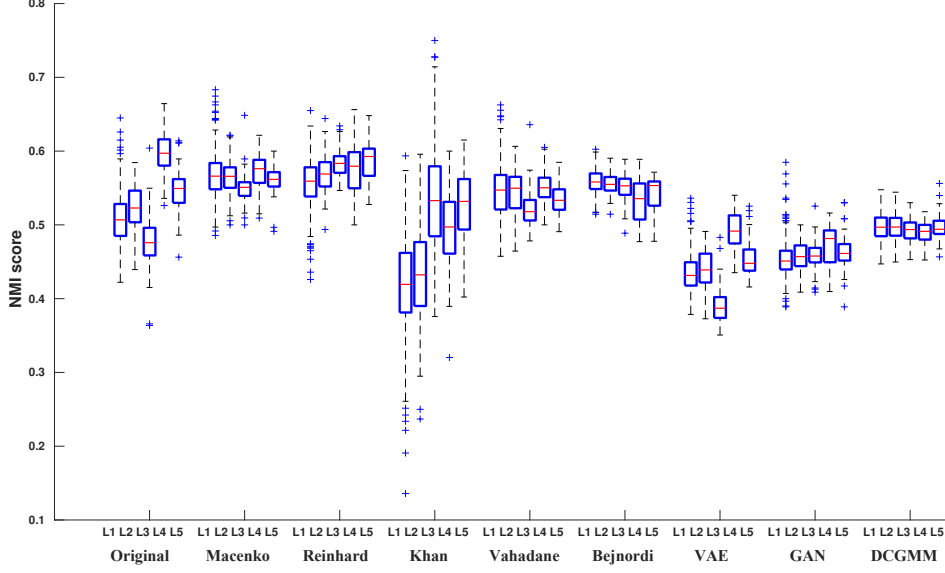
Although the responsibility term is predicted by the network from only the density channel of the image, the mean and covariance matrices ( $\mu_k$  and  $\Sigma_k$ ) are estimated from all three channels of the HSD color image. The randomly initialized parameters of the network are updated by ADAM gradient-based optimization with a fixed learning rate of 0.001. The diagram of the proposed model is depicted in Figure 3.5.

**B2. Color conversion** in DCGMM involves the transformation of two multivariate Gaussian distributions. By training the model in the color normalization task, two GMMs are fitted to the source and template images, individually. Afterwards, a set of transformations are applied to align the multivariate Gaussian distributions between the source and the template. These transformations consist of three operations: *mean centering*, *whitening* and *coloring* transformations. We assume that  $(\mu^s, \Sigma^s)$  and  $(\mu^t, \Sigma^t)$  are the predicted parameters of the two distributions for the source and template image, respectively. By shifting the mean of the source image to the origin (*centering*) and then whitening which involves the SVD algorithm, the source image will have a zero mean and an identity covariance matrix. Consequently, by applying a “coloring” transformation (which is the inverse of whitening), but after replacing the Eigenvalues ( $\Lambda$ ) and Eigenvectors ( $\Phi$ ) of the source distribution with the template distribution, the whitened Gaussian distribution of the source image is scaled and rotated to obtain the same covariance matrix as the template image ( $\Sigma^t$ ). Finally, the distribution is shifted to obtain the same mean as the template distribution. For clarifying this procedure, pseudocode in Algorithm 2 further describes these steps in detail.

#### 3.2.5 Empirical Evaluation of Neural-GMM method

##### A. Dataset

We focus on inter-laboratory variations of the H&E staining, as it is a major concern in the large-scale application of CAD in pathology. For better comparison with recent studies, we use the same data set as has been introduced in [52]. The data set contains 625 images (each  $1388 \times 1040$  pixels) from 125 digitized



**Figure 3.6** Boxplot of NMI scores in hematoxylin regions for the original images from different laboratories and their color-normalized versions by different methods.

H&E stained WSIs of lymph nodes from 3 patients and was collected from five Dutch pathology laboratories, each using their own routine staining protocols (more details can be found in [52]). Our model is trained on randomly cropped patches ( $576 \times 576$  pixels) and evaluated on full-sized images by using leave-one-out cross-validation based on the above laboratories.

## B. Results

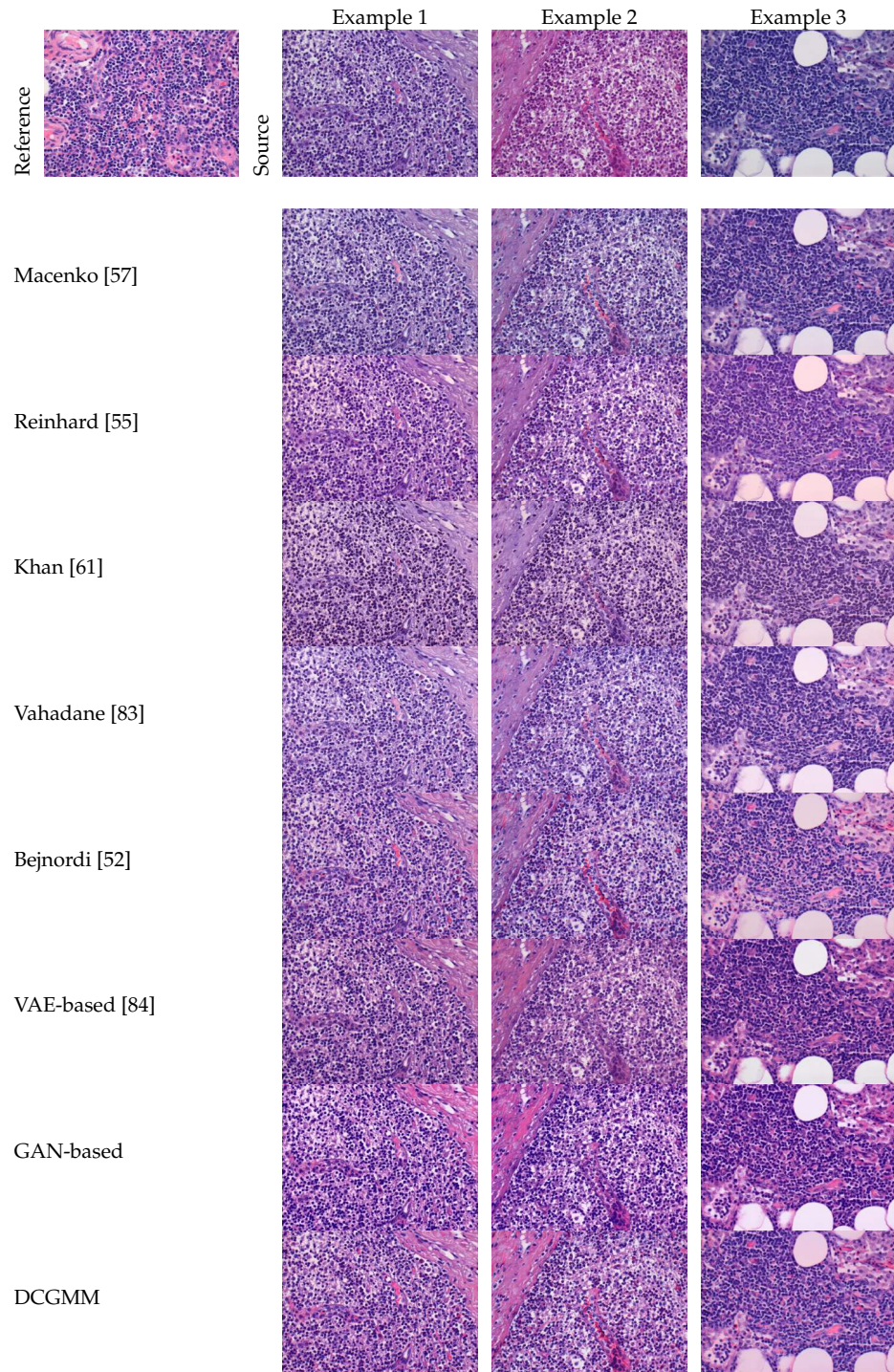
We trained the DCGMM on the data set. The model converges relatively fast in a few minutes. The inference computation time for each image in its original size ( $1388 \times 1040$  pixels) is about 0.6 seconds, implemented in the TensorFlow library and running on a TITAN Xp GPU. The performance of our method is compared to that of five competing state-of-the-art algorithms: linear appearance normalization by Macenko *et al.* [57], statistical color properties alignment by Reinhard *et al.* [55], nonlinear mapping for stain normalization by Khan *et al.* [61], sparse non-negative matrix factorization by Vahadane *et al.* [83] and WSI color standardizer by Bejnordi *et al.* [52]. The normalized median intensity (NMI) measure (Eq. 3.3) is used to evaluate color constancy of normalized images. Quantitative analysis is based on independently evaluating the color constancy in the regions that show mostly absorbed hematoxylin or eosin. Since nuclei mostly absorb hematoxylin, they first are detected automatically by using a fast radial symmetry transform and a marker-controlled watershed algorithm [52]. Since our generated hema-

toxylin masks slightly differ from what was used in [52], the obtained NMI scores in our benchmark are not exactly the same as reported in [52]. However, the results are in agreement with [52]. For evaluation of the eosin analysis, several regions are manually annotated for 25 images. The evaluation results of different methods for assessing hematoxylin and eosin regions are shown in Table 3.4. For a better comparison, the evaluations of the hematoxylin regions are visualized in Figure 3.6. The results clearly indicate that our proposed method results in the lowest variation in color after normalization of the images and it outperforms competing state-of-the-art methods. Figure 3.7 illustrates an example of the template image, a source image and the outcomes of color normalization by the different methods.

#### 3.2.6 Discussion

We presented two unsupervised deep learning models for stain-color normalization in histopathology images. The proposed methods implicitly learn to extract different image structures that have the same chromatic characteristics. They lack any threshold, ground truth or any other assumptions about the shape and color of structures present in histopathology images. Having fewer assumptions about the image contents enables the methods to be generic and applicable to different histopathology images corresponding with different tissue types. By experiments, we show that both GAN-based and DCGMM models outperform competitive methods on stain-color normalization tasks.

### 3.2. Stain-color normalization



**Figure 3.7** Qualitative evaluation of different stain color-normalization methods for three H&E images from three distinct labs with respect to a reference image.

#### 3.3 Cancer detection and classification

Advanced image analysis can lead to an automated examination of histopathology images which is essential for objective and fast cancer diagnosis. Recent deep learning methods, in particular ConvNets, have shown exceptionally successful performance on medical image analysis as well as computational histopathology. Since WSIs have a very large size, the ConvNets are commonly applied to classify WSIs per patch. Although a network is trained on a large population of the input patches, the spatial dependencies between patches are typically ignored. Consequently, the inference is limited only to the level of visual contexts that have appeared in the individual patches. As a result, the prediction of the neighboring regions can be inconsistent.

In the study of this section, the dependencies among neighboring patches are modeled by graphical models with the objective to exploit a broader context in the inference for each individual patch. To this end, we apply Conditional Random Fields (CRFs) over latent spaces of a trained deep ConvNet, in order to jointly assign labels to the patches. In the proposed approach, the extracted compact features from intermediate layers of a ConvNet are considered as the observations in a fully-connected CRF model. This leads to performing inference on a wider context rather than the appearance of individual patches. The conducted experiments show an improvement of approximately 3.9% on average FROC score for tumorous region detection in histopathology WSIs. The proposed model is trained on the Camelyon17[85] ISBI challenge dataset and has won the 2<sup>nd</sup> place in this international competition with a kappa score of 0.876 at the patient-level for pathology lymph-node classification to detect breast cancer. This section explains the proposed model in detail.

##### 3.3.1 Metastases detection and grading using ConvNets

Detecting tumor cells in microscopic examination of stained histological slides can be tedious work for pathologists; this procedure is error-prone and depends on qualification and skills. The emergence of fast digital scanners producing Gigapixel pathology images is a major driving force for research in the automation of cancer detection by developing CAD systems. Creating a CAD system that approaches human performance, can be a highly challenging task because of the large variations in morphology, color patterns of stained slides and inter-patient visual contents. The CAD system can assist pathologists by comprehensive evaluation of WSIs in a short amount of time. The main challenge in designing such a CAD system lies in precisely detecting regions that contain tumorous cells. In this section, we study the automated detection and classification of breast cancer metastases in WSIs. Investigation on improving the performance of the current CAD system to be comparable to or even better than humans is highly clinically relevant and can reduce the subjectivity in diagnosis.

Given the complexity of the data, a supervised training of a ConvNet can be the best choice, since it shows outstanding results in medical image classification and segmentation tasks [86], particularly in computational histopathology [87].

Because of the very large size of WSIs, training the network as a patch classifier is a common approach [88, 87, 89]. In such a framework, all patches inside a WSI are classified individually. Therefore, the prediction over neighboring patches in WSI may be incoherent and the labels of these patches may be inconsistent. Furthermore, the inference would be limited to the contextual information of each patch. For addressing this problem, recently Kong *et al.* [89] integrated a ConvNet with a 2D Long Short-Term Memory (LSTM) network for modeling the spatial dependencies among patches, as well as their visual contents. However, their approach requires a two-stage training setup, which makes it heavily dependent on available data and is computationally expensive.

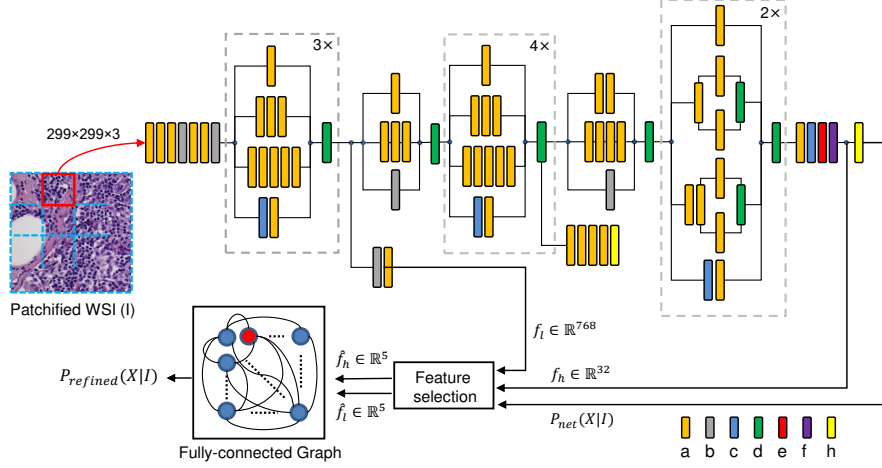
Instead of using an RNN for modeling the dependencies across patches, we propose an alternative approach for incorporating the spatial dependencies by using probabilistic graphical modeling and in particular, Conditional Random Fields (CRFs). Among different variants of the proposed CRF models for image segmentation, a fully-connected (dense) pairwise CRF shows higher performance, since it considers all possible dependencies among nodes/pixels by constructing a complete graph over the image [90]. Similar to the *DeepLab* architecture [38], the research of this section proposes using a dense CRF as a post-processing stage for refining the ConvNet predictions (i.e. labels). The conducted research of this section contributes in two ways.

- *Learned patch descriptor as node attribute*: Firstly, in the proposed model, each node in the complete graph is not a pixel, but it represents an image patch that is located inside the WSI domain. Therefore, assigning pixel attributes (e.g. color) to the nodes of the graph is not relevant anymore. Instead, the representation of a patch in the hidden layers of a deep ConvNet is used as a region descriptor and being assigned to its node.
- *Multi-level feature fusion*: Secondly, we leverage the representations of patches in both lower and upper layers of the network for improving segmentation performance by embedding them jointly into the graph. Combining lower and upper-layer features, corresponding to primitive and complex visual contexts from distinct layers of network, allows analysis of the images at different levels.

In the remainder of this section describes the proposed approach and evaluates its performance in detecting tumor regions in histopathology WSIs of breast lymph nodes. In Section 3.3.2 the method based on combining ConvNets with CRF models is presented. In Section 3.3.3 the experiments are shown on a breast cancer dataset. Finally, this study is finalized with a discussion on the results.

### 3.3.2 Cancer detection using CRFs on deep embedded spaces

The block diagram of the approach is depicted in Figure 3.8. The method consists of three main stages: pre-processing of the images, training a ConvNet on input image patches, and post-processing by using CRF and clustering methods.



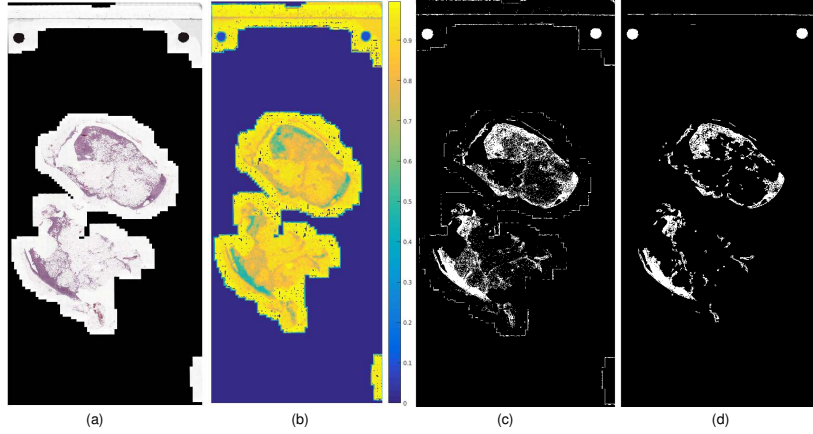
**Figure 3.8** Block diagram illustrates the proposed method in the cancer detection framework. Initially, the Inception-v3 network is trained for patch-level cancer classification. Subsequently, the network is utilized for enhancing prediction accuracy at the whole slide image (WSI) level. The neural features predicted for each patch are represented as nodes within a fully-connected graph, facilitating pairwise inference. To incorporate WSI-level contextual information, a Conditional Random Field (CRF) is applied to refine the probability vectors of the nodes on the graph. This process ensures improved accuracy by leveraging contextual information at a higher level. The colorful rectangular blocks indicate: (a) convolution layer, (b) max pooling, (c) mean pooling, (d) concatenation, (e) dropout, (f) fully-connected layer, and (h) softmax layer.

The pre-processing aims to find the Region of Interest (ROI) in WSIs and subsequently extract patches for training a classifier. Afterwards, a deep ConvNet is trained for binary classification of the patches, by assigning labels (e.g. normal or tumor) to the patches. Finally, in the post-processing stage, a feature selection algorithm is performed on the extracted features from the hidden layers of the trained network. The selected features along with the network probability map are used in a CRF model for jointly labeling all the regions in the WSI.

#### A. Pre-processing

The pre-processing stage consists of ROI detection, patch extraction and data augmentation. The ROI detection provides the tissue images by removing void spaces in the WSIs image. These ROIs are suitable for later analysis and the irrelevant empty areas are rejected for further processing. Approximately 80% of a WSI area contains a background region [87] that makes the ROI-extraction stage necessary for performing efficient computation. For achieving this, we apply the adaptive Otsu thresholding technique [91] inside a sliding window with a predefined fixed stride value. The obtained threshold value is assigned to the central pixel of the window. After finding a threshold map for the WSI, we limit the min-





**Figure 3.9** ROI detection steps prior to patch extraction; (a) original WSI; (b) Otsu threshold map; (c) binary mask image; (d) filtering out the isolated regions.

imum and maximum values of the map by considering a global Otsu threshold which is computed over the whole image. Figure 3.9 (a) and Figure 3.9 (b) show an original WSI and the corresponding threshold map, respectively. Using the computed threshold map, the original WSI is binarized (see Figure 3.9 (c)).

Finally, the ROI mask is obtained by applying morphological operations for removing small isolated objects. By considering the computed ROI mask, we randomly crop patches within the tissue region from the original slides. For better training of the ConvNet, common data augmentation techniques are used, such as randomly rotating the patches by 4 multiples of  $90^\circ$  and flipping them vertically and horizontally. For increasing the generalization ability of the classifier against minor changes in chromatic information of the images, we also apply color augmentation to training data. This leads to training a broader range of color variations compared to what occurs in the training set. For this reason, we add a little noise to the lightness and saturation channels of the patches after they are converted from the RGB to the HSL color space.

### B. ConvNets as Patch Classifier

As a patch classifier, the Inception-v3 network [92] is used, which forms a 48-layer deep ConvNet. This network has shown higher performance in image classification with a much lower number of parameters, compared to its preceding versions, due to its convolution factorization strategy [92] and the resulting lower number of training parameters. Wang *et al.* [87] reported promising results of a first version of this ConvNet architecture, applied to histopathology images in comparison with other well-known convolutional architectures.

The network inputs are color patches with a size of  $299 \times 299$  pixels and the outputs are binary labels (tumor vs. normal), which are presented by a one-hot encoding vector. For the benefits of using pre-trained parameters of the network



on natural images, we preserve the input image size of the network, identical to its original version that was pre-trained on the ImageNet dataset. However, we have slightly modified the original Inception-v3 architecture, by changing the dimensionality of the last logits from 2,048 to 32 elements. This results in extracting a 32-dimensional feature vector  $\mathbf{f}_h \in R^{32}$  prior to the output softmax layer of the network. The lower dimensionality of this layer provides a more compact representation of the patches, which later leads to higher computational efficiency in the CRF model. Furthermore, along with the  $\mathbf{f}_h$  feature vector, we export another feature vector  $\mathbf{f}_l \in R^{768}$ , which is sampled from the feature map of size  $17 \times 17 \times 768$  of the Inception-v3 network, after a  $5 \times 5$  max-pooling with stride 3 and a  $5 \times 5$  convolution kernel (see Figure 3.8). In contrast to the vector  $\mathbf{f}_h$ , its counterpart  $\mathbf{f}_l$  yields a low-level representation of the visual data.

### C. Conditional Random Fields (CRFs) on embedding spaces

This part starts with a brief overview of the well-known CRF method for pixel labeling in an image. A CRF model is defined over a complete undirected graph. The nodes of the graph represent the pixels in the image [90]. By considering a random field for the labels assigned to the pixels, which is defined over a set of random variables  $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ , we assign to each random variable a label from the set of  $\{tumor, normal, void\}$ , conditioned on observations  $\mathbf{I} = \{I_1, I_2, \dots, I_N\}$ , where  $\mathbf{I}$  represents a WSI with  $N$  patches. In the case of WSI data, the nodes of the graph are not pixels but patches, so the observation  $I_j$  is defined as a set of region descriptors like features of the network for the patch  $j$ , while  $x_j$  denotes its label.

Based on standard definition, the joint distribution of  $(\mathbf{X}, \mathbf{I})$  is a CRF if the random variables of set  $\mathbf{X}$  conditioned on  $\mathbf{I}$ , satisfy the Markov property. This CRF is characterized by a Gibbs distribution, which is specified as

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-E(\mathbf{X}|\mathbf{I})), \quad (3.7)$$

where the energy is defined as

$$E(\mathbf{X}|\mathbf{I}) = \sum_{c \in C_G} \psi_c(X_c|\mathbf{I}), \quad (3.8)$$

and  $G$  is a graph on  $\mathbf{X}$ , where each clique  $c$  in the set of cliques  $C_G$  induces a potential  $\psi_c$ . In a fully-connected pairwise CRF model,  $G$  is a complete graph on  $\mathbf{X}$  and  $C_G$  denotes the set of all unary and pairwise cliques. The corresponding Gibbs energy can be written as

$$E(\mathbf{X}|\mathbf{I}) = \sum_i \psi_u(x_i) + \sum_{i < j} \psi_p(x_i, x_j), \quad (3.9)$$

where  $i$  and  $j$  range from 1 to  $N$ . The unary potential ( $\psi_u$ ) is computed independently and comes from the probability output of a softmax layer in the

trained network. The key idea for computationally efficient inference in a fully-connected CRF model lies in defining the pairwise edge potentials ( $\psi_p$ ) to be a linear combination of Gaussian kernels in arbitrary feature space (see Eq. (3.10)) [90]. The pairwise potential is defined in the following form:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^K [w^{(m)} \cdot k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)], \quad (3.10)$$

where  $\mu(x_i, x_j)$  is a label-compatibility function between labels, for example, it can be defined by using a Potts model [90], in its simplest form. The vectors  $f_i$  and  $f_j$  are feature vectors,  $k^{(m)}$  is a Gaussian kernel for measuring similarity in feature space. In our case, this feature space consists of the spatial coordinates of the patch's center ( $\mathbf{p}$ ) in WSI and patch representations in latent spaces of the network ( $\hat{\mathbf{f}}_h \in R^5$  and  $\hat{\mathbf{f}}_l \in R^5$ ) that we are going to define them later. For the pairwise potential in the CRF formulation, we introduce three kernels ( $K=3$ ):

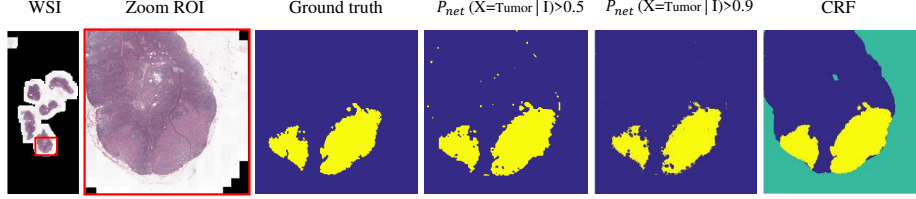
$$\begin{aligned} k(\mathbf{f}_i, \mathbf{f}_j) = & w^{(1)} \exp\left(-\frac{|\mathbf{p}_i - \mathbf{p}_j|^2}{2\sigma_\alpha^2}\right) \\ & + w^{(2)} \exp\left(-\frac{|\mathbf{p}_i - \mathbf{p}_j|^2}{2\sigma_\alpha^2} - \frac{|\hat{\mathbf{f}}_l i - \hat{\mathbf{f}}_l j|^2}{2\sigma_\beta^2}\right) \\ & + w^{(3)} \exp\left(-\frac{|\mathbf{p}_i - \mathbf{p}_j|^2}{2\sigma_\alpha^2} - \frac{|\hat{\mathbf{f}}_h i - \hat{\mathbf{f}}_h j|^2}{2\sigma_\gamma^2}\right) \end{aligned} \quad (3.11)$$

The first term on the right-hand side of the Eq. 3.11 is called *smoothness kernel* [90] that removes the small isolated regions. The next two terms can be considered as *appearance kernels* that emphasize the observation that nearby pixels with similar features (by considering both low-level  $\hat{\mathbf{f}}_l$  and high-level  $\hat{\mathbf{f}}_h$  features) are likely to be in the same class. The *width* of the Gaussian kernels ( $\sigma_\alpha$ ,  $\sigma_\beta$  and  $\sigma_\gamma$ ) and their contribution weights ( $w^{(m)}$ ) are adjusted by performing a grid search and measuring the performance of the model on a validation set. For minimizing the energy (see Eq. 3.9), we apply a proximal-gradient based approach for mean-field inference with ADAM optimization [93] that shows faster convergence and often finds better optima compared with standard mean-field variational inference. A fixed number of 30 iterations is used for mean-field variational inference. Figure 3.11-i shows the Kullback–Leibler (KL) divergence as the state of convergence in the inference algorithm [90] for an example processed WSI. After doing inference, the CRF updates the label of each patch of WSI. In our case, the labels correspond to three classes: tumor, normal and void.

#### D. Feature selection

Designing a proper feature space in a CRF model is crucial for its performance. For introducing a suitable feature space, two points should be considered. Firstly, in a pairwise CRF model, the potential between two nodes is computed w.r.t.

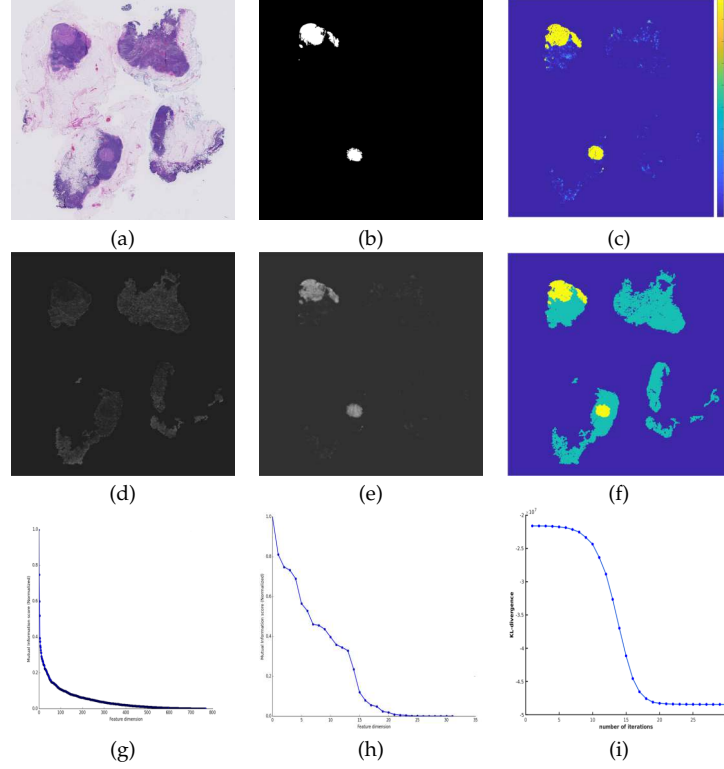
### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS



**Figure 3.10** Detection of tumor regions (yellow areas), before and after applying the CRF model. The figure shows the binarized network prediction with two different thresholds.

their similarity, which is evaluated by the Gaussian kernels (Eq. (3.10)). The assignment to the nodes of the graph of a high-dimensional redundant feature vector that lacks sufficient discriminative elements, deteriorates the similarity measurement. Secondly, we use approximation techniques for efficient inference in the CRF by using a permutohedral lattice [90], which is a filtering technique for reducing the complexity of the convolution operation in message passing of the mean-field approximation. Nevertheless, the complexity of such an inference is still in the order of  $\mathcal{O}(Nd)$  [90], where  $N$  represents the number of nodes in the graph and  $d$  is the dimension of feature space. When the length of the feature vector is high ( $d \gg 10$ ), the method becomes practically infeasible due to the limitation of memory and computational power.

For efficiency in computation and similarity measurement, a feature selection procedure is employed for which we have adopted a supervised univariate feature selection method. We perform the CRF post-processing only for WSIs which have at least one patch per slide that is classified as a positive class. Therefore, we can be confident that candidates for both classes are present in the examined WSI. Here, the positive class is defined by comparing the ConvNet output probability value with a predefined threshold (Figure 3.10). For speeding up the feature selection stage, we only use a subset of patches from both classes by random selection. Afterwards, the mutual information between each dimension of the feature vectors and the probability value of the network is measured. Finally, by sorting the measured scores, we can rank the most important dimensions of feature vectors (see Figures 3.11(g) and 3.11(h)). By choosing  $n$  of the most important elements of each  $\mathbf{f}_h$  and  $\mathbf{f}_l$  vectors (e.g.  $n=5$ ), we reduce the feature vector length to a total of  $2n$  elements for each patch in the WSI. The two new down-sampled feature vectors for logits and hidden layers are called  $\hat{\mathbf{f}}_h$  and  $\hat{\mathbf{f}}_l$ , respectively. According to the experiments for finding the optimal length of the feature vector ( $n$ ), considering the sorted values of mutual information is not helpful because a significant gap between them (for finding cutoff) is not observed (see Figures 3.11(g) and 3.11(h)). Therefore, the feature vector length is empirically adjusted to be equal to 5 in the experiments. It is worth mentioning that using feature vectors with a larger length ( $n > 5$ ) demands higher computational resources that for the large histopathology WSIs would be beyond the available computing power and memory at the time of conducted research.



**Figure 3.11** Refinement of tumor detection with the CRF model. (a) Original WSI. (b) Ground truth. (c) CNN probability map. (d) Average of top-5 CNN features ( $\mathbf{F}_{H5}$ ) in hidden layer of the network. (e) Average of top-5 CNN features in logits layer ( $\hat{\mathbf{f}}_h$ ). (f) CRF segmentation result. (g-h) Mutual information scores (sorted and normalized) between the CNN probability map and each of the  $\mathbf{f}_l$  and  $\mathbf{f}_h$  vectors, respectively. (i) Kullback-leibler divergence as the state of convergence during successive iterations of inference in the CRF algorithm [90].

### 3.3.3 Empirical Evaluation

#### A. Experimental settings

Although the proposed method has won the 2nd place in the Camelyon17 [94] challenge for metastases detection and grading, here, for evaluation of the post-processing technique we have performed the evaluation experiments based on the Camelyon16 [95] data. This is mainly due to the availability of pixel-level annotations for WSIs that are required for our assessment. Whereas only a small part of the Camelyon17 data was annotated at the pixel-level, Camelyon16 includes pixel-level annotation for a total of 400 WSIs, split into 270 for training and 130 for testing. The training set consists of 160 positive (tumor) and 110 negative (normal) WSIs. The test set consists of 49 positive and 81 negative WSIs. For all positive WSIs, a pixel-level annotation mask has been provided by patholo-

gists. The free-response receiver operating characteristic (FROC) curve, which is the ratio of sensitivity versus the average number of false-positives per image is measured for evaluation. As the final score, the average FROC (Ave. FROC) which is the average of the sensitivities at 6 predefined false-positive (FP) rates: 1/4, 1/2, 1, 2, 4 and 8 FPs per WSI are reported. A higher value of Ave. FROC presents a better performance.

For studying the performance of the proposed method, we evaluate the results of metastasis detection before and after applying the CRF model. Thus, we first train the network on the training set and evaluate its performance on the test set as a *baseline* method. Then, by applying the CRF post-processing to the outputs of network, we assess the improvement, compared with the baseline. After pre-processing the data and finding the tissue regions in WSIs, according to what was explained earlier, we select a large number of patches from both normal and tumor regions in 20X magnification level of the training set. The number of extracted patches is approximately equal for all positive WSIs (2,000 patches per class per WSI). We also sample approximately 2,000 patches from each negative WSI.

#### B. Experimental results on tumor detection

We train the network with initial parameters that have been adapted to the ImageNet 2012 Challenge data. This helps to accelerate the convergence time. The learning rate is changed by monitoring the loss function value on the validation set. The batch size is equal to 32 and a weight decay has been used for prohibiting the over-fitting issue. After false-positive bootstrapping on the training set and by an inference stride equal to 64, the network achieves Ave. FROC score of 0.812 on the test set.

After inference on each test WSIs by applying the network, we store the network output probability along with the extracted features,  $\mathbf{f}_h$  and  $\mathbf{f}_l$ . Then, as already explained, a feature selection is applied to downsample these two feature vectors. By assigning the computed probability to the unary potential in the CRF model and the feature vectors to the pairwise potential (see Eq. (3.9)), we do inference again on test WSI by applying the CRF model. Adding the CRF post-processing, the measured Ave. FROC equal to 0.851 on the test set was obtained. This is 3.9% improvement compared with the baseline. The result for an example test WSI is shown in Figure 3.11.

In comparison with the state-of-the-art, the study of Kong *et al.* [89] is the most similar work to ours in terms of modeling the spatial dependencies in histopathology WSIs. They trained a deep residual network on the same training data as we introduced earlier. Afterwards, the network was considered as a feature extractor and was used for inference on patches, of which their central position is localized on a 2D grid over the input WSI. Subsequently, by passing the extracted feature vectors into a 2D-LSTM model, they reported an improvement of about 5% on Ave. FROC score (by performing a fivefold cross-validation on the training set but not evaluated on test set). Although the proposed method

shows slightly lower improvement (about 1%) compared with Kong *et al.*[89], the proposed approach does not need training of a secondary network for learning the spatial dependencies in WSIs. However, energy-based learning methods such as the CRF algorithm, are relatively slow in inference time. It is worth noting that by applying the proposed post-processing method along with adding some morphological and clustering techniques, we have obtained a quadratic-weighted kappa score equal to 0.876 in patient-level metastasis classification of the breast lymph nodes. This has resulted in the 2<sup>nd</sup> place in the Camelyon17 ISBI challenge [94].

### 3.3.4 Discussion and Conclusion

Histopathology image analysis can significantly benefit from CAD. We expect that patch-based classification using ConvNets will become a common approach in analyzing WSIs. Besides several advantages of this approach, due to its local inference mechanism, it has some shortcomings like ignoring both short- and long-distance spatial dependencies, as well as contextual dependencies among image patches. For modeling such dependencies across patches, in this section, the CRF technique has been proposed as a post-processing step. Empirical evaluations show that incorporating a CRF model for the dependencies between all patches inside a WSI and imposing a joint labeling strategy improves the tumor region detection. Since our post-processing method is solely based on the features that have been already incorporated for the prediction by the ConvNet, the classification improvement can be explained by the inference performed on a broader context rather than the patch-based analysis. This can be observed by achieving a high kappa score equal to 0.876 in patient-level metastasis classification on the breast lymph data.

### 3.4 Impact of JPEG compression on deep neural networks

#### 3.4.1 Introduction and related work

*Medical image compression:* The availability of massive amounts of data in histopathological Whole-Slide Images (WSIs) has enabled the application of deep learning models and especially ConvNets, which have shown a high potential for improvement in cancer diagnosis. However, storage and transmission of large amounts of data is challenging such as for Gigapixel histopathological WSIs. The positive impact of using compression technique is that it lowers the requirements with memory access and bandwidth. However, the exploitation of lossy compression algorithms for medical images is controversial yet acceptable, provided that the clinical diagnosis is not affected by compression distortion.

In this section, we study the impact of JPEG 2000 compression on the proposed deep learning model [96], which has produced a comparable performance to that of pathologists and which was ranked on the 2<sup>nd</sup> place in the CAMELYON17 challenge [94]. Detecting tumor metastases in hematoxylin and eosin-stained tissue sections of breast lymph nodes is evaluated and compared with the pathologists' diagnoses in three different experimental setups. The experiments show that the ConvNet is robust against compression ratios up to 24:1, when it is trained on uncompressed high-quality images. We demonstrate that a model trained on lower quality images – i.e. lossy compressed images – depicts a classification performance that is significantly improved for the corresponding compression ratio. Moreover, it is also observed that the model performs equally well on all higher-quality images. These properties will help to design cloud-based CAD systems (e.g. telemedicine) to employ deep learning models that are more robust to image quality variations resulting from the required compression. However, the results presented in this study are specific to the CAD system and application described in this manuscript, and further work is needed to examine whether they generalize to other systems and applications.

*Compression for computational pathology:* Emerging new scanners for digital microscopic imaging make it possible to acquire Gigapixel histopathological images at a large scale. These large-scale digital datasets make digital pathology a perfect use case for deploying data-greedy, deep learning models. The availability of these massive amounts of data in combination with recent advances in deep learning models and more specifically convolutional neural networks, results in a situation where for many clinical image-analysis tasks, computational pathology solutions have comparable performance to that of humans [86]. For example in pathology, recent deep learning-based techniques are comparable to or even outperform humans in detecting and localizing breast-cancer metastases in lymph node WSIs [97].

Although increasing the number of image samples enhances the performance of a deep network by better learning of the image-content diversity [98], the intrinsic image quality of the used samples will also impact the model performance. Furthermore, dealing with a large database for storage and the associated

transmission for cloud-based computing is challenging, while for the design of a CAD system it is even critical. For example, working on big data in the cloud requires reconciling two contradictory design principles. On one hand, cloud computing is based on the concepts of consolidation and resource pooling, while on the other hand, big data systems (such as Hadoop) are based on the shared nothing principle, where each node is independent and self-sufficient [99]. These issues are more crucial in telemedicine and cloud-based computation, regarding privacy and security issues. For example, in the CAMELYON17 challenge [85, 94], which is an international competition on designing the best CAD algorithm for automated breast-cancer metastases detection, about 1000 histopathological WSIs (more than 3 Terabytes of image data) have been made publicly available. Downloading the whole data on a local machine for training a CAD model is cumbersome and required a significant amount of time and network bandwidth.

*Lack of standard for medical image compression:* Given the large size of WSIs, the use of compression algorithms forms a very appealing solution. Particularly, lossy compression that can support larger compression ratios is interesting. Fortunately, it is generally not prohibited by the main regulatory bodies in the European Union, United States, Canada and Australia, provided that it does not impair the diagnostic quality and does not cause new risks compared with conventional practice [100]. Hence, it is important to define a strategy or protocol for an efficient parameterization of the deployed compression techniques to yield a high compression ratio without jeopardizing the classification performance. The issue of a higher compression ratio with lower encoding time has been recognized as well in recent efforts for creating the DICOM standard in the field of digital pathology [101].

*Existing research on the impact of image compression:* For studying the impact of lossy compression on the diagnostic performance of human experts, several studies have been reported [102, 103, 104, 100, 105, 106, 107, 108]. Mostly, these studies reported that the human visual perception is to some extent robust against image quality degradation. However, there is not a generally accepted tolerance level with respect to diagnostic accuracy. In addition, because clinical evaluations can be subjective and have a bias regarding the task at hand and the skill of experts, different studies have suggested varying compression ratios corresponding to the addressed clinical task. For example, Kalinski *et al.* [100] reported that the impact of a JPEG 2000 compression factor up to 20 did not show significant influence on the detection of *Helicobacter pylori* gastritis in gastric histopathological images, performed by three pathologists. In another work by Krupinski *et al.* [103], involving six pathologists, a compression factor of up to 32 did not cause a noticeable difference in distinguishing benign from malignant cancer in breast tissue. At the same time, the authors reported that increasing the compression ratio to 64:1 affected the diagnostic performance significantly. Marcelo *et al.* [105] studied the accuracy of diagnosis and confidence level of 10 pathologists between



non-compressed and JPEG-compressed pathology images (reduced 90% in file size). The authors reported no statistically significant difference in diagnostic accuracy at the 95% confidence interval. Johnson *et al.* [106] reported a threshold of about 13:1 compression ratio for a human observer to discriminate using the JPEG 2000 compression versus employing uncompressed breast histopathological images. In the work of Pantanowitz *et al.* [104], a compression ratio of 200:1 was reported as an acceptable threshold for measuring the HER2 score in immunohistochemical images of breast carcinoma evaluated by a conventional image processing algorithm [109]. Lopez *et al.* used a cell-counting CAD system as a reference for statistical evaluation of cell-counting error in the uncompressed and JPEG-compressed histopathological images. The authors exploited three different compression ratios of 3, 23 and 46 and concluded that increasing the compression ratio deteriorated the performance of cell-counting in images. The authors concluded that the significant factors influencing the classification-performance degradation of a CAD system are the compression ratio and the intrinsic image complexity. According to their study, a more complex image is known as an image with a higher number of nuclei.

Although all the above works study the impact of lossy compression on diagnostic performance, they do not involve a complex model such as a deep ConvNet as a model observer. Furthermore, their experiments with a CAD observer are limited to training on the high-quality input data and evaluating both high- and low-quality image data, while they have not considered the performance of a model observer that was trained on low-quality input data.

*JPEG 2000 as gold standard compression:* JPEG 2000 [110] was introduced as a follow-up standard for JPEG (ISO/IEC 10918-1 — ITU-T Rec. T.81) bringing improved rate-distortion performance and additional functionality, such as resolution and quality scalability [111]. One of the main differences between JPEG 2000 and the JPEG algorithm is the exploitation of the discrete wavelet transform (DWT) instead of a block-based discrete cosine transform (DCT). In term of visual artifacts, JPEG 2000 produces soft *ringing* and *blocking* artifacts at high compression ratios, whereas JPEG generates both artifacts, but dominantly more visible the *blocking* artifacts [112]. Nonetheless, while both algorithms visually perform comparably for higher bit rates, at mid and lower bit rates, JPEG 2000 clearly outperforms JPEG in terms of rate-distortion performance [113]. With the use of the JPEG 2000 algorithm, it also becomes possible to store different parts of the same picture with different qualities, which makes it attractive for the compression of WSIs [114]. This is because approximately 80% of a WSI area contains an empty (white) background region [87] that does not contain any tissue. Helin *et al.* [115] showed that applying a very high degree of JPEG 2000 compression on the background part of WSIs and applying a conventional ratio of compression (e.g. 35:1) on the tissue-containing part results in a high overall compression ratio. An additional compression gain of up to a factor 3 was reported compared to conventional, non-adaptive compression with JPEG 2000.

With respect to natural images, some studies have assessed the impact of the quality of such images in terms of compression on the performance of a deep ConvNet [116, 117]. In the work of Dodge *et al.* [116], a VGG-16 network, which was trained on the ImageNet 2012 dataset [118], was found resilient to JPEG and JPEG 2000 compression up to a compression factor of 10 and down to 30 dB Peak Signal-to-Noise Ratio (PSNR), respectively. In similar work by Dejean *et al.* [117], again an experiment on the ImageNet dataset was performed, where a CNN showed only a drop of one unit on the classification ranking for object categorization after applying the compression (with the ratio of 16:1). Here, the classification ranking was defined by sorting (in descending order) the output probabilities of the assigned class labels, given an input image by the network. So, ideally, the true class should be recognized with rank one, while categorizing in any lower-ranking can be considered as a greater error in the classification performance.

*Considerations for our research:* Although these studies involve a ConvNet for the assessment of classification performance on compressed natural images, to the best of our knowledge, no similar study has been carried out on the histopathological images/WSIs. The outcome of such a study can be different from the obtained results for the natural images, as the histopathological image contents have a high inter-component correlation and can be processed differently.

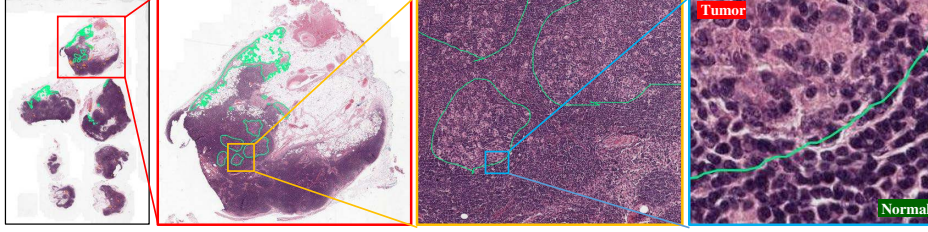
In this section, we investigate the impact of the compression ratio to evaluate the performance of a deep ConvNet applied to JPEG 2000-compressed histopathological WSI data. We employ a recently proposed CAD model that produces comparable performance to that of pathologists in detecting cancer metastases in breast lymph nodes. Since the proposed CAD system exploits a learning model, we study the impact of degradation in image quality due to compression, both by varying the quality (i.e. compression ratio) of the training data in the training phase, and by varying the quality of the test data during the testing phase. Such a study reveals the adaptivity of the network for preserving its high performance on low-quality images when its parameters are re-trained.

The following sections first introduce the data that is used in this study. Afterwards, the deep learning model is briefly explained. Then the experimental set-up and the obtained results are detailed. Finally, the results are discussed and conclusions are presented.

#### 3.4.2 Materials and Methods

##### A. Dataset

We use the CAMELYON16 dataset [52] for the experiments. This data set, which is the preceding version of CAMELYON17, contains tumor annotations at the pixel level (see Figure 3.12 for a visual example). The task in the CAMELYON16 challenge was detecting and segmenting the metastases in WSIs, while in CAMELYON17 the task was changed to the categorization of each detected metastatic



**Figure 3.12** Example WSI from the CAMELYON16 dataset. A histopathology WSI (left) of size  $220k \times 98k$  pixels is shown at multiple zoom levels. (Most right) An image patch of size  $300 \times 300$  pixels. The metastases have been delineated by the pathologists (green contours).

region into four types (i.e. grades), according to their area<sup>1</sup> of metastases. This can be considered as a post-processing stage to what was defined in CAMELYON16. In this study, we base our evaluation of the CAD performance on the task that was defined in the CAMELYON16 challenge, as it obtains more accurate quantified measures in comparison with the slide-level categorization task of CAMELYON17.

The CAMELYON16 dataset consists of WSIs having pixels acquired with a resolution equal to  $0.243 \mu m$ , collected from two clinical centers in the Netherlands. Originally, the dataset was split into a training and a testing set. The original training set consists of 111 WSIs with and 159 without metastases. The original testing set consists of 129 WSIs, 49 with and 80 without metastases. We have removed one slide (namely tumor slide number 114) from the testing set because it does not have an exhaustive annotation for all its metastasis regions, as was also mentioned by the data provider. For ground truth, the pixel-level annotation for the positive (i.e. containing tumor) WSIs was provided by a group of pathologists. The original WSIs were stored in the TIFF file format that was already compressed by the JPEG compression with 80% quality and 4:2:2 YCbCr chroma subsampling. The WSIs are stored in a pyramidal level structure with different factors of magnification. For the research in this chapter, we adopt the data level of  $20 \times$  magnification factor, since it has shown the highest performance for tumor detection [119]. In this study, the uncompressed high-quality data (also labeled with 1:1 ratio) refers to this dataset with the mentioned magnification factor. More details about this dataset can be found in the paper of Bejnordi *et al.* [52].

#### B. Data Sampling

Since processing all the regions inside a WSI is redundant and inefficient for training a deep learning model, a data-sampling stage is applied, which consists of

<sup>1</sup>For the used WSI data, the grading of the metastatic region is dependent of the area of the affected regions.

two parts: region of interest (ROI) detection and patch extraction. As mentioned earlier, about 80% of a WSI area contains an empty background region [87], which can be easily detected by using a conventional image processing technique such as Otsu thresholding [96]. By detecting the empty regions of each WSI, they are ignored for further analysis by the ConvNet. Because of the very large image-frame size of WSIs, directly using them as input to the network is impractical. A common approach is therefore the processing of image patches and employing the network as a patch classifier [120]. In a patch-classification approach, the input to the model is a patch image with predefined dimensions and the output is the predicted class of the central pixel inside the image patch (this is a common approach to transfer pixel-level labels to the patch level). After training the network on image patches, prediction on WSIs can be performed by sliding a window over the entire WSI and consequently predicting the central pixel of the window.

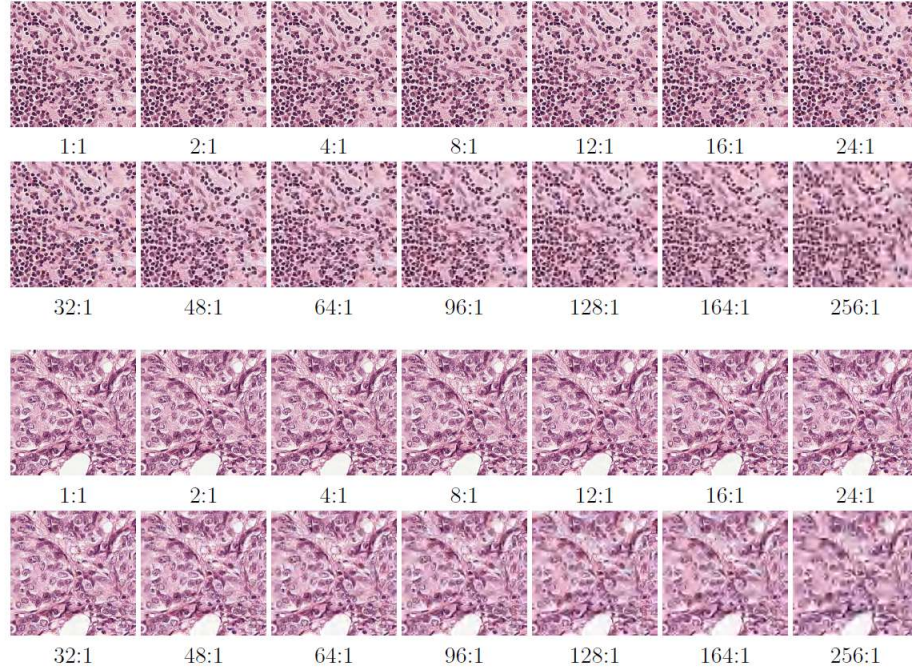
Training the model on all possible extracted patches is redundant and the population of samples between two classes would be highly imbalanced because in most cases only a small portion of the examined tissue contains tumor cells. For compensating the problem of highly imbalanced data in patch sampling, we only randomly select a limited number of negative patches, while all the extracted positive patches from the training set are used [121]. Here, the negative and positive patches refer to the patches that have been labeled as normal and tumor patches, respectively.

In total, 650,000 patches of size  $300 \times 300$  pixels are extracted. A patch is labeled as a positive sample (tumor) if more than 20% of its pixels are annotated as positive, otherwise, it is labeled as a negative (normal) sample. For better training of the model, the extracted patches are on-the-fly augmented (during training) by applying random rotation, using multiples of  $90^\circ$  (i.e. rotation angles of  $\{0, 90, 180, 270\}$  degrees). Afterwards, the images are also randomly chosen to be flipped or not. Flipping may be operated vertically, horizontally or in both directions.

For increasing the generalization ability of the classifier against minor changes in chromatic information, we apply color augmentation on the training image patches, which has become a common practice in training a deep ConvNet [52, 94]. This leads to training a wider range of color variations, compared to what typically occurs in the training set. As a result of this, we insert some noise into the lightness and saturation channels of the HSL color coordinates, by adding a (uniformly distributed) random value to the pixels of each patch (or subtracting a random value from them). The maximum magnitude of such additive noise is equal to 0.25% of the highest value of the channel (e.g.  $0.25 \times 255$  in the standard YCrCb system).

#### C. JPEG 2000 and Image Quality

We have deployed JPEG 2000 with 6 wavelet decomposition levels and 14 different compression ratios. The extracted patches from the data sampling stage are



**Figure 3.13** Examples of normal (top) and tumor (bottom) image patches compressed with JPEG 2000 at different compression ratios. Data patches are the highest resolution.

compressed. Figure 3.13 shows the quality of a normal and a tumor patch, both compressed at different compression ratios. The outcome of this trade-off between quality and compression factor is extensively discussed later in this chapter.

#### 3.4.3 Automated Tumor Detection

The recently proposed ConvNet-based model [96] is adopted as an automated cancer metastases detection system in breast lymph node WSIs. This model uses the *Inception-v3* [92] architecture, a 48-layer deep network, as patch classifier. The input data to the model are full-color RGB image patches and its output is a 2-element vector with one-hot encoding, representing a binary classification. For speeding up the training, the parameters of ConvNet are initialized by using the parameters which have been trained on the ImageNet 2012 dataset [118]. The Inception-v3 architecture has shown to have a better performance for image classification with a much lower number of parameters, compared with its preceding versions, due to its convolution factorization strategy [92]. In computational pathology, this model has shown human-level performance in detecting tumor cells [119] and has won the 2<sup>nd</sup> place in the CAMELYON17 challenge [85, 94],

which includes the CAMELYON16 dataset used in this study. As a state-of-the-art deep learning baseline, we have employed it as an observer for studying the impact of image compression on the performance of such a model.

#### 3.4.4 Experiments and Evaluation

##### A. Experiments

First, ROI selection and patch extraction are performed on the training and test WSIs, as described earlier. About 150,000 positive samples are extracted from the regions with metastases in positive WSIs, according to the provided ground truth. For compensating the problem of severe class imbalance in samples, we have extracted only about 500,000 negative samples from the normal regions of the negative and positive WSIs. Afterwards, these samples are compressed by the JPEG 2000 algorithm using 14 different compression ratios.

The impact of changing the compression ratio of JPEG 2000 on the performance of the network is extensively evaluated for the following three distinct scenarios.

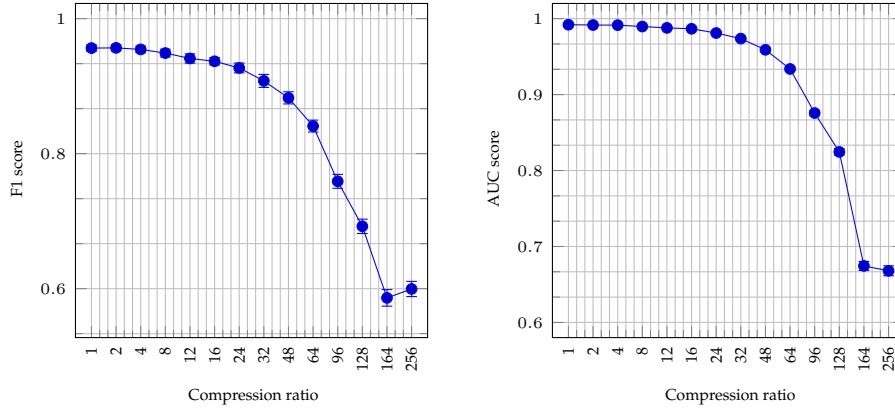
- *Scenario 1 (telemedicine)*: The CAD system is trained on high-quality uncompressed images and is evaluated on compressed low-quality images with several compression ratios.
- *Scenario 2 (cloud computing)*: The CAD system is trained and tested on the same level of compression. For example, if the model is trained on images compressed by a factor of 32, it is also evaluated on images that are compressed by that factor.
- *Scenario 3 (medical center communication)*: The CAD system is trained on images compressed with the maximal compression ratio that still allows the classification performance to be above a predefined threshold (e.g. less than 10% drop from the maximum  $F_1$  score, which is obtained when using uncompressed data). Thereafter, it is evaluated with test images with both lower and higher compression ratios.

The first scenario is highly applicable to telemedicine and in particular *telepathology* [122], where e.g. a primary diagnosis can be obtained by transmitting the compressed images to a remote CAD system. Such a remote CAD system can have already been trained on high-quality input data. The second scenario is more relevant in cloud-based computing and training, where several pathology labs share their data to a remote server for training and evaluation. The third scenario is valid for a case where a powerful computation engine is locally available, e.g. exploiting a supercomputer in a clinical institute or large hospital, so that the transmission of high-quality images is not an issue internally, but utilizing external images from remote data sources for training still has limitations due to transmission bandwidth constraints.

### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS

**Table 3.5**  $F_1$  scores, AUC values, and their 95% upper ( $\uparrow$ ) and lower ( $\downarrow$ ) bounds confidence intervals (CI) for the CAD model, tested on the images with different compression ratios, when is trained on the uncompressed (1:1) high-quality training images (Scenario 1).

Metric		JPEG 2000 compression ratio														
		1:1	2:1	4:1	8:1	12:1	16:1	24:1	32:1	48:1	64:1	96:1	128:1	164:1	256:1	
$F_1$ score	CI $\uparrow$	0.954	0.954	0.952	0.946	0.938	0.934	0.923	0.903	0.878	0.836	0.754	0.687	0.580	0.594	
	Mean	<b>0.956</b>	<b>0.957</b>	<b>0.954</b>	<b>0.949</b>	<b>0.941</b>	<b>0.937</b>	<b>0.927</b>	<b>0.908</b>	<b>0.883</b>	<b>0.841</b>	<b>0.759</b>	<b>0.692</b>	<b>0.586</b>	<b>0.600</b>	
	CI $\downarrow$	0.959	0.959	0.957	0.952	0.944	0.940	0.931	0.913	0.887	0.845	0.764	0.698	0.592	0.605	
AUC	CI $\uparrow$	0.991	0.991	0.991	0.989	0.987	0.985	0.980	0.972	0.957	0.931	0.872	0.820	0.668	0.661	
	Mean	<b>0.992</b>	<b>0.992</b>	<b>0.991</b>	<b>0.990</b>	<b>0.988</b>	<b>0.987</b>	<b>0.981</b>	<b>0.974</b>	<b>0.959</b>	<b>0.934</b>	<b>0.876</b>	<b>0.825</b>	<b>0.674</b>	<b>0.668</b>	
	CI $\downarrow$	0.993	0.992	0.992	0.991	0.989	0.988	0.983	0.976	0.961	0.936	0.879	0.829	0.680	0.675	

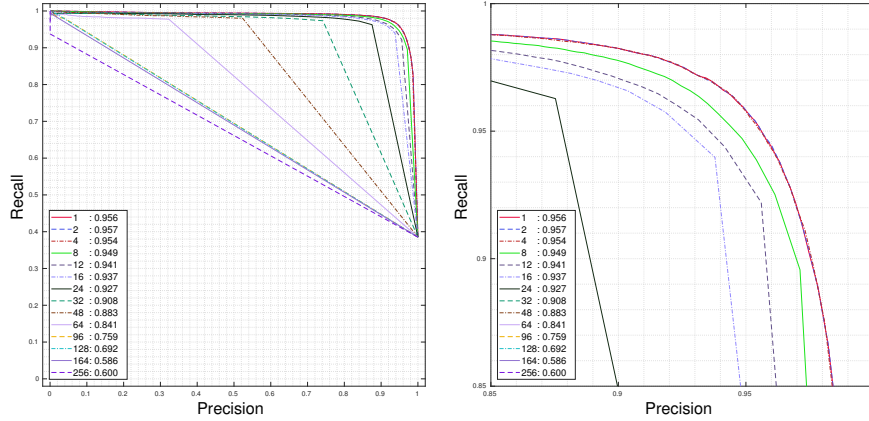


#### B. Evaluation method

The performance of the binary classification between tumor and normal image patches is evaluated by reporting the  $F_1$  score and area under the receiver operating characteristic (ROC) curve, called AUC. Evidently, the configuration of a CAD system with a higher AUC and  $F_1$  score represents better performance.

Since the discrimination threshold of the binary classification system is varied, we report the diagnostic capability of the system with a complementary measurement of the precision-recall (PR) curve. In comparison with alternative measures such as the ROC curve, a PR curve can better expose the differences between algorithms, especially when highly-skewed cancer detection data are studied [123]. The PR curve visualizes the performance of a classifier by ignoring the true negative samples. This property highlights well the change in classification performance when imbalanced data are processed. It is worth mentioning that even if the training set contains an equal number of patches per class, the data originally are considered imbalanced, since the area of tumor region is often smaller than the normal region in a pathology slide.





**Figure 3.14** Evaluation of the CAD model for Scenario 1; (left) precision-recall (PR) curves; (right) enlarged view on the PR curves. The  $F_1$  scores are shown in the legend of the graphs.

### 3.4.5 Results

#### Scenario 1) Train on high- and infer on low-quality data

In this experiment, the network is trained on uncompressed images for a fixed number of iterations equal to 10,000. Afterwards, we evaluate its performance on the test set, which has been compressed with 14 different compression ratios, including the original uncompressed test images (compression ratio 1:1). The obtained  $F_1$  score and AUC values are depicted in Table 3.5 and the PR curves are plotted in Figure 3.14. As expected, by degrading the image quality due to increasing the compression ratio, the model performance is decreased. It can be observed that up to a factor of 24, the performance does not show considerable changes, but for a ratio of 32:1, the  $F_1$  score drops to 0.908. As the  $F_1$  scores and the PR curves illustrate, a factor of 24 presents a decent trade-off between performance and compression. The deterioration of the performance is even more clear by inspecting the PR curve in Figure 3.14 where compression with more than factor 24 drops the classification performance significantly.

#### Scenario 2) Training and doing inference on the same quality

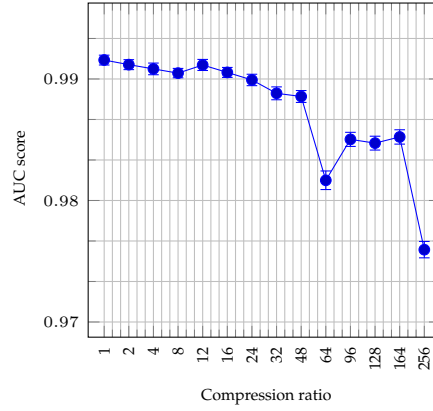
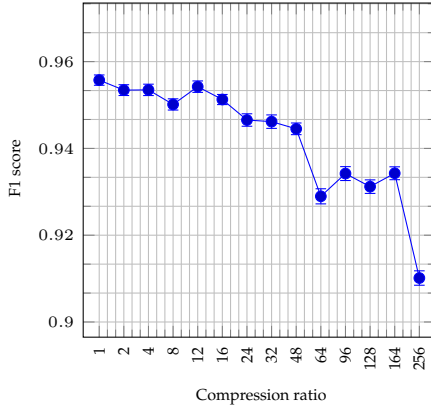
In this experiment, the network is trained several times. Each time, the training images are compressed with a different ratio. Afterwards, the model is evaluated on the test set, which has been compressed with the same ratio as applied to its corresponding training set. Table 3.6 and Figure 3.15 show the obtained results. The outcome drastically differs from the previous experiment. As can be observed, the performance of the network is more tolerant against the compression ratios, implying that a ConvNet can be trained to handle larger compression ra-



### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS

**Table 3.6**  $F_1$  scores, AUC values, and their 95% upper ( $\uparrow$ ) and lower ( $\downarrow$ ) bounds confidence intervals (CI) for the CAD model, tested on the compressed images which have been compressed with the same ratio as applied to the training images (Scenario 2).

Metric		JPEG 2000 compression ratio														
		1:1	2:1	4:1	8:1	12:1	16:1	24:1	32:1	48:1	64:1	96:1	128:1	164:1	256:1	
$F_1$ score	CI $\uparrow$	0.955	0.952	0.952	0.949	0.953	0.950	0.946	0.945	0.943	0.927	0.933	0.930	0.933	0.908	
	Mean	<b>0.956</b>	<b>0.953</b>	<b>0.953</b>	<b>0.950</b>	<b>0.954</b>	<b>0.951</b>	<b>0.947</b>	<b>0.946</b>	<b>0.944</b>	<b>0.929</b>	<b>0.934</b>	<b>0.931</b>	<b>0.934</b>	<b>0.910</b>	
	CI $\downarrow$	0.957	0.954	0.955	0.952	0.956	0.953	0.948	0.947	0.946	0.930	0.936	0.933	0.936	0.912	
AUC	CI $\uparrow$	0.991	0.991	0.990	0.990	0.991	0.990	0.989	0.988	0.988	0.981	0.984	0.984	0.985	0.975	
	Mean	<b>0.992</b>	<b>0.991</b>	<b>0.991</b>	<b>0.990</b>	<b>0.991</b>	<b>0.991</b>	<b>0.990</b>	<b>0.989</b>	<b>0.989</b>	<b>0.982</b>	<b>0.985</b>	<b>0.985</b>	<b>0.985</b>	<b>0.976</b>	
	CI $\downarrow$	0.992	0.992	0.991	0.991	0.992	0.991	0.990	0.989	0.989	0.982	0.986	0.985	0.986	0.977	

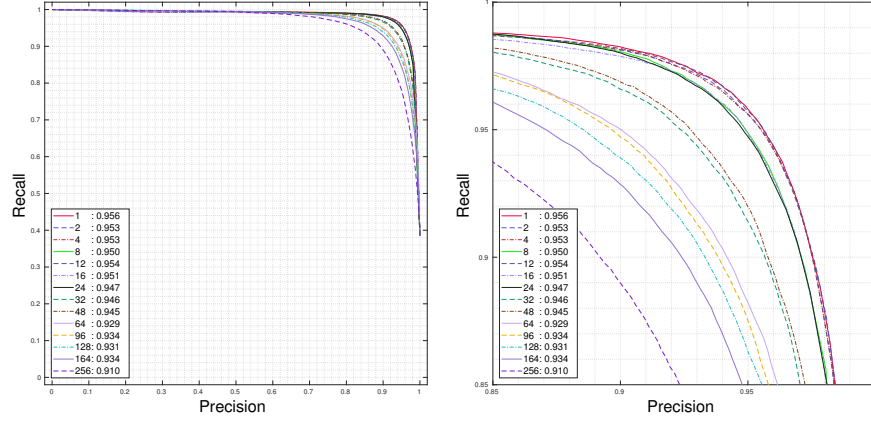


tios. The differences between the performances of the model under different compression ratios are minimal. In comparison with the performance of the model in the previous experiment (Scenario 1), the improvement is significant per compression factor. For example, the  $F_1$  score for compressed images with the factor of 164 is equal to 0.934 whereas when the model is trained on high-quality images, its  $F_1$  score is only 0.586. This represents about 59% improvement in the  $F_1$  score. A possible explanation for such a strong improvement is the adaptation of the network parameters to the distortion and degradation of the image quality, which are broadly present in its training set.

#### Scenario 3) Making inference on varying-quality images with fixed compressed trained images

In this experiment, the performance of the model, which is trained on images compressed with a factor of 48, is evaluated on a compressed test set with various compression ratios as well as uncompressed images. The compression ratio of 48:1 has been selected, since it shows a maximum compression ratio where the  $F_1$  score of the model drops less than 10% of its maximum, according to the previous experiment (Scenario 2). As can be observed from Table 3.7 and Fig. 3.16, the results improve for the higher compressed images (higher than 48), compared with the first experiment when the model is trained on uncompressed

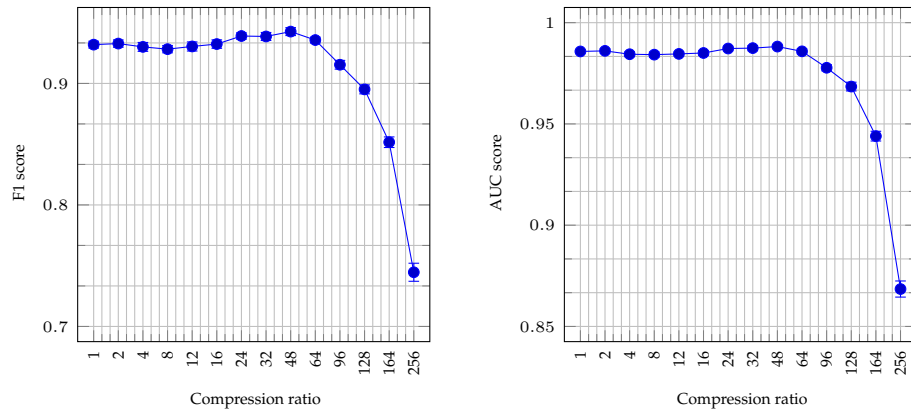
### 3.4. Impact of JPEG compression on deep neural networks



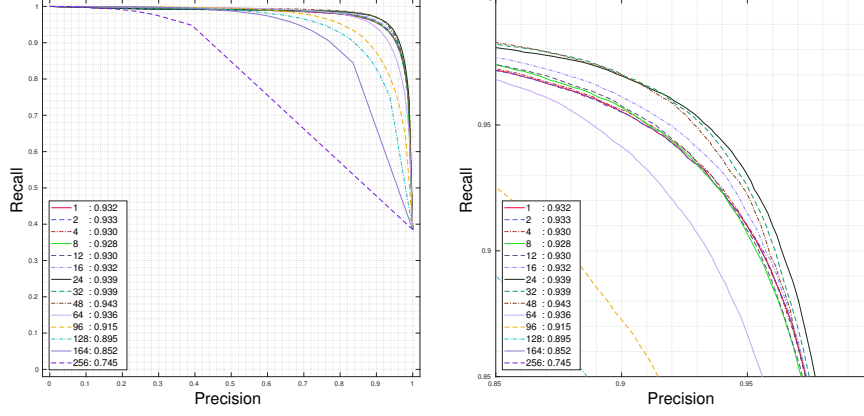
**Figure 3.15** Evaluation of the CAD model on compressed test images with the same compression ratio as applied to the training images; (left) precision-recall (PR) curves; (right) enlarged view of the PR curves at the bending position. The  $F_1$  scores are shown in the legend of the graphs. In contrast with Scenario 1, the drop in the performance of the network by increasing the compression ratio is very smooth and does not portray distinct turning point(s). The choice remains on the absolute quality level to be sufficient for clinical cancer detection.

**Table 3.7**  $F_1$  scores, AUC values, and their 95% upper ( $\uparrow$ ) and lower ( $\downarrow$ ) bounds confidence intervals (CI) for the CAD model tested on the images with different compression ratios while it was trained on 48:1 compressed training images (Scenario 3).

Metric		JPEG 2000 compression ratio														
		1:1	2:1	4:1	8:1	12:1	16:1	24:1	32:1	48:1	64:1	96:1	128:1	164:1	256:1	
$F_1$ score	CI $\uparrow$	0.929	0.930	0.926	0.925	0.927	0.929	0.936	0.935	0.939	0.933	0.912	0.891	0.847	0.737	
	Mean	0.932	0.933	0.930	0.928	0.930	0.932	0.939	0.939	0.943	0.936	0.915	0.895	0.852	0.745	
	CI $\downarrow$	0.935	0.936	0.934	0.931	0.934	0.936	0.942	0.942	0.946	0.938	0.919	0.899	0.856	0.752	
AUC	CI $\uparrow$	0.985	0.985	0.983	0.983	0.984	0.984	0.986	0.986	0.987	0.985	0.976	0.966	0.942	0.864	
	Mean	0.986	0.986	0.984	0.984	0.985	0.985	0.987	0.987	0.988	0.986	0.978	0.968	0.944	0.868	
	CI $\downarrow$	0.987	0.987	0.986	0.985	0.986	0.986	0.988	0.989	0.989	0.987	0.979	0.971	0.946	0.872	



### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS



**Figure 3.16** Evaluation of the CAD model which was trained on 48:1 compressed images, while tested with variable compressed images; (left) precision-recall (PR) curves; (right) enlarged view on the PR curves. The  $F_1$  scores are shown in the legend of the graphs.

images. The reason can be similar to what has been observed in Scenario 2 because the system has learned the compression artifacts from the training samples. In comparison with Scenario 2, the performance slightly decreases at both sides of the trained compression ratio. In a nutshell, from the results, it can be observed that a trained ConvNet on the low-quality images, e.g. with a compression ratio of 48:1, can perform almost equally well on all higher-quality images and even on the slightly lower-quality samples. Furthermore, in the PR curves with the magnified view, the performance curve are very close to each other around the ratio of 48.

#### 3.4.6 Discussion

Compression of histopathology images has not yet been approved by regulatory agencies in the USA for clinical applications. The contribution of this work is that the investigation provides evidence that compression may be used from the CAD point of view, but this does not involve a guarantee for the clinical acceptance of using compression. We will discuss several aspects of these experiments and its possible acceptance.

*The nature of experiments:* In this investigation, the aim is to evaluate the impact of JPEG 2000 compression on the diagnostic performance of a deep ConvNet model for detecting metastases in breast lymph node histopathological WSIs. Hence, three series of experiments have been set up.

*Scenario 1:* When employing uncompressed high-quality images for training the deep neural network, the performance drops significantly for a compression ratio of 48:1 and higher, but it does not change significantly for the ratio of 24:1 and lower. Our findings of the robustness of the ConvNet against JPEG 2000

compression is in near agreement with the cut-off quality threshold of 32:1 ratio in the work of Krupinski *et al.* [103], which involved pathologists as observers for classifying benign versus malignant breast cancer in WSIs. However, this work did not examine any effects on human performance in cancer detection when reviewing such images.

*Scenario 2:* Training and predicting on equal-quality images drastically produces better results compared with the previous scenario in which the model only has been trained on uncompressed data. The outcome is remarkably improved for high compression ratios, whereas it does not change for low compression ratios. As an example of such an improvement, the performance of the model on compressed images with a factor of 164 is on par with the results of the previous experiment with a factor of 24. This mainly happens because the ConvNet parameters have been optimized by observing the low quality (distorted) training images. Hence, it can be robust to some extent to the presence of compression artifacts.

*Scenario 3:* Finally, we have empirically shown that training the networks on 48:1 compressed images increases the performance for somewhat lower and higher compression ratios. These findings can help in designing a more efficient CAD system, mainly when a constraint exists for transmission and storage, such as in a system with a cloud-based computation or telepathology. Also, we have shown that for better training of the ConvNet model, the availability of high-quality uncompressed images is not a necessity.

*Generalization of the applied CAD:* Generally speaking, given the robustness and learning capability of various networks, it is expected that the obtained results will not largely change if a different network would have been used. We expect that employing a different convolutional-based network architecture does not change the trend of variations in the experiment. However, depending on the learning capability of the network, the absolute scores might be somewhat different.

*Clinical acceptance:* Clinical diagnostics of cancer in pathology images are based on human perception. Previous studies have broadly investigated the robustness of human observers against appearing distortion resulting from applying compression techniques and in particular JPEG compression. However, this study has been concentrating on the CAD as observer, trying to diagnose the cancer using a neural network. This inherently can result in different findings and conclusions. The clinical acceptance of this study has not been evaluated and is not part of the current study. Similar to the acceptance of CAD algorithms [124] by clinicians, this work should be extended by several thoroughly conducted clinical experiments where a multitude of clinicians of variable experiences are participating to compare the diagnosis results of differently (un)compressed data.

### 3.5 Conclusions

In this chapter, we have studied three important problems in automated histopathology, including stain-color normalization, metastases detection and

### 3. TWO-DIMENSIONAL HISTOPATHOLOGICAL IMAGE ANALYSIS

---

their classification, and the impact of compression-aware neural network training and inference. For each of these problems, we have proposed individual deep learning-based solutions that advance the borders of this field by outperforming state-art-the-art methods. We have gathered the following findings related to these problems.

*Stain-color normalization:* In Section 3.2, we have discussed that stain-color variations can be critical when the color profile of the test sample deviates too much from the training set. The color variations are due to many factors, such as different characteristics of histopathology scanners, different staining procedures among laboratories, etc. We have shown that the proposed generative neural models are able to reduce the color variations between samples and this led to a higher likelihood between the training and test sample distributions. Such adaptation of sample distributions is considered as *domain adaptation*, which is known to reduce the gap between the color profiles of training and test samples and consequently increases classifier performance in the inference of unseen examples.

At the time of performing this research, the work was to proposed the first GAN-based approach to color normalization that inspired several follow-up studies.

*Cancer detection and classification:* In Section 3.3, cancer detection and segmentation as another important problem in computational pathology is studied. We have proposed a cancer detection and classification framework, based on recent advances in deep learning architectures and graphical modeling. Because of the huge size of whole slide images, patch classification is a common approach in data-driven methods. We have shown that modeling dependencies between patches by using conditional random fields, combined with updating the predictions of a neural network on image patches, can increase the overall performance in metastases detection and categorization. This is mainly due to broadening the decision-making of the model over all patches and not only the individual prediction. Here, expanding the field of view of the model is performed through a fully-connected graph that is constructed and optimized over all patches in the image.

The proposed technique has proven to be effective on achieving a rather high performance in pathology metastases detection when tested on a large-scale dataset. One of the main reason of this high performance score is the full exploitation of visual contextual information in WSIs. This result has obtained the second place in the international CAMELYON17 competition on a large-scale dataset, including samples from several laboratories.

*Histopathology image compression:* Lastly, in Section 3.4, we have studied the impact of a state-of-the-art image compression technique (namely JPEG 2000) on the performance of cancer detection by a neural network. The compression aspect is hardly explored for this domain and is therefore highly relevant for

designing a CAD system. Due to the enormous size of WSIs, the compression cannot be avoided for storage or transfer of histopathology scans via a communication channel. We have conducted three series of experiments, referring to different scenarios in training of or inferencing with a CAD model across various compression ratios. The empirical studies imply that training on high quality images and testing on compressed data results in decreasing the performance significantly when the compression factor exceeds 24. This low tolerance of the model against compression is mainly explained by absence of distortion artifacts in the training set, so that the parameters of the model are not adapted to such effects. However, when the training does include the same amount of compression artifacts that exist in the test data, it shows a constantly high performance across various compression ratios. Lastly, we have shown that compressing the data with a factor of 48, the model is considerably tolerant to increase the compression ratios to higher values above 48, as well as decreasing the compression ratios below 48 with a reasonable range.

The next chapter will address a study on a 3D convolutional network for needle detection in ultrasound volumes. This study is an extension of employing orthogonal 2D convolutional operators to a 3D volume for efficiency in computation.



## Needle Localization in Volumetric 3D Ultrasound Imaging

### 4.1 Introduction and related work

The previous chapters have discussed data-driven model development for one-dimensional and two-dimensional medical imaging data. This chapter represents a contribution to 3D volumetric Ultrasound (US) imaging.

Ultrasound imaging is broadly used to visualize and guide the interventions that involve the percutaneous advancing of a needle to a specific tissue or blood vessel vicinity inside the patient body. However, for a typical 2D US system, bi-manual coordination of both the needle and the US transducer is required and quite challenging, since the limited US field of view obscures the visualization of the complete needle, while an inadequate view leads to an erroneous placement of the needle tip. Therefore, while advancing the needle, continuous manipulation of the transducer is necessary to search for the needle in the imaging data and identifying and visualizing the best needle plane. As an alternative, 3D US transducers with an image-based needle-tracking system can overcome these limitations and minimize the manual coordination, while preserving the use of a conventional needle, signal generation and transducers [125]. In such a system, the needle is conveniently placed in the larger 3D US field of view and the processing unit automatically localizes and visualizes the entire needle by adopting the best planar view. Therefore, the manual skills are significantly simplified when the entire needle remains visible in the visualized plane, after the needle is further advanced or the transducer is moved slightly.

Several image-based needle localization techniques have been proposed, based on maximizing the intensity over parallel projections [126]. Due to the complexity of realistic data, methods that solely rely on the brightness of the needle are not robust for localizing thin objects in a cluttered background. There-



fore, information regarding the line-like structure of a needle is used by Hessian-based line filtering methods [127]. Although shown to be limited in localization accuracy [125], they can be beneficial for reducing the imaging artifacts. Other techniques involve exploiting the intensity changes caused by needle movement to track the needle in the US data [128]. Nevertheless, large movements of the transducer or the patient will increase the difficulty of motion-based tracking and therefore, we aim at repeated detection of the needle in static 3D volumes. When realizing the real-time operation, tracking of the needle is implemented by repeated detection with sufficient time resolution. This will result in detection per volume in a 4D US sequence, which allows for arbitrary inter-volume movements.

More recently, attenuation of the US signal due to energy loss beyond the US beam incident with the needle (i.e. the shadow behind/below the needle), is used to detect the position of the needle [129, 130]. However, signal loss due to the presence of other attenuating structures may degrade the accuracy of estimation and should be explicitly handled. Alternatively, supervised needle-voxel classifiers that employ the needle shape and its brightness have shown to be superior to the traditional methods [125]. Nevertheless, since the needle is assumed to be already inserted in the volume up to a considerable length, the operator typically does not achieve high detection precision and therefore cannot localize the needle when it is partially inserted into the volume. Moreover, when the target tissue is deep, the degraded resolution and possible needle deflections further complicate the interpretation of data and reduce voxel classification performance, which should be addressed by better modeling of both local and contextual information.

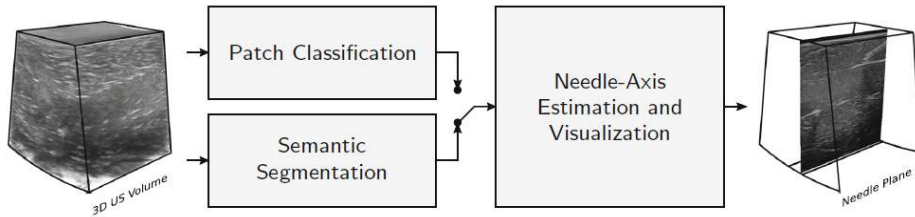
This chapter shows how the Convolutional Neural Networks (ConvNets) can achieve a substantial improvement in the detection accuracy of needle voxels in 3D US data by better learning and extracting discriminating features for this task [131]. Afterwards, adopting a patch classification in this framework can further improve the US data segmentation for needle localization. When using sector, curved and phased-array transducers in US imaging, the insonification angle of the US beams is changed throughout the volume, which creates varying angles with different parts of the needle. Therefore, the needle can be partially invisible, due to the lack of received US reflections from parts of the needle. The missed data enforces a trade-off between patch size for richer needle context information against localization accuracy. In ConvNet standard architectures, larger patches require more down-sampling layers (e.g. through max-pooling) that reduce the localization accuracy, while small patches allow the network to infer only parts of the needle.

As an alternative to patch training, semantic segmentation methods can generate dense prediction maps, by learning a transformation between input image and the target annotation map. Examples of such networks are fully-convolutional networks (FCN) [132, 133], and/or context modeling by employing atrous convolutions [134, 36]. Although integrating atrous (or dilated) convolutions in the deep layers of the network increases the field of view while preserving the relatively high spatial dimensions, performing convolution on a large

number of such high-dimensional feature maps is computationally expensive and memory costly. However, original FCN architectures can simultaneously exploit the global and local information in the data and remain more memory-efficient by introducing skip connections from higher-resolution feature maps of encoder to the decoder layers. Initial attempts of applying these networks on US data are presented for fetal heart segmentation in 2D US [135]. Further improvement is shown for segmentation of fetus, gestational sac, and placenta in 3D US volumes by integrating the sequential information [136]. However, first a drawback of using such a 3D+time model exponentially increases the computational complexity of the 3D convolution operations. Secondly, a very large dataset is required for training the increased number of network parameters. Thirdly, the sequential modeling will be sub-optimal in the early time intervals after large movements of the transducer or subject.

The main contributions of this chapter are as follows.

- A new approach for segmentation and localization of partially inserted and partly invisible needles in 3D US data, using ConvNet models.
- An original update strategy for ConvNet parameters using the most aggressive non-needle samples, which significantly improves the performance on highly imbalanced datasets.
- A new method for modeling 3D US context information using 2.5D (thick slice) data, which enables an accurate training of the network using limited training samples.
- Extensive evaluation of the proposed methods on two types of ex-vivo data giving a very high average of 81.4% precision at 88.8% recall rate.



**Figure 4.1** Block diagram of the proposed needle localization frameworks in 3D US data

## 4.2 Deep learning methods for needle localization

### 4.2.1 Proposed frameworks

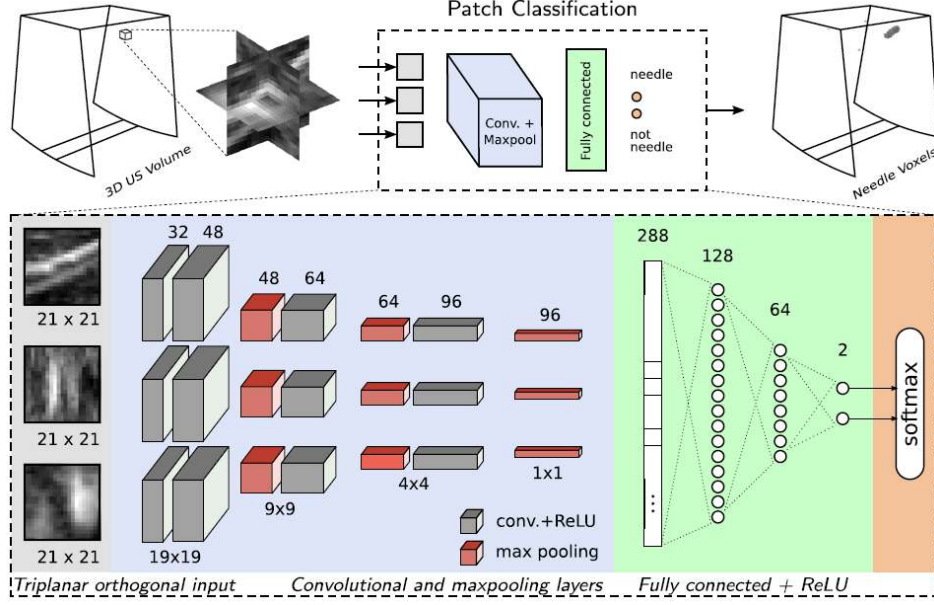
This chapter introduces two deep learning frameworks for localization and segmentation of needle voxels in 3D US volumes. The first framework is based on 3D patch classification. For efficiency in computations, instead of processing the

3D patches by using 3D kernels, the triplanar orthogonal views of each 3D patch are processed by the network. In the second framework, we extend such a multi-view analysis from patch-level into slice level. By processing multi-slice including both lateral and elevational planes in the volume, the network segments each slice in its output binary maps. The block diagram of the proposed needle localization CAD is shown in Figure 4.1. Apart from which deep learning framework is employed, for estimating the needle axis, a predefined geometric model is fitted to the position of positively-detected voxels in the volume. For clarity, we emphasize here that we detect the plane where the needle and its tip are maximally visible but do not explicitly detect the needle tip. This localization processing is done for every data volume individually. For a 3D+time US sequence, this would effectively mean repeated detection for every volume or image.

This chapter is organized to introduce and investigate each of these frameworks separately and to present the proposed solutions in detail. To this end, each framework is elaborated in a separate subsection of this chapter. In Subsection 4.2.2 introduces a 3D patch classification method as a primary technique for needle localization in the 3D US volume. Subsection 4.2.3 presents a segmentation approach as a secondary technique for needle localization, which explores the 3D volume by processing the sampled multi-view image planes of the 3D volume. In Section 4.2.4, a robust fitting of a geometric model to the detected voxels is explained which serves as a post-processing step. In Section 4.3, the conducted experiments and the obtained results for both deep learning frameworks as well as axis estimation performance on different datasets are presented. In Section 4.4 and Section 4.5, the chapter is finalized and the observations are discussed to conclude this study.

##### 4.2.2 Patch classification - primary technique

The block diagram of the proposed patch classification technique is shown in Figure 4.2. A ConvNet is trained to robustly classify the needle voxels in the 3D US volumes from other echogenic structures, such as bones and muscular tissues. Our voxel classification network predicts the label of each voxel from the raw voxel values of local proximity. In a 3D volume, this local neighborhood can simply be a 2D crosssection in any orientation, multiple cross-sections, or a 3D patch. Here, we use three orthogonal cross-sections centered at the reference voxel, which is a compromise with respect to the complexity of the network. The size of triplanar crosssections is chosen based on the diameter of a typical needle (0.7–1.5 mm), voxel size, and spatial resolution of the transducer, to contain sufficient context information. We extract triplanar cross-sections of  $21 \times 21$  pixels ( $4.2 \times 4.2$  mm), which provides sufficient global shape information and remains spatially accurate. For low-frequency transducers, more context is required for discriminative modeling, as the structural details of a needle will be distorted at low spatial resolutions.



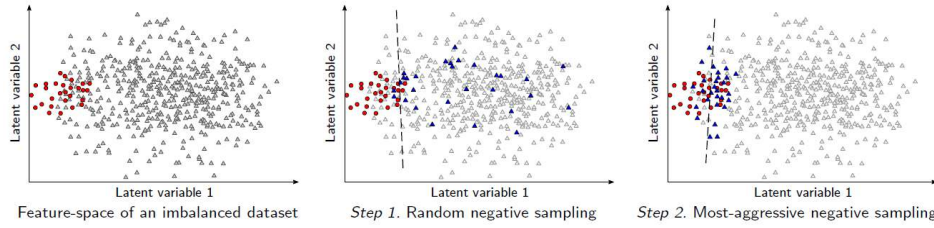
**Figure 4.2** Block diagram of the patch classification approach using CNN

*Network architectures:* For our experiments, we evaluate two CNN architectures based on shared convolutional (SharedCNN) and independent convolutional (IndepCNN) filters. In SharedCNN, a single convolutional filter bank is trained for the three input planes to have the same set of filters for all the planes. In IndepCNN, three sets of filter banks are trained independently, each to be convolved with one of the three planes. As depicted in Figure 4.2, both architectures consist of four convolutional layers having 32, 48, 64 and 96 filters of  $3 \times 3$  kernel size, three fully connected layers having 128, 64 and 2 neurons, and one softmax layer. According to the given number of filters, SharedCNN and IndepCNN architectures have 2160 and 6480 parameters in their convolutional layers, respectively. In both architectures, extracted feature maps after the last convolutional layer are concatenated prior to the fully connected layers [137].

*Training:* Our dataset is significantly imbalanced due to the small size of a needle compared to the full volume, i.e., approximately only 1 voxel out of 3000 voxels in a volume belongs to the needle. This is common in the representation of an instrument in 3D US volumes. Therefore, to avoid a prediction bias toward the majority class, we downsample the negative training data to match the number of needle samples. For an informed sampling of the negative (non-needle) set, we propose an iterative scheme based on bootstrapping [138] to achieve the maximum precision. In the first step, we train our network with uniformly sampled needle and non-needle patches. Training patches are rotated arbitrarily by  $90^\circ$  steps around the axial axis to improve the orientation invariance. The trained

network then classifies the same training set for validation. Finally, misclassified false positives are harvested as the most-aggressive non-needle voxels, which are used to update the network. Figure 4.3 shows how the iterative and informed sampling can increase the precision of the network. It is worth mentioning that commonly used methods for imbalanced data, like weighted loss function, do not necessarily improve precision. For example, the majority of our negative set consists of *easy* samples that can be classified beyond the model’s margin and will influence the loss function in their favor.

The CNN parameters are trained using stochastic gradient descent (SGD) and the categorical cross-entropy cost function. All activation functions are chosen to be rectified linear units (ReLU). Furthermore, for optimization of the network weights, we divide the learning rate by the exponentially weighted average of recent gradients (RMSProp) [139]. Initial learning rates are chosen to be  $1 \times 10^{-4}$  and  $1 \times 10^{-5}$  for train and update iterations, respectively. To prevent overfitting, we implement the dropout approach [140] with a probability of 0.5 in the first two fully connected layers. The trained network computes a label per voxel indicating whether it belongs to the needle or not.



**Figure 4.3** Example of the iterative sampling strategy to increase the precision of the network. The red circles represent the positive data points, gray and blue triangles are the negative and sampled data points, respectively. The dashed line represents the decision boundary of a classifier.

#### 4.2.3 Semantic segmentation - secondary technique

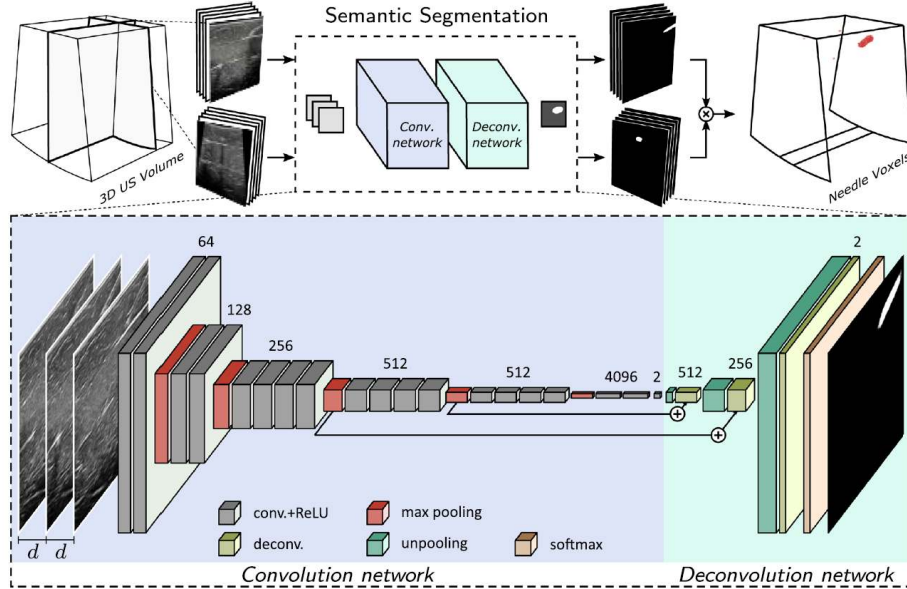
As discussed earlier, semantic segmentation of a needle using FCN architectures is advantageous compared to patch classification because the context information is modeled, while the spatial dimensions are preserved. Furthermore, in contrast to the patch-based methods, where redundant processing of voxels is inevitable, FCN models are more computationally efficient, since they exploit the one-time extracted features to simultaneously label all the data points using deconvolutional networks.

Figure 4.4 depicts the architecture of the proposed semantic needle segmentation technique in 3D US volumes. The proposed method is based on decomposing the 3D volume into 2D cross-sections for labeling the needle parts and reconstructing the 3D needle labels from multiple views. Therefore, in the proposed approach, the number of parameters in the convolution kernels decreases exponentially compared to the 3D kernels, so that the network requires fewer

training samples and executes faster. The proposed strategy for selecting the cross-sections is now presented below.

The 2D cross-sections are selected in multiple directions while all being perpendicular to the transducer with the step size equal to the voxel size. Since in a 3D US volume, the needle can enter the field of view from either the lateral or elevational directions, we consider cross-sections perpendicular to these axes. The segmentation outcome of each cross-section is mapped onto its corresponding position in 3D. Afterwards, the resulting probability volume from the two directions is combined using multiplicative averaging to create the final labeling outcome in 3D.

To exploit the 3D structural information in the proposed model, instead of only using 2D planar data, we opt for processing the consecutive cross-sections before and after the processing plane as additional inputs to the network. This is motivated by leveraging the spatial dependencies among neighboring cross-sections. In this study, we have added two additional cross-sections and have evaluated several spacing gaps  $d$ , between them. Therefore, as shown in Figure 4.4, a 3-channel input to the network is formed from the 2.5D (thick slice) US data at a specific position, which is used to create a 2D segmentation map of the corresponding cross-section.



**Figure 4.4** Block diagram of the semantic segmentation approach using FCN. The top part of the figure depicts the high-level architecture of the network, while the bottom part shows the detailed architecture of the semantic segmentation network. The numbers in the bottom figure indicate the size of feature map in different layers.

*FCN architecture:* Figure 4.4 depicts the FCN architecture used in the proposed system, comprising two stages of convolution and deconvolution networks. Inspired by SharedCNN, we use shared convolutional filters (SharedFCN) for both lateral and elevational planes. The convolution network is identical to the design of the VGG 19-layer ConvNet [141]. The deconvolution network consists of three unpooling masks of 2, 2 and 8 pixels with convolution layers having 512, 256 and 2 filters, respectively, with  $3 \times 3$  kernel size, and 1 softmax layer. Therefore, the receptive field of the network is equal to a window of  $96 \times 96$  pixels, which is equivalent to approximately  $19.2 \times 19.2$  mm for the high-resolution VL13-5 transducer and  $34.5 \times 34.5$  mm for the low-resolution X6-1 transducer, achieving a large context modeling at the same inference resolution of the input data. Convolution layers are stacked together followed by an activation function. As discussed, the network takes a 3-channel 2D input and the output layer indicates the class-assignment probability for each pixel in the corresponding cross-section.

*Training:* The training set consists of 3-channel cross-sections extracted with a gap of  $d$  mm in both elevational and lateral directions. The training volumes are augmented by 10 arbitrary rotations around the axial  $z$ -axis prior to extraction of the cross-sections. Therefore, several views of the needle are exploited to train the network, including in-plane, out-of-plane and cross-sections with partial visibility of the needle. Similar to the approach presented in subsection 4.2.2, we downsample the negative training data to match the number of cross-sections from the needle. These negative samples are the sections which do not contain the needle, so that the training samples are being balanced. However, since the initial training samples are not highly imbalanced, we do not perform bootstrapping for training the FCN parameters.

We have trained the network parameters using the Stochastic Gradient Descent (SGD) optimization with a batch size of 1 sample and softmax cross-entropy loss function. The learning rate is adaptively computed using the ADAM optimization method [46] with an initial learning rate equal to  $1 \times 10^{-4}$ . Furthermore, dropout layers with a probability of 0.85 are added to the layer numbers 17 and 18 of the convolution network (these are the deep layers of the 19-layer network).

#### 4.2.4 Needle axis estimation and visualization - post-processing

In order to robustly detect the instrument axis in the presence of outliers, we fit a model of the needle to the detected voxels using the RANSAC algorithm [142]. The needle model can be represented by a straight cylinder with a fixed diameter. In cases of large instrument deflection, the model can be adapted to define a parabolic segment, as shown in [143]. Using the RANSAC algorithm, the cylindrical model that contains the highest number of voxels is chosen to be the estimated needle. As the experimented needle diameters are less than 2 mm, we set the cylindrical model diameter to be approximately 2 mm.

After successful detection of needle axis, the 2D cross-section of the volume that contains the plane with the entire needle with maximum visibility, is visual-

ized, which is also perpendicular to coronal ( $X - Y$ ) planes. This cross-section is actually the in-plane view of the needle that is the most intuitive for physicians to interpret. This selection ensures that while the physician is advancing the needle, the entire instrument is visualized as much as it is visible and any misalignment of the needle and target is corrected without requiring to maneuver the transducer.

#### 4.2.5 Implementation details

The developed Python implementations of the proposed patch classification and semantic segmentation methods take on average 74 and 0.5  $\mu$ s for each voxel (1180 and 15 ms for each 2D cross-section), respectively, on a standard PC with a GeForce GTX Titan-X GPU. Therefore, when implementing a full 3D scan to process all voxels and cross-sections in the volume, patch classification executes in 4–5 mins., whereas semantic segmentation takes only 2–3 secs. Nevertheless, further optimization is possible using conventional techniques, such as a coarse-to-fine search strategy [144, 145] with a hierarchical grid, to achieve real-time performance. Furthermore, the execution time of RANSAC model-fitting is negligible, as the expected number of outliers is very small. The required computational power for realization of the proposed method is expected to be broadly available on high-end Ultrasound devices that can benefit from parallel computing processors, such as a GPU. However, for implementation in mid-range and portable systems, more efficient architectures should be investigated. Still, the continuously increasing computational capacity of mobile processors, as well as fast development and availability of on-board embedded units with pre-programmed convolutional modules, will make such computer-aided applications much more affordable and readily accessible to the majority of US devices.

### 4.3 Experimental results

The evaluation dataset consists of four types of ex-vivo US data acquired from chicken breast and porcine leg, using a VL13-5 transducer (motorized linear-array) and a X6-1 transducer (phased-array). The experiments with two types of transducers and tissue types investigate the robustness of the proposed methods in various acquisition settings and conditions. Properties and specifications of the applied datasets are summarized in Table 4.1. Each volume from the VL13-5 transducer contains on average  $174 \times 189 \times 188$  voxels (lateral  $\times$  axial  $\times$  elevation), at 0.2 mm/voxel and from the X6-1 transducer contains  $452 \times 280 \times 292$  voxels, at approximately 0.36 mm/voxel. Ground-truth data is created by manually annotating the voxels belonging to the needle in each volume. Testing evaluation is performed based on fivefold cross-validation separately for each transducer across its 20 ex-vivo 3D US volumes. For each fold, we use 4 subsets for training and 1 subset for testing, to make the training and testing data completely distinct.



#### 4. NEEDLE LOCALIZATION IN VOLUMETRIC 3D ULTRASOUND IMAGING

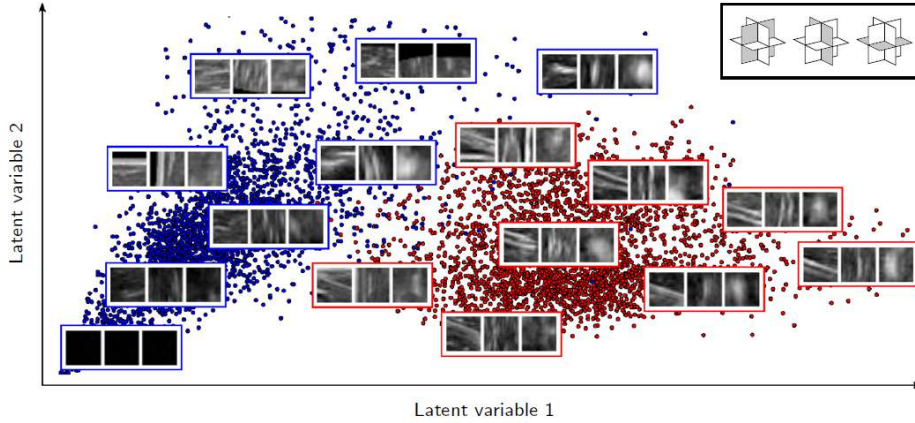
**Table 4.1** Dataset specifications and experimental settings of 3D US volumes, used for evaluation of the proposed methods.

Data and Instruments		Experimental settings			
Tissue type (transducer)	Needle type (diameter)	Volumes (counts)	Max. length (mm)	Steepness angles ( $^{\circ}$ )	Voxel size (mm)
Chicken breast (VL13-5 <sup>a</sup> )	17G (1.47 mm)	10	30	$10^{\circ} - 30^{\circ}$	0.20
	22G (0.72 mm)	10	30	$05^{\circ} - 50^{\circ}$	0.20
Porcine leg (X6-1 <sup>a</sup> )	17G (1.47 mm)	10	45	$55^{\circ} - 80^{\circ}$	0.36
	22G (0.72 mm)	10	35	$20^{\circ} - 65^{\circ}$	0.36

<sup>a</sup> Available from Philips Healthcare, Bothell, WA, USA

##### 4.3.1 Patch classification

We use Chicken breast dataset to evaluate the performance of the proposed patch classification method. The capability of the network to transform the input space to meaningful features is visualized using a multi-dimensional scaling that projects the representation of feature space onto a 2D image. For this purpose, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) [146] to the first fully-connected layer of the network. The result of the multi-dimensional projection of the test set in one of the folds is depicted in Figure 4.5, where neighboring points have similar characteristics in the feature space. As can be observed, the two clusters (needle and non-needle samples) are clearly separated, based on the features learned by the network.



**Figure 4.5** Multi-dimensional projection of voxels in the test set onto a 2D plane, using the t-SNE algorithm. Red and blue points represent needle and non-needle voxels, respectively.

Performance of the proposed methods is evaluated in the full volumes and the results are shown in Table 4.2, listing voxel-level recall, precision, specificity and F1-score. Recall is the sensitivity of detection and is defined as the num-

ber of correctly detected needle voxels divided by the actual number of voxels belonging to the needle. Precision or the positive predictive value is defined as the number of correctly detected needle voxels divided by the total number of detected needle voxels. Specificity is defined as the number of voxels that are correctly detected as non-needle divided by the actual number of voxels that are not part of the needle. Finally, the  $F_1$  score is calculated as the harmonic mean between the voxel-based recall and precision and is used to measure the similarity between the system detections and the ground-truth labels.

**Table 4.2** Average voxel classification performances of three methods in the full volumes of Chicken breast, measured by the indicated metrics, expressed in percentages (%). The  $\pm$  values indicates standard deviation.

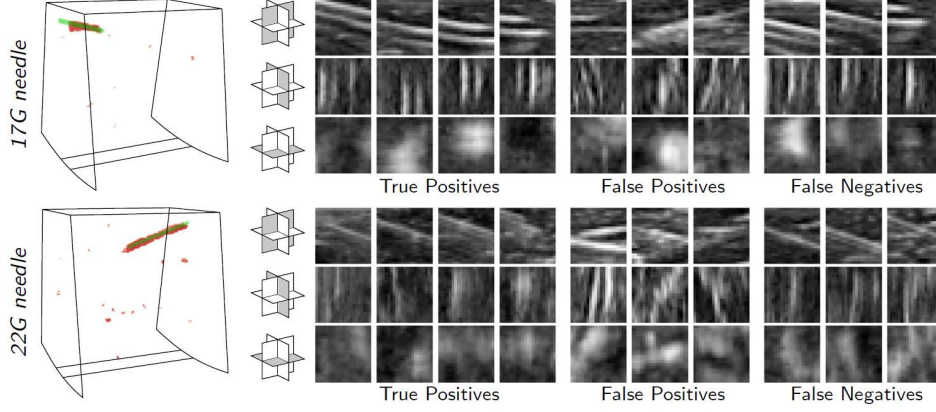
Method	Recall	Precision	Specificity	F1-score
Gabor transformation <sup>a</sup> [147]	47.1	48.2	-	53.7
SharedCNN <sup>b</sup>	76.3 $\pm$ 5.8	83.2 $\pm$ 5.6	99.98 $\pm$ 4 $\times$ 10 <sup>-5</sup>	78.5 $\pm$ 5.3
IndepCNN <sup>b</sup>	78.4 $\pm$ 5.3	64.7 $\pm$ 4.8	99.97 $\pm$ 6 $\times$ 10 <sup>-5</sup>	66.1 $\pm$ 4.9

<sup>a</sup> Two models trained separately for each needle (averaged)

<sup>b</sup> Single models trained directly for both needles

Furthermore, we compare the results with the approach of [147], which is based on supervised classification of voxels from their responses to handcrafted Gabor wavelets. As shown, both SharedCNN and IndepCNN architectures outperform the Gabor features yielding a 25% improvement on F1-score. Furthermore, SharedCNN achieves higher precision than IndepCNN at approximately similar recall rate. The degraded performance of IndepCNN can be explained by the large increase in the number of network parameters using multiple branches against our small-sized data.

Figure 4.6 shows examples of the classification results for 17G and 22G needles. As shown in the left column, the detected needle voxels correctly overlap the ground-truth voxels, which results in a good detection accuracy. Furthermore, example patches from true and false positives are visualized, which show a very high local similarity. Most of the false-negative patches belong to the regions with other nearby echogenic structures, which distort the appearance of the needle.



**Figure 4.6** Examples of classification results for 17G and 22G needles. (Left) Detected needle voxels in 3D volumes shown in red and ground-truth voxels in green. (Right) Example of triplanar orthogonal patches which were classified as true positives, false positives, and false negatives.

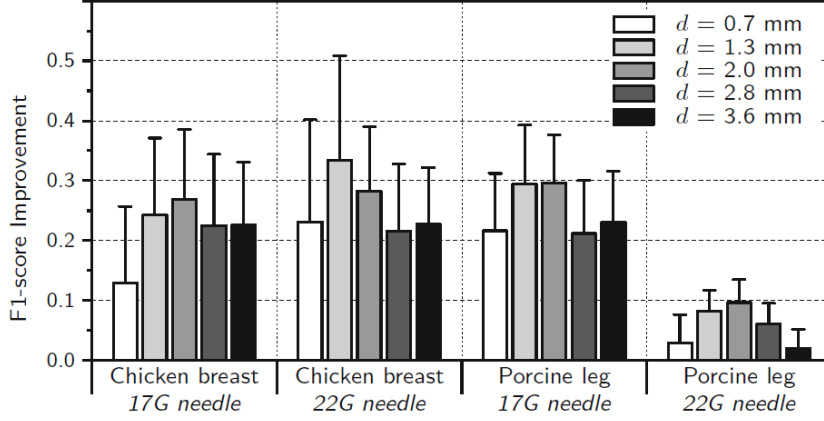
#### 4.3.2 Semantic segmentation

The performance of the proposed semantic segmentation method is evaluated on both datasets from Chicken breast and Porcine leg. As shown in Table 4.3, the data of porcine leg are acquired using a phased-array transducer, in which the needle appearance will be more inconsistent, due to the variable reflection angles of the backscattered beams. This phenomenon results in slightly lower recall values and more degradation of the precision score.

**Data representation in 2.5D:** As discussed in Subsection 4.2.3, we use a 3-channel input to the FCN network for better modeling of the 3D structures from 2.5D (thick slice) US data. The three channels consist of parallel cross-sections having a gap of  $d$  mm between them. In this subsection, we investigate the contribution of the proposed multi-slicing approach for increasing the segmentation accuracy of individual cross-sections and identifying the optimal  $d$  value for each type of data and needle.

Figure 4.7 depicts the bar chart of the measured improvement of the  $F_1$  scores for each dataset and choice of parameter  $d$  compared to a one-channel single-slice input. The  $F_1$  scores are calculated after cross-validation of the predictions on parallel cross-sections to the lateral and elevational axes. As shown, adding extra consecutive cross-sections for segmentation of the needle increases the performance in all cases. However, when the distance  $d$  is too large, the visible structures in the extracted cross-sections cannot be co-related to each other any longer and therefore the performance gain will decrease. As shown in Figure 4.7, the spacing values of 1.3 and 2.0 mm yield the highest improvement in the  $F_1$  score, while the results for 2.0 mm are more stable. Therefore, we adopt  $d = 2.0$  mm

as the optimal spacing among the consecutive cross-sections and use it in the following experiments.



**Figure 4.7** Improvements of  $F_1$  scores for different values of  $d$ , which is the gap between the consecutive slices used as input to the three-channel FCN network.

**Voxel segmentation performance:** The proposed method based on dense needle segmentation in multi-slice (2.5D) thick US planes is evaluated in terms of recall, precision and specificity, as mentioned earlier. Table 4.3 shows the obtained voxel-wise performances on both Chicken breast and Porcine leg datasets. As can be observed, the proposed SharedFCN architecture achieves very high recall and precision scores in both Chicken breast and Porcine leg datasets.

**Table 4.3** Average voxel classification performances of semantic segmentation approach in the full volumes (%).

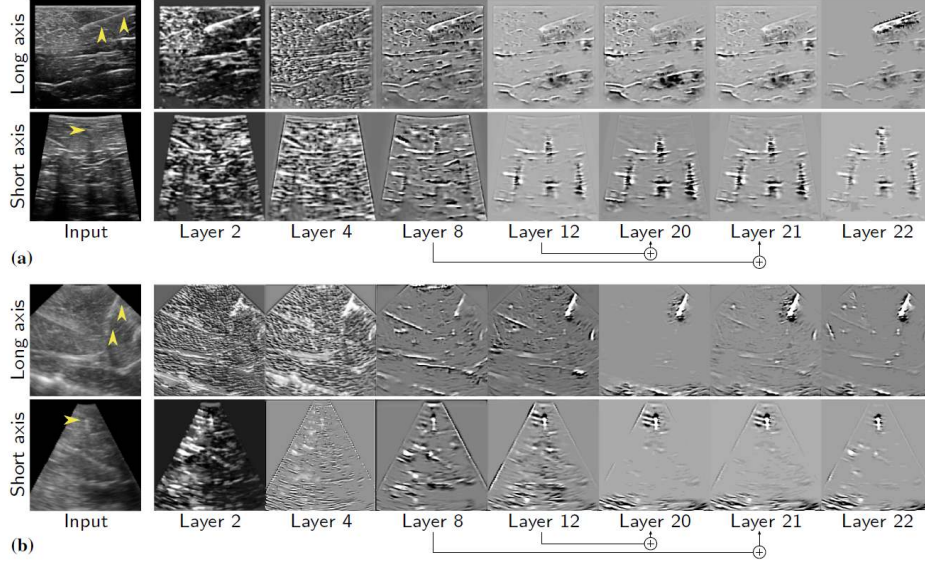
Method	Tissue	Recall	Precision	Specificity	F1-score
SharedFCN <sup>a</sup>	Chicken breast	$89.6 \pm 4.2$	$79.8 \pm 5.5$	$99.97 \pm 1 \times 10^{-4}$	$80.0 \pm 4.7$
	Porcelin leg	$87.9 \pm 4.2$	$83.0 \pm 3.7$	$99.99 \pm 1 \times 10^{-5}$	$84.1 \pm 3.4$

<sup>a</sup> Single model trained directly for both 17 and 22G needles

To study the performance of our trained networks in segmenting needle voxels, we visualize the response of the intermediate feature layers to needle cross-sections. For this purpose, the reconstructed patterns from the evaluation set that cause high activations in the feature maps are visualized using the Deconvnet, as proposed by Zeiler et al. [148]. Figure 4.8 shows the input stimuli that creates largest excitations of individual feature maps at several layers in the model, as well as their corresponding input images. As shown, both networks trained for VL13-5 and X6-1 transducers improve the discriminating features of the needle and remove the background as the network depth increases. However, it is interesting to notice the different modeling behavior of the network in convolution layers 12, 20 and 21 for the two transducers. In the dataset acquired using the

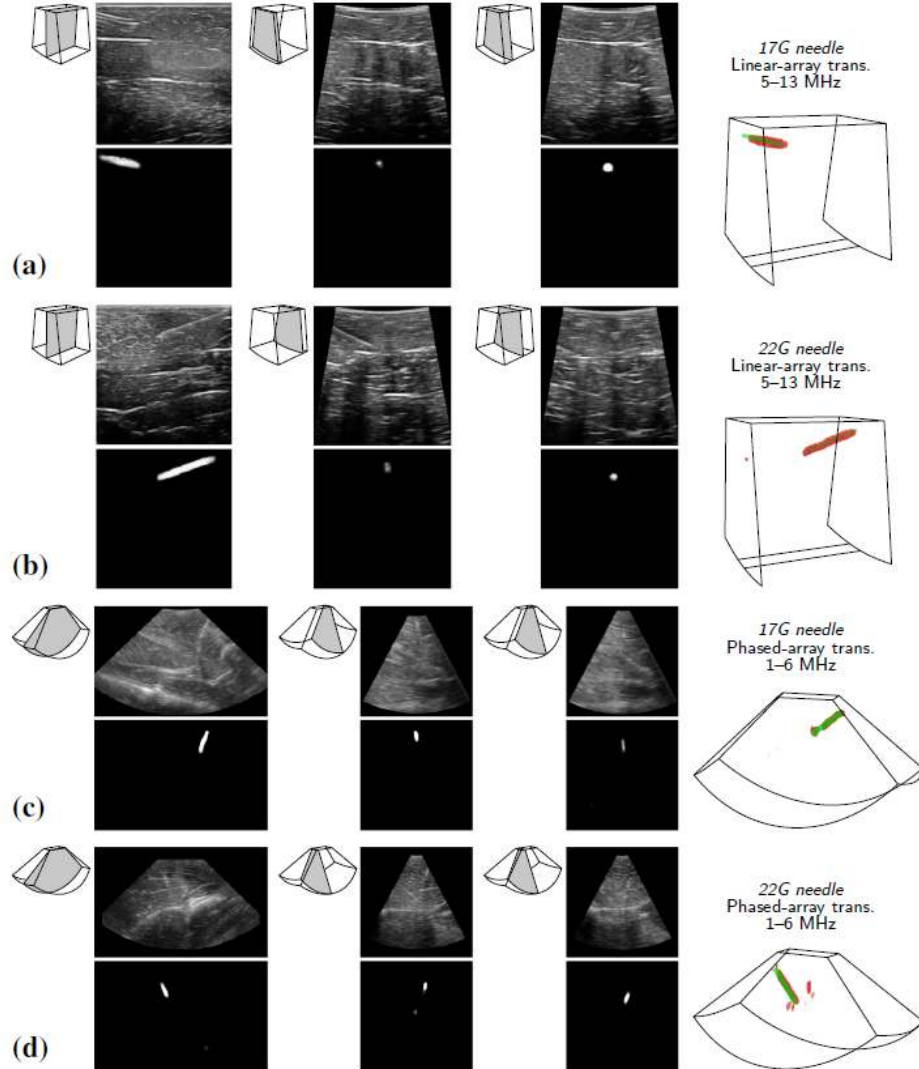
#### 4. NEEDLE LOCALIZATION IN VOLUMETRIC 3D ULTRASOUND IMAGING

VL13-5 transducer, the high-frequency range creates more strong shadow castings below the needle in the data. Therefore, as can be observed in Figure 4.8(a), the trained network additionally models the dark regions in layers 12 and 20 and fuses them to the shape and intensity features extracted in the shallower layers of the network.



**Figure 4.8** Visualization of features projected into the input space in the trained model; (a) for Linear-array VL13-5 transducer and (b) for Phased array X6-1 transducer. Reconstruction of the input image is shown by using only the highest activated features after the convolutional layers 2, 4, 8, 12, 20, 21, and 22. Note that the skip connections in the layers are intended to fuse coarse, semantic and local features. The ground-truth needle is marked with a yellow arrowhead in the input images.

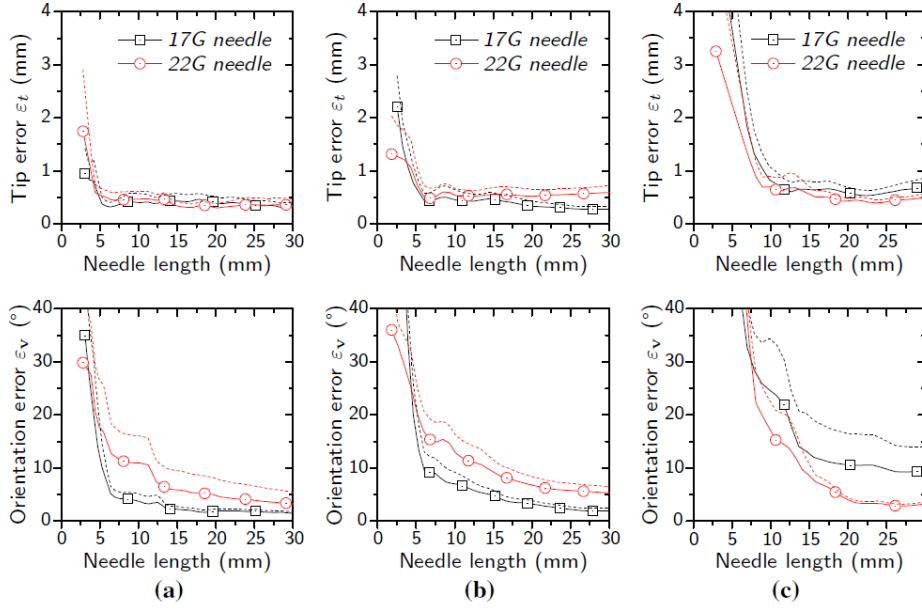
Figure 4.9 shows examples of the segmentation results in cross-sections, perpendicular to the lateral and elevational axes. As shown, the segmentation is very accurate for all the cases of a needle being entirely visible in a cross-section, partially acquired or being viewed from the out-of-plane cross-sections. In particular, Figure 4.9(d) depicts a case of a needle with a relatively large horizontal angle with the transducers, which results in the needle being partially acquired in all the processed cross-sections. As can be observed, visible parts of the needle at each cross-section are successfully segmented and after combining the results, the needle voxels are recovered and detected in 3D.



**Figure 4.9** Examples of segmentation results of (a) 17G and (b) 22G needles in the Chicken breast dataset, acquired with a linear-array transducer, and (c) 17G and (d) 22G needles in the the Porcine leg data, acquired with a phased-array transducer. Images in the top row of the sub-figures are input cross-sections to the network and images in the bottom row are the segmentation results. Volumes at the right side show the segmented needle voxels after combining the results from both lateral and elevational directions (in red) and the ground-truth voxels (in green).

### 4.3.3 Axis estimation accuracy

Since both SharedCNN and SharedFCN models detect the needle with a high precision, estimation of the needle axis is possible, even for short needle insertions after a simple RANSAC fitting. The accuracy of the proposed SharedCNN and SharedFCN models in localizing the needle axis is evaluated as a function of the needle length, as portrayed by Figure 4.10. Two measurements for evaluating the spatial accuracy of the detection are used. The needle tip error ( $\epsilon_t$ ) is calculated as the point-plane distance between the ground-truth needle tip and the detected needle plane. The orientation error ( $\epsilon_v$ ) is the angle between the detected and the ground-truth needle. As discussed earlier, the proposed models do not explicitly detect the needle tip, but detect the plane where the needle and its tip are maximally visible. This localization processing is done for every individual 3D volume data, leading to repeated detection in cases of a temporal US sequence (i.e. 4D data).



**Figure 4.10** Needle tip position error ( $\epsilon_t$ ) and orientation error ( $\epsilon_v$ ) as a function of needle length. Dashed lines represent standard errors of the measured values. (a) SharedCNN results for Chicken breast dataset with voxel size of  $\approx 0.20$  mm. (b) SharedFCN results for Chicken breast dataset with voxel size of  $\approx 0.20$  mm. (c) SharedFCN results for Porcine leg dataset with voxel size of  $\approx 0.36$  mm

As shown in Figure 4.10 (a) and (b), both SharedCNN and SharedFCN models perform accurately in the Chicken breast datasets, reaching a tip error of less than  $\epsilon_t < 0.7$  mm for needle lengths of approximately 5 mm or larger. In both approaches, the  $\epsilon_v$  shows more sensitivity to shorter needles and varies more for the 22G needle, which is more difficult to estimate, compared to a thicker



needle. Furthermore, for the SharedCNN model, voxels in the first 2 mm are undetectable, as the minimum distance of extracted 3D patches from the volume borders corresponds to half of a patch length.

In the Porcine leg datasets, the voxel size is reduced to 0.36 mm, due to the lower acquisition frequency of the phased-array transducer. Therefore, longer lengths of the needle are required for accurate detection. As shown in Figure 4.10 (c), for needles of approximately 10 mm or longer, the  $\epsilon_t$  reduces to 0.7 and 0.6 mm for 17G and 22G needles, respectively. In contrast to the Chicken breast dataset, the  $\epsilon_v$  is generally larger for short 17G needles in the Porcine leg data. Most importantly, in all of the experiments, the needle tip error,  $\epsilon_t$  remains lower than 0.7 mm. This shows that after insertion of only 5 mm for higher-resolution linear-array transducers and 10 mm for lower-resolution phased-array transducers, the tip will be always visible in the detected plane because their distance is less than the thickness of US planes.

#### 4.4 Discussion

*Comparison of two models:* Comparing the results in Tables 4.2 and 4.3, it can be observed that the performance of SharedFCN is comparable and only slightly better than the patch-based SharedCNN on Chicken breast data acquired from the higher-frequency range VL13-5 linear-array transducer. However, a major benefit of dense segmentation using SharedFCN is related to the data of the lower-frequency range X6-1 phased-array transducer. The resulting lower spatial resolution of these transducers distorts the appearance and obscures the structure details of a needle. In these cases, training a discriminant model of the needle requires deeper and more complex networks, which increases the computational complexity. Therefore, a more computationally efficient network such as the proposed SharedFCN is preferred over the patch classification networks.

*Influence of patch size:* Furthermore, as discussed in Section 4.1, the US beam steering angle of a phased-array transducer varies for each region in the field of view. Consequently, US reflections from different parts of a needle will vary largely, such that a considerable portion of the needle shaft can be virtually invisible in the data. Therefore, the receptive field of a convolutional network needs to be large enough to model the contextual information from the visible parts of the needle. In a patch-based classification technique, a larger receptive field can be achieved by, e.g., increasing the patch size, increasing the number of convolutions and max-pooling layers, or employing normal or atrous convolutions with larger kernel sizes. In all of these methods, the computational complexity increases exponentially as more redundant calculations have to be computed for adjacent patches, and the spatial accuracy decreases as small shifts of patches cannot be translated to two different classes. This is mainly because the context of two large patches with a small shift only include minor changes.



## 4.5 Conclusion

Ultrasound-guided interventions are increasingly used to minimize risks to the patient and improve health outcomes. However, the procedure of needle and transducer positioning is extremely challenging and possible external guidance tools would add to the complexity and costs of the procedure. Instead, an automated localization of the needle in 3D US can overcome 2D limitations and facilitate the ease of use of such transducers, while ensuring accurate needle guidance.

Conventional methods including Gabor wavelet analysis for association of the needle 3D shape in the US volume have established the first results of needle detection and localization in 3D US volumes. This chapter has introduced two different deep learning techniques for detecting needles in 3D US data, which achieve very high precision at a low false-negative rate. Comparison with the conventional technique, shows that deep learning-based solutions significantly outperform the conventional approach with a large margin (see Table 4.2). The largest part of improvement is explained by using learning filters, instead of fixed Gabor filters. Furthermore, the high precision is achieved by exploiting dedicated convolutional networks for needle segmentation in 3D US volumes. The proposed models are based on ConvNets, which are improved by proposing a new update strategy to handle highly imbalanced datasets by informed resampling of non-needle voxels. Furthermore, novel modeling of 3D US context information is introduced using 2.5D data of multi-view thick-sliced FCN.

The proposed patch classification and semantic segmentation systems are evaluated on several ex-vivo datasets and outperform the classification of the state-of-the-art handcrafted features, achieving 78 and 80%  $F_1$ -scores in the chicken breast data, respectively. This shows the capability of CNN in modeling more semantically meaningful information in addition to simple shape features, which substantially improves needle detection in complex and noisy 3D US data. Furthermore, our proposed needle segmentation method based on 2.5D US information achieves a 84%  $F_1$ -score in the porcine leg datasets that are acquired with a lower-resolution phased-array transducer. These results show strong semantic modeling of the needle context in challenging situations, where the intensity of the needle is inconsistent and even partly invisible.

The performance analysis of the proposed system is also discussed here from the clinical point of view. The accuracy of localization and needle tip finding are essential for clinical acceptance. Quantitative analysis of the localization error with respect to the needle length shows that the tip error is less than 0.7 mm for needle insertions of only 5 mm long and 10 mm long at the voxel size of 0.2 and 0.36 mm, respectively. Hence, the system is able to accurately detect short needles as well, enabling the physician to correct inaccurate insertions at early stages in both higher resolution and lower-resolution US volumes. Furthermore, the needle is visualized intuitively by its in-plane view while ensuring that the tip is always visible, which eliminates the need for advanced manual coordination of the transducer.

The future work may evaluate the proposed method in even more challenging in-vivo datasets with suboptimal acquisition settings. Due to the complexity of data from interventional settings, larger datasets need to be acquired for training more sophisticated networks. Moreover, further analysis is required to limit the complexity of ConvNets with respect to their performance for embedding this technology as a real-time application.

The next chapter builds further on exploring 3D medical data, but for non-Euclidean 3D point clouds that have been less explored by the deep learning community for medical applications. In particular, we study semantic teeth segmentation in 3D point cloud data of intra-oral scanners and the existing challenges for using such data and the application of deep learning for the proposed solutions.



## Teeth Instance Segmentation in 3D Point Cloud Data

### 5.1 Introduction

The previous chapter has presented the analysis of noisy 3D US volumetric data for needle localization during medical interventions. The data-driven solutions clearly outperformed the conventional approaches. This chapter aims at using a different 3D data modality for automating a tedious and laborious task in dentistry, which is teeth instance segmentation in 3D point cloud data.

Dentistry has witnessed a rapid growth of technological innovations in advanced imaging methods. Such advances in imaging systems are playing an important role in efficient diagnoses, treatment, and surgeries. Computational dentistry involves computerized methods for automated analysis of digital dental images. It utilizes both mathematical and data-driven models to facilitate data analysis for e.g. accurate treatment planning and diagnostic purposes.

Computational dentistry may incorporate multiple sources of imaging data obtained by both extra-oral (e.g. X-ray panoramic, cephalometric and cone-beam computed tomography) and intra-oral optical imaging (e.g. laser or structured light projection scanners). The emergence of digital imaging equipment has been a driving force for developing computer-aided design (CAD) systems. The CAD systems can analyze the imaging data for highly accurate treatment planning and provide useful clinical information to support better treatment. For supporting an automated clinical workflow in implantology and orthodontic fields, such a CAD system should be able to resolve some fundamental problems of which accurate semantic segmentation of teeth and gingiva (gums) from imaging data is highly desirable.

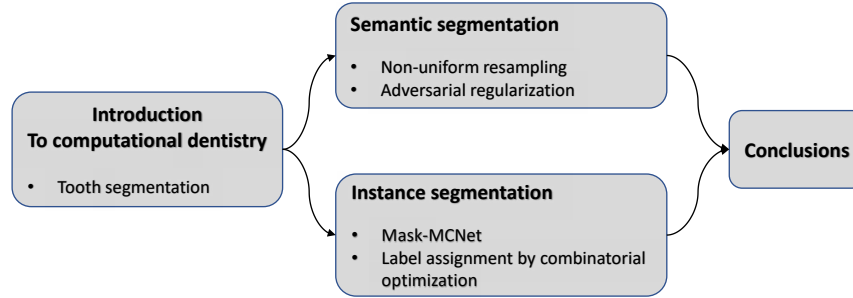
Intra-oral scanners are advanced imaging devices for optical measurement of the surface profiles of anatomical structures inside the oral cavity of the patient.

Similar to other 3D scanners, intra-oral scanners project a light source (laser, or structured light) on the surface of objects to be scanned, in this case, the dental arches. Based on the underlying technique, the time-of-flight of the laser or the deformation of the projected pattern on the subject's surface is measured by the imaging sensors and processed by the scanning software. This software generates a highly accurate 3D point cloud, in the order of 30-80 points/mm<sup>2</sup> [149] with a spatial accuracy of less than 20  $\mu$ m. The obtained point cloud can be further processed and converted to a 3D surface model (mesh) by using *triangulation* techniques. Such a precise 3D model is widely used for implant treatment and orthodontic planning. A scan of one dental arch consists of a large set of points (e.g. hundreds of thousands) in the 3D Cartesian coordinate system.

As mentioned earlier, one challenging problem in designing a CAD system for computational dentistry is accurate semantic segmentation of intra-oral scans (IOS). This problem refers to assigning a label, based on the *Federation Dentaire Internationale* (FDI) standard. In more technical details, this involves labeling all points as belonging to a specific tooth crown, or as belonging to gingiva within the recorded IOS point cloud. Each point is represented by a coordinate in the 3D Cartesian coordinate system, which is not universally acquired (i.e. the latter can be different between two IOS). The FDI specifies 32 labels for adult dentition, referring to 16 teeth in each upper and lower jaw. In this study, we treat the teeth on the upper and lower jaw in the same way, so that we only consider 16 separate labels to be classified. This changes the problem to finding 16 classes (+1 extra for the gingiva), which facilitates better learning. Automated processing of such a large 3D point cloud by a computational model that preserves highly detailed information from anatomic structures of teeth crowns is highly beneficial for many clinical dental applications.

The more specific challenges of the problem statement for this chapter are as follows.

- *Inherent 3D point cloud issues*: Handling of 3D point cloud data is a specialism on its own. In contrast with 2D images or 3D volume data, the point cloud data modality is not structural data. In other words, there is no spatial ordering or arrangement among elements of the data. Furthermore, in practice the IOS scans may have some missing data due to non-reflected areas of the mouth cavity. Also, there may be some outliers such as implants on some teeth crowns of the patient. Proper handling of the missing data and outliers by the learning model is a significant challenge.
- *Efficient deep learning model*: 3D IOS data include a huge number of samples which contains fine geometrical information of teeth-crown shapes. Reducing the spatial resolution of data degrades the spatial accuracy and leads to inaccurate 3D modeling. Hence, processing the data samples in its original resolution is essential but very computationally intensive.
- *Highly accurate segmentation*: Segmentation has become a regular task for deep learning models, even including 3D point cloud data modality. However in this case, the dentist expects to generate a tooth crown that smoothly



**Figure 5.1** schematic of the chapter layout

fits to the cavity and aligns properly with surrounding teeth in terms of shape. Such crowns are segmented from the IOS data. This smooth and glueless fit requires a high accuracy in the predictive segmentation.

As shown in Figure 5.1, in this chapter, we introduce two deep learning methods for tooth segmentation in the 3D point cloud of IOS. The purpose of this study is to propose a new methodology, based on recent advances in deep learning for the automation of the clinical workflow and to improve the quality thereof. This new method starts already at beginning stages of the model where data analysis takes place at the full resolution of the point cloud, in order to prevent loss of accurate information. To sustain high-accuracy processing, a new deep learning architecture called Mask-MCNet is proposed. The performance of the proposed models are compared against existing competitive models for 3D point cloud segmentation. The empirical results show that state-of-the-art performance is achieved by the proposed models.

The sequel of this chapter is structured as follows. Section 5.2 briefly introduces the related work on IOS segmentation, recent advances of deep learning in semantic instance segmentation on point cloud data, and the proposed contributions to both. Afterwards, Section 5.3 and Section 5.4 explain the two proposed deep learning models for accurate teeth instance segmentation and the obtained results in detail. Lastly, Section 5.4.5 provides discussions and conclusions.

## 5.2 Related work

Related literature to our work can be divided into two parts: conventional IOS segmentation methods (Section 5.2.1) and existing deep learning models for object-instance segmentation in 3D point clouds (Section 5.2.2).

### 5.2.1 Conventional IOS segmentation approaches

The literature on IOS segmentation covers mostly conventional computer vision/graphics techniques, which are limited to finding the best handcrafted features, manual tuning of several parameters and lack of generalization and robustness [150]. Among the proposed methods, one generic approach is first projecting the 3D IOS mesh on one or multiple 2D plane(s) and then applying standard computer vision algorithms. Afterwards, the processed data is projected back into the 3D space. For example, Kondo *et al.* [151] propose gradient orientation analysis and Wongwaen *et al.* [152] apply a boundary analysis on 2D projected panoramic depth images for finding teeth boundaries.

Most of the other studies are based on curvature analysis [153, 154, 155, 156, 157], *fast marching watersheds* [158], *morphological operations* [157], 2D [159] and 3D [160] active contour (snake) analysis and tooth-target harmonic fields [161] for segmenting the teeth and gingiva. Some other works follow a semi-automated approach by manually setting a threshold [154], picking some representative points [156], or interactively involving a human operator for the analysis [155, 157]. As already mentioned, the performance of these methods is mostly limited to the expressiveness of the used handcrafted features.

### 5.2.2 Deep learning approaches

The available deep learning approaches for structured learning on geometric point clouds can be roughly categorized into four types: *feature-based deep neural networks (DNNs)*, *volumetric*, *2D projection*, and *point cloud* methods.

*Feature-based DNNs* first extract a set of standard shape features (e.g. based on computer graphic algorithms) and then apply a neural network (e.g. a ConvNet) for feature classification [162, 163]. The performance of this approach is limited to the discriminating properties of the handcrafted features [164]. The *volumetric* approach first voxelizes the shape and then applies 3D ConvNet models on the quantized shape into a 3D grid space [165, 166]. As expected, spatial quantization limits the method's performance, especially when fine, high-frequency details need to be preserved in shape curvatures for accurate prediction. The *2D projection* approach first renders the 3D data into one/multiple 2D plane(s) and then applies the 2D convolution operator for the 2D-image pixel classification. Afterwards, the processed data are projected back into the 3D data [167]. *Point cloud* deep learning models work directly with raw point clouds [164, 168, 169, 170]. Each point has some attributes, mainly their 3D coordinates and sometimes additional attributes, like the normal of a surface they may represent, color, etc. Currently, deep learning-based point cloud models form a very active research track. This last approach does not suffer from some shortcomings that occur when using handcrafted features, quantization errors or high processing demands, as is the case with earlier mentioned approaches.

The next two sections introduce the two different proposed deep learning frameworks for the 3D point cloud analysis of IOS. The first approach that defines tooth segmentation as a semantic segmentation problem for 3D point cloud

of IOS is the first published deep learning-based solution for this problem. The second approach formulates the tooth segmentation as an instance segmentation problem that has some advantages over the first approach. The following describes each of these two methods in detail and shows how they perform on the real-world clinical dataset.

### 5.3 Model 1: Semantic segmentation in 3D point cloud scans

#### 5.3.1 Technical aspects and contributions of method 1

In this study, we improve IOS semantic segmentation by means of end-to-end learning of an accurate segmentation model. Building an accurate segmentation model involves two aspects of complexity. Firstly, complexity originates from dentition (teeth arrangement) and data acquisition. Since the shape of two adjacent tooth crowns (e.g. two molar teeth) may appear to be similar, assigning a correct label demands additional information such as relative position with respect to other teeth on the dental arch. Furthermore, the presence of *abnormalities in dentition and shape deformation*, makes IOS segmentation a challenging task for a segmentation model. Examples of such an abnormality may be lacking teeth (e.g. wisdom teeth). Additional challenges may arise from acquisition issues such as partially missing data (e.g. because of occlusion in scanning), lack of a universal coordinate system, presence of noise, outliers, etc. The interaction between these challenges is important for successfully applying computer vision algorithms.

The second aspect of complexity relates to the *3D geometrical representation of data* by a point cloud that is not well suited to the decent deep learning models that are highly performant on 2D/3D images. Application of such deep learning models (e.g. ConvNet architectures) to point cloud analysis would require three main issues to be addressed. These are (1) data irregularity, (2) permutation-invariance, and (3) re-sampling-invariance. These issues are briefly discussed below.

**A. Irregularity** of the point cloud means that the data elements are not organized on a 2D/3D grid, like the data in 2D/3D images. This mainly originates from the pseudorandom nature of recording (sampling) of the external surface of an object, recorded by e.g. a laser scanner. The irregularity results in ineffective use of convolutional filters for capturing the spatial-local correlation in data [169], as they work best on organized data.

**B. Permutation-invariance** refers to the geometrically unordered presentation of a point cloud. If we present a point cloud by a matrix in which each row contains a point, alternating the order of the rows does not change the data semantics, while it does affect the numerical computation in deep learning architectures.

**C. Re-sampling-invariance** is a property that means random selection of a sufficiently large subset of the points, preserving the global structure of the object captured by the overall point cloud. The IOS data contains tens of thousands of points. The number of points can vary considerably between two scans, or even



between different acquisition cycles of the same object. Processing such large-scale and variable-size data is challenging for a deep learning model. Hardware limitations (e.g. memory of the GPU) and working with fixed-rank matrices require a re-sampling stage. However, a naive re-sampling approach can invoke the loss of important information and is highly application-dependent.

Since 2016, several studies have investigated point cloud analysis by artificial neural networks for object classification/segmentation tasks. *PointNet* [164] and *DeepSets* [168] are two pioneering works from recent years, based on the multi-layer perceptron (MLP) network, recently followed by other researchers [170, 169]. Available deep learning models include some inventive techniques for the joint handling of the first two mentioned issues (i.e. irregularity and permutation invariance), while still addressing the third issue by applying a uniform re-sampling for fixing the number of points. Although such an approach is sufficient for many applications like object classification (e.g. classifying the chairs vs. tables), it does not preserve the finer details of data, which is important for the segmentation tasks (e.g. classifying a point close to the borderline of a tooth and gingiva). This last issue if not addressed causes significant performance loss in semantic segmentation tasks.

**Contributions of Method 1:** This section presents an end-to-end learning framework for IOS segmentation, based on recent point cloud deep learning models. Our contributions are threefold.

1. To the best of our knowledge, this is the first end-to-end learning study, proposed for IOS point cloud segmentation.
2. We propose a unique non-uniform re-sampling mechanism, combined with a compatible loss function, for training and deploying a deep network. The non-uniform re-sampling facilitates the training and deployment of the network on a fixed-size re-sampled point cloud which contains different levels of spatial resolution, involving both local fine details and the global shape structure.
3. In addition to a point-wise classification loss, we employ an adversarial loss for imposing a realistic arrangement of teeth labels to the network prediction. So, the segmentation network learns the realistic layout of the labeling space and improves the classification of points, by involving the high-level semantics and preserving the valid arrangement of the teeth labels on the dental arch. In contrast to the existing methods, the discriminator network is applied only to the statistics, which are computed from the spatial distributions of labels and the predictions. Therefore, taking advantage of the abstraction with statistics only, enables to employ a shallow network as a discriminator that facilitates the training.

### 5.3.2 Non-uniform re-sampling

Because of the mentioned re-sampling-invariance property of the point cloud, training a deep learning model on a whole set of points of an IOS poses several issues. Two examples of such issues are handling of the variable-rank matrices (the number of points in our IOS dataset varies between [100k, 310k]) and the hardware limitations (e.g. computation memory) for processing of the large point cloud. Applying a patch-classification technique, which is common for large-size 2D/3D images, degrades the quality of segmentation results because the extracted patches (i.e. local subsets of points) lack contextual information about the whole surface structure that is required for semantic labeling of a tooth. Therefore, the patch-sampling technique is deprived of the existing strong dependency between the label of each point and its location in the point cloud. As we already mentioned, the alternative solution based on common (uniform) down-sampling techniques does not lead to an accurate analysis of data at its highest available resolution.

Recently, various non-uniform re-sampling methods have been proposed by means of optimization of different metrics that preserve either high-frequency contents [171, 172], or local directional density [173]. However, the effectiveness of using such data abstraction methods on the performance of a deep network cannot be easily established and is in contrast with our interest in designing an end-to-end learning scheme that operates directly on the raw data. In essence, it is preferable to have such an abstraction of information to be performed by the network itself with respect to its objective function. The proposed non-uniform re-sampling method is based on the Monte Carlo sampling technique and results in a locally-dense and globally-sparse sampling of points that is beneficial for training the deep learning model on IOS data.

We now state the re-sampling problem more formally, such that data from one set is re-sampled into another set where the point cloud is locally dense and globally sparse. The advantage and further background of this method will be explained later after the formal description.

We assume a matrix representation for the point cloud ( $X = [x_1, x_2, \dots, x_N]$ ) with  $N$  points of which each point has  $D$  attributes ( $x_i \in \mathbb{R}^D$ ), where  $D = 3$  for the 3D geometric points. By introducing a radial basis function (RBF), denoted by  $\mathcal{K}$ , which is positioned on a randomly chosen point ( $x_f \in X$ ), the geometrical similarity (spatial distance) to the point  $x_f$  can be measured with a weighted distance metric, as specified in Eq. (5.1). In accordance with the *foveation* as defined in the work of Cirean *et al.* [120], we call this point the *fovea*. The RBF kernel is specified by:

$$\mathcal{K}(x_i, x_f) = \exp\left(-\frac{\|x_i - x_f\|^2}{2\sigma^2}\right), \quad (5.1)$$

where  $\sigma$  is a free parameter that controls the bandwidth (*compactness*) of the kernel. By re-sampling, we aim to choose a subset  $Y$  out of  $X$  with  $M$  points ( $M < N$ ) that has a dense sampling around the fovea and a sparse sampling for farther locations. According to the Monte-Carlo sampling, by randomly drawing

(with replacement) a point  $x_i$  from the set  $X$ , we accept to insert such a point into the subset  $Y$ , only if the condition  $\mathcal{K}(x_i, x_f) > r_\delta$  is satisfied, otherwise, it is rejected. The variable  $r_\delta$  is a random number from a uniform distribution within the unity interval according to the Monte-Carlo technique. This process iterates until  $M - 1$  unique points are accepted. Algorithm 3 shows these steps in detail. As a result, the subset  $Y$  includes  $M$  total points at different levels of granularity (see Figure 5.2). By random selection of the fovea in every training batch ( $x_f \in_R X$ ), the model trains on the whole point cloud in its highest available resolution with a fixed number of points. It is worth mentioning that as the point cloud is normalized to have a variance of unity, the uniform-re-sampling and patch sampling both can be considered as two extreme cases of the proposed algorithm by setting  $\sigma \gg 1$  and  $\sigma \ll 1$ , respectively. The aforementioned formal description of the non-uniform re-sampling is summarized below in the form of Algorithm 3 as pseudocode.

---

**Algorithm 3** Non-uniform re-sampling

---

```

input : point cloud
output: non-uniform re-sampled point cloud

 $X \leftarrow \{x_1, x_2, \dots, x_N\};$  // Whole point cloud
 $Y \leftarrow \emptyset;$  // Initialized empty set
 $x_f \leftarrow x \sim X;$  // Randomly draw one sample as fovea
Function  $\mathcal{R}(x_f, X)$ :
    while  $|Y| < M;$  // Check the size of Y
    do
         $x_i \leftarrow x \sim X;$  // Randomly draw sample
         $r_\delta \sim \text{uniform}(0, 1);$  // Draw a random value
        if  $\mathcal{K}(x_i, x_f) \geq r_\delta;$  // The RBF kernel Eq. 5.1
        then
            if  $x_i \notin Y$  then
                 $Y \leftarrow x_i \cup Y;$  // Insert to the subset
            end
        end
    end
    return  $Y$ 

```

---

### 5.3.3 Model architecture

The proposed model includes two networks: the *segmentation* network ( $\mathcal{S}$ ) and the *discriminator* network ( $\mathcal{D}$ ). The PointCNN [169] architecture is used for implementing the  $\mathcal{S}$  network. The inputs to the  $\mathcal{S}$  network are the re-sampled points and its output is a 17-element vector for each point, which represents the class probability.

**A. Weighted point-wise cross-entropy loss:** As the input is non-uniformly sampled, training the network by computing an equally weighted loss for each point is not efficient. Since the re-sampled point set contains various levels of granularity, equally penalizing the output errors for dense and sparse regions

prevents the model from optimally adapting its convolutional kernels to capture the fine-detailed content in the data, as the error on sparse points increases relatively equally. Figure 5.2 shows the uncertainty values for each point, predicted by the network with an equally weighted loss function. As expected, the sparse regions with a lower sampling rate yield a high uncertainty during the learning process. This is mainly because the missing context makes it difficult for the network to perform as accurately as it performs in dense regions. For optimizing the performance of the learning algorithm, we have to trade-off the preservation of the sparse points (which contain the global dental arch structure) and learning of the fine curvature in point cloud data by parameter-tuning. To do so, we apply different weights per point, which are computed with the distance metric of the RBF kernel (see Eq. (5.1)).

By assuming the posterior probability vector ( $\mathbf{p}_i$ ) for point  $x_i$ , which is computed at the output softmax layer of the segmentation network with the transfer function of  $\mathcal{S}(x_i, \theta)$ , the weighted loss value ( $\mathcal{L}_p$ ) for each point  $i$  is formulated by:

$$\mathcal{L}_p = - \sum_{i=1}^M w_i \cdot \sum_{j=1}^L y_i \cdot \log(p_{ij}), \quad (5.2)$$

where  $\mathbf{p}_i = \mathcal{S}(x_i, \theta_S)$  is the posterior probability of labels, weights  $w_i = \mathcal{K}(x_i, x_f)$  is computed from the RBF kernel, the elements of  $\mathbf{p}_i$  are  $\mathbf{p}_i = [p_{i1}, \dots, p_{iL}]$  and associate point  $x_i$  with class label  $j$ . The probabilities  $p_{ij}$  sum up to unity over all classes  $j$ . The variable  $y_i$  denotes the one-hot encoded target label for the  $i$ th point with  $x_i$  in 3D coordinates. In our experiments,  $L = 17$  and  $M=3 \times 10^4$  denote the number of labels and the number of re-sampled points, respectively.

**B. Adversarial loss:** Training the segmentation network only by applying a pixel-wise cross-entropy loss (Eq. (5.2)) has an important shortcoming. The label of each point in the cloud has a high dependency on the label of its adjacent points. For example, if a point belongs to an incisor tooth, its adjacent points can only belong to the same or another incisor, a canine tooth or the gingiva, but certainly not belong to a molar tooth. Although such a strong structural constraint exists in the data, it is ignored when the optimization problem is only formulated by Eq. (5.2). As discussed in [174], such a formulation is ill-posed, since the semantic segmentation is inherently not a pixel-based (point-wise) classification problem. For improving the higher-level semantic consistencies, Luc *et al.* [175] employed adversarial training in addition to supervised training of the segmentation network. According to such an approach, a discriminator network provides a supervisory signal (feedback) to the segmentation network, based on differences between distributions of labels and the distribution of predictions. Such an effective mechanism was later followed for medical image analysis [176, 177, 178, 179, 180, 181].

Inspired by prior works, we use a discriminator network for imposing the spatial distribution of labels as an inductive bias for training the segmentation network. However, different from the common approach, instead of employ-

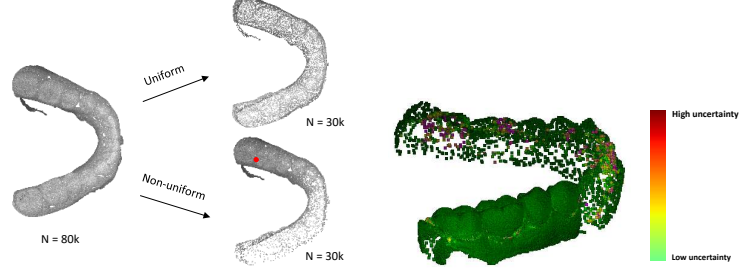
ing a costly training of a discriminator network on labeled samples in its high-dimensional input space, we train a shallow network on statistics that have been computed over predicted and ground-truth labels. We compute two statistical variables from both the predicted labels and the real labels. Afterwards, by training a shallow MLP network as a discriminator, the segmentation network learns not only to predict the label of each point through its cross-entropy loss but also produces a realistic arrangement of labels over the dental arch by minimizing the discrepancy with ground-truth labels. So, the input to the discriminator is composed of statistics over predicted labels and can be specified as:

$$\hat{\mathbf{u}} \in \mathbb{R}^{(L-1) \times 6} = \left[ \hat{\mu}_1, \hat{\sigma}_1^2 \parallel \hat{\mu}_2, \hat{\sigma}_2^2 \parallel \dots \parallel \hat{\mu}_{L-1}, \hat{\sigma}_{L-1}^2 \right], \text{ where} \quad (5.3)$$

$$\hat{\mu}_j = \sum_{i=1}^M p_{ij} \cdot x_i \quad \text{and} \quad \hat{\sigma}_j^2 = \sum_{i=1}^M p_{ij} \cdot (x_i - \hat{\mu}_j)^2.$$

for  $j = 1, 2, \dots, L - 1$ . Here, the symbol  $\parallel$  denotes a vertical vector concatenation (stacking) and  $L$  denotes the number of labels in the data. The statistics that we used simply consist of the mean ( $\hat{\mu}$ ) and variance ( $\hat{\sigma}^2$ ) of the coordinates of all points with the same label, as given by the segmentation network. The stacked feature set ( $\hat{\mathbf{u}}$ ) can be read as a *soft* computation of the central positions of teeth and their variance (i.e. their soft bounding boxes) in the 3D space, according to the predicted labels ( $p_{ij}$ ). The statistical mean and variance are computed only for  $L - 1$  classes of teeth while excluding the gingiva class, as its point cloud is spread almost across the whole input space and the applied non-uniform re-sampling stage alters its resulting statistics severely across training batches. By replacing the  $p_{ij}$  values in Eq. (5.3) with the one-hot encoded values of the ground-truth labels ( $y_i$ ), the associated feature set of  $\hat{\mathbf{u}}$ , denoted by  $\mathbf{u}$ , is obtained. For any absent label in the point cloud, we simply insert a vector consisting of zeros instead. The feature set  $\mathbf{u}$  represents a *realistic* statistical measurement of the labeled data.

The discriminator network ( $\mathcal{D}$ ) is trained to classify the feature set  $\mathbf{u}$  and  $\hat{\mathbf{u}}$ . The network consists of two cascaded parts. The first part estimates an affine transformation and is applied to an input 96-element ( $16 \times 6$ ) input vector. The second part consists of a 3-layer MLP network that maps the transformed input vector into a scalar value by a sigmoidal activation function at its output node. In effect, the network tries to produce the scalar of unity value at its output if the network is applied on ground-truth statistics  $\mathbf{u}$ , while the scalar 0 should be produced if the network is applied on the predicted  $\hat{\mathbf{u}}$ . The architecture of the first part of the network is identical to what is proposed in the PointNet model [164], called a *T-Net*. More details about the T-Net can be found in [164]. In adversarial training, the discriminator is trained by minimizing the cross-entropy loss for binary classification of  $\hat{\mathbf{u}}$  and  $\mathbf{u}$ . After updating the parameters of the discriminator network in each iteration of training, the expectation of network over predicted  $\hat{\mathbf{u}}$  is used as an additional loss for training the segmentation network. These two



**Figure 5.2** Example of uniform versus non-uniform re-sampling (left). The fovea is shown by a red dot. The uncertainty (represented by color) of the prediction for dense and sparse regions (right) indicates that the sparse regions contain spurious red points. The color bar shows the range of uncertainty, estimated by the network.

losses are specified as:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\mathbf{u}, \hat{\mathbf{u}}; \theta_{\mathcal{D}}, \theta_S) &= \mathbb{E}_{\mathbf{u}}[\log \mathcal{D}(\mathbf{u})] + \mathbb{E}_{\hat{\mathbf{u}}}[\log(1 - \mathcal{D}(\hat{\mathbf{u}}))], \\ \mathcal{L}_{\text{Adv}}(\hat{\mathbf{u}}; \theta_{\mathcal{D}}, \theta_S) &= \mathbb{E}_{\hat{\mathbf{u}}}[\log \mathcal{D}(\hat{\mathbf{u}})].\end{aligned}$$

Hence, the total loss for the segmentation network is a contribution of the individual losses  $\mathcal{L}_p$  in Eq. (5.2) and  $\mathcal{L}_{\text{Adv}}$  in Eq. (5.4). To avoid the need for manual hyperparameter tuning for the contribution weights ( $\lambda$ ) between two loss terms, we used *adaptive loss weighting*, as introduced in the work of Kendall *et al.* [182]. After initializing  $\lambda = [\lambda_1, \lambda_2]$  with a vector of ones, we add the regularization term  $\mathcal{R}(\lambda)$  to the total loss function for the segmentation network ( $S$ ), giving:

$$\mathcal{L}_S = \frac{1}{\lambda_1^2} \cdot \mathcal{L}_p + \frac{1}{\lambda_2^2} \cdot \mathcal{L}_{\text{Adv}} + \mathcal{R}(\lambda), \quad \text{where} \quad \mathcal{R}(\lambda) = \lambda_1^2 \cdot \lambda_2^2. \quad (5.4)$$

**C. Inference on the whole point cloud:** Since the segmentation network is trained on non-uniformly re-sampled data, for inference on the whole point cloud, we need to extract several subsets of points according to the non-uniform re-sampling algorithm. So, for inference on the whole point cloud, first, we marked all the points in the input point cloud as unprocessed samples. In the first step, a point is selected randomly as fovea among the unprocessed points and consequently by using non-uniform re-sampling, the network predicts the label for a subset of points. At this stage, we only accept the predictions for the points which lay inside the dense region of re-sampled data. The labels are stored and these points are marked as processed samples. The rest of the points in the sampling set are considered unprocessed points. Afterwards, by randomly selecting another fovea point out of the unprocessed points in the original point cloud, another point set is sampled and given to the network, this procedure is repeated until all points in the point cloud are marked as processed (i.e. selected at least once in the dense region). The probability vectors of points that are processed more than once are stacked and finally, an *argmax* operator yields the final predicted label for the point. Here, we define the dense region for those

points which fall within the  $\pm\sigma$ , close to the fovea point, based on the RBF kernel (Eq. (5.1)). The pseudocode of Algorithm 4 clarifies this procedure in detail.

---

**Algorithm 4** Inference on the whole point cloud

---

**input** : point cloud  
**output**: predicted label per point

```

 $X \leftarrow \{x_1, x_2, \dots, x_N\};$  // Whole point cloud
Function Inference ( $X$ ):
     $U \leftarrow \emptyset;$  // Initialized an empty set
     $P_X \leftarrow 0_{N \times 17};$  // Initialized probability vectors
    while  $|U| < |X|$  do
         $x_f \sim \{x \in X \mid x \notin U\};$  // Select from the unprocessed points
         $Y \leftarrow \mathcal{R}(x_f, X);$  // Non-uniform re-sampling (Algorithm 3)
         $P_Y = \mathcal{S}(Y, \theta_S);$  // Prediction of  $\mathcal{S}$  Net.
         $\{x_i\} \leftarrow \{x \in Y \mid \mathcal{K}(x_f, x) < \sigma\};$  // Only dense region is valid
         $P_X(x_i) = P_X(x_i) + P_Y(x_i);$  // Aggregate the probabilities
         $U \leftarrow \{x_i\} \cup \{U\};$  // Mark as processed
    end
    return  $\text{argmax}(P_X);$  // Labels on whole point cloud

```

---

### 5.3.4 Experiments and evaluation of Model 1

#### A. Data

A dataset of 120 optical scans of dentitions from 60 adult subjects, each containing one upper and one lower jaw scan, is used for evaluation of the proposed method. The data includes scans from healthy dentitions and a variety of abnormalities among subjects. The optical scans were recorded by a 3Shape d500 optical scanner (3Shape AS, Copenhagen, Denmark). On average, an IOS contains 180k points (varying in a range between [100k, 310k]). The precision of each data point is based on floating 32-bit numbers, where three of such numbers compose a 3D coordinate vector. All the optical scans were manually segmented and their respective points were categorized into one of the 32+1 classes by a dental professional and reviewed and adjusted by one dental expert (DAM) with Meshmixer 3.4 (Autodesk Inc, San Rafael CA, USA). Labeling of the tooth categories was performed according to the international tooth numbering standard (FDI). Segmentation of each optical scan took 45 minutes on average, which shows that this is an intensive laborious task for a human.

#### B. Experimental setup

The performance of the model is evaluated by fivefold cross-validation and the results are compared making use of the average Jaccard Index, also known as intersection over union (IoU). On top of the IoU, we report the precision and recall for the proposed multi-class segmentation task. For computing the precision and recall, each class is treated individually (one-versus-all), as a binary prob-

**Table 5.1** Performance of the proposed model within different experimental setups in comparison with state-of-the-art models.

Method			Metric			Exec.time
Network Arch.	Non-uniform	Adv. setting	IoU	Precision	Recall	(sec.)
PointNet [164]	-	-	0.76	0.73	0.65	<b>0.19</b>
PointGrid [170]	-	-	0.80	0.75	0.70	0.88
PointCNN [169]	-	-	0.88	0.87	0.83	0.66
Proposed (I)	✓	-	0.91	0.90	0.87	6.86
Proposed (II)	-	✓	0.91	0.91	0.89	0.66
Proposed (III)	✓	✓	<b>0.94</b>	<b>0.93</b>	<b>0.90</b>	6.86

lem and finally, the average scores are reported. Our experiments are partitioned into following three parts. (1) Benchmarking the performance of the PointCNN in comparison with two other state-of-the-art deep learning models, capable of IOS segmentation. These models include PointNet [164] and PointGrid [170]. (2) Evaluating the impact of applying the non-uniform re-sampling versus using naive uniform re-sampling. For the purpose of a fair comparison, the number of resampled points is kept identical ( $M=30k$ ). (3) Evaluating the effectiveness of involving the adversarial training in the proposed model.

All models are trained to utilize stochastic gradient descent and the Adam learning adaptation technique for 1000 epochs with a batch size of unity. The initial learning rate is equal to  $5 \times 10^{-3}$ , which decreases each 20k iterations by a factor of 0.9. We empirically adjust the free parameter of the re-sampling kernel (see Eq. (5.1)) to 0.4 (i.e.  $\sigma=0.4$ ). Since the point cloud is normalized to have unity variance, we have found that the resampled point cloud by such a chosen setting of  $\sigma$  would encompass at least two teeth in its dense region.

### C. Results of Model 1

Table 5.1 depicts the obtained results from our different experimental setups. As it can be observed from Table 5.1, the PointCNN performs better than two other state-of-the-art deep learning architectures when a naive uniform re-sampling is applied. This is mostly because of the inclusion of the spatial-correlation information by the  $\chi-Conv$  operator in the PointCNN and its lower amount of parameters, which is less prone to over-fitting. The PointGrid which samples points inside a predefined grid utilizes convolutional operators, but its performance is still limited to the spatial resolution of the quantized 3D volume. The PointNet performance is also constrained, as it omits the processing of spatial correlations in the point cloud.

So, with the choice of basing our experiments on PointCNN, next we evaluate the effectiveness of applying non-uniform re-sampling and adversarial training on tooth segmentation. As the results in Table 5.1 show, each of the pro-



posed techniques improves the performance of tooth segmentation with respect to different reported metrics. Finally, by incorporating both techniques jointly, the highest performance is achieved. Figure 5.3 shows a number of exemplary results from the proposed model.

### 5.3.5 Discussion

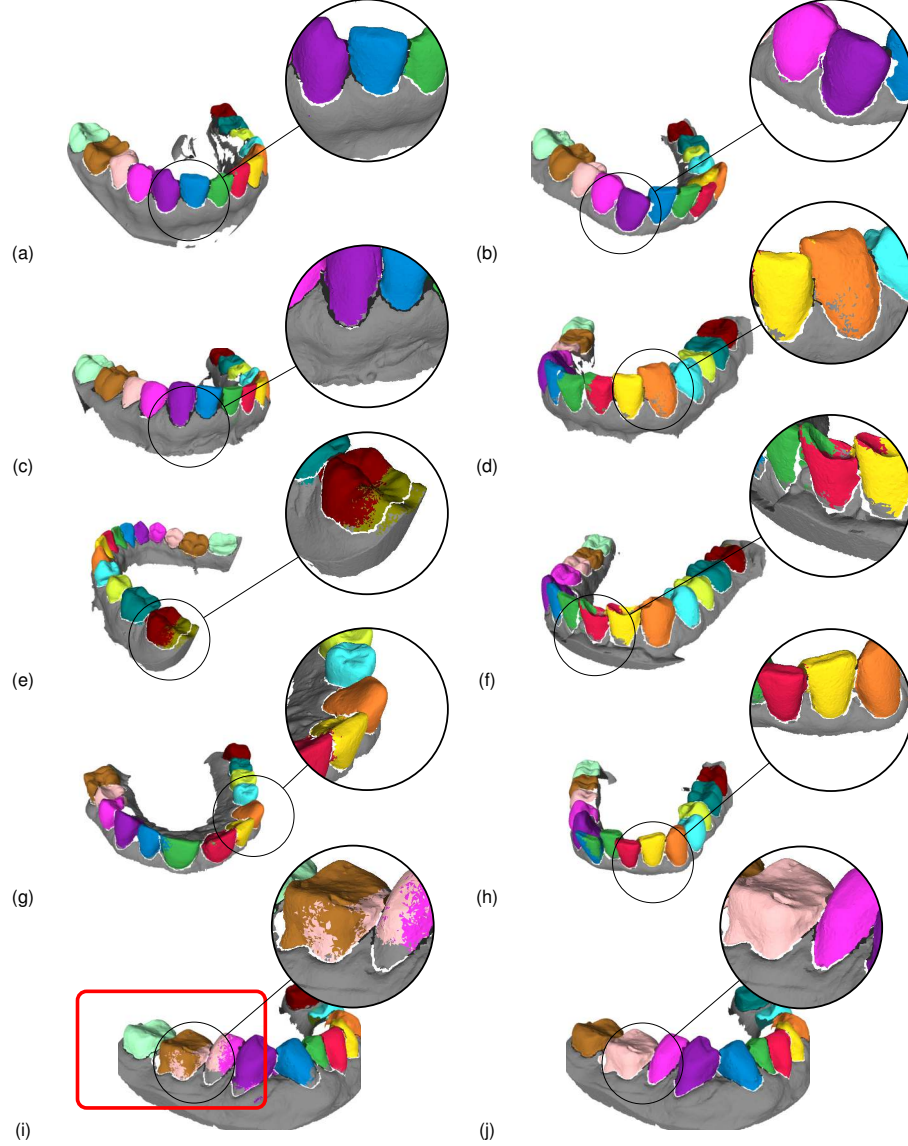
In this section, we have presented an end-to-end learning approach for semantic tooth segmentation from point clouds derived from IOS data. The proposed segmentation network is based on a state-of-the-art PointCNN model, which has been proposed for point cloud classification/segmentation tasks. For analysis of point clouds in their original spatial resolution (resulting in predictions for all points), we have introduced a non-uniform re-sampling technique and a compatible loss weighting, based on foveation and Monte Carlo sampling. This re-sampling approach includes both local, fine-detail information and the sparse global structure of data, which is essential for an accurate prediction of each individual point.

Different evaluation metrics have shown that the proposed non-uniform re-sampling improves the segmentation performance. This technique allows to process an IOS scan in its highest resolution, while the CAD system is limited to a certain computational and memory budget. Processing the IOS data in high-resolution amounts for preserving the fine-detailed information, like the curvatures in borderlines between teeth that are important for accurate segmentation.

Furthermore, we have shown that semantic segmentation of teeth can be improved when the distribution of labels is used in the form of adversarial training. This is mainly due to the high dependency of the semantic label of each tooth on its location on the dental arch with respect to other teeth. Our experiments confirm that by involving the high-level data semantics, through training a discriminator network for learning the realistic layout of labels in data, the results are improved. In contrast with common adversarial training that involves training a deep discriminator network on labeled data in the high-dimensional input space, we propose using only the statistics (mean and variance) from spatial distributions of labels as input to the discriminator. Such an abstract representation is beneficial, since only a shallow network can be employed as a discriminator and this facilitates the optimization process. As a consequence of using adversarial training, a heavy post-processing stage (e.g. applying conditional random fields (CRF) on a constructed graph) is not required for incorporating dependencies and locality constraints into the model.

## 5.4 Model 2: Instance segmentation in 3D point cloud scans

In the previous section, we have investigated the tooth segmentation problem as a multi-class semantic segmentation task for a deep learning model. In this approach, each tooth instance is considered a semantic class. The points in the point cloud are assigned to one of those classes, i.e. a *tooth number* or the *gingiva*.



**Figure 5.3** (a-h) Examples of segmentation by the proposed method (Model 1). (i) Example of a failure case when the adversarial loss is not involved in the training of the segmentation network. The assigned label inside the circle is unrealistic (i.e. invalid). Consequently, the model assigned a set of invalid labels to other neighbouring teeth (inside the red rectangle) by their maximum likelihood. (j) Ground truth.

Considering gingiva and a maximum of 16 teeth on each dental arch, each point has a probability of belonging to each of 17 semantic classes. This probability is expected to be predicted at the output of the segmentation network.

Although training a neural network by defining the tooth segmentation problem as a semantic segmentation task is straightforward, the model performance for tooth segmentation suffers from two main shortcomings. The first shortcoming originates from the fact that there is low inter-class variability between the crown shape of neighboring teeth, especially among the molar and premolar teeth. Hence, an accurate prediction of the labels requires not only the local geometrical information (i.e. crown shapes), but also the global context e.g. the relative position, teeth arrangement and possible absence of other teeth. For mitigating this issue, we have introduced adversarial training to regularize the predicted labels at the output of the network.

The second shortcoming is that because of preserving the global context, employing patch-based analysis and processing those patches individually is not feasible. Hence, due to hardware limitations (e.g. GPU memory), training and inference on the whole point cloud require down-sampling of the point cloud. Such a down-sampling would be detrimental to a precise teeth segmentation where preserving high-frequency information such as curvature at the borders of teeth is crucial. Employing *non-uniform* re-sampling of the point cloud has been proposed for addressing this issue, which was studied in the previous section.

In this section, we follow a different approach to the teeth segmentation problem. We introduce *Mask-MCNet*, a deep learning model for object detection and localization in a 3D point cloud. We formulate the tooth-segmentation task in the context of an instance localization and segmentation problem. It means that only one semantic class is defined that is the class of *tooth*. The points that do not belong to any tooth (e.g. to the gingiva), are considered as the undefined class. In the instance segmentation, apart from assigning a semantic label to each point, indicating whether it belongs to a tooth, the model should assign a secondary unique *arbitrary* label to those points as an indication of individual tooth instance. It is worth mentioning that in contrast to the first approach, the assigned labels to the instances convey no semantic meaning (for example they convey no information regarding whether the tooth is an incisor or molar) and are used only to distinguish/indexing the teeth from each other.

In contrast to the semantic segmentation task, formulating the problem as instance localization and segmentation does not suffer from missing global context and dependencies between the labels, since it localizes each tooth by fitting a 3D bounding box and simultaneously assigns a unique label to all points belonging to each instance inside the detected 3D bounding box. Accordingly, the model processes the cropped 3D patches locally and localizes all the teeth in the patch. Later, by aggregating all detections in the processed patches, the model performs the inference on the whole point cloud. As a consequence, the point cloud data are divided in local data processing actions in a natural way, so that the later processing of patches is possible in full quality, preserving the original spatial resolution of the data (without down-sampling). This property greatly facilitates the machine learning algorithms for performing automated parallel instance segmentation. In this approach, the teeth labels lack any semantics and only distinguish the points that belong to different tooth instances. Hence, for clinical

purposes and consistency of the tooth label-assignment, we introduce a postprocessing stage for translating (via a look-up table) the instance labels predicted by the network into the FDI standard labels. In the presence of missing teeth and abnormality in dentition, the assignment does not have a straightforward solution. For solving this assignment problem and finding the correspondences, we employ a combinatorial search algorithm that searches for the most likely label assignment, which satisfies a predefined constraint (prior measurements on training data) formulated by the constraint satisfaction problem (CSP) [183].

In the remainder of this chapter, first, a brief survey is presented for the existing deep learning models for object-instance segmentation in the 3D point cloud. Then the Mask-MCNet is introduced in detail for object-instance segmentation. Afterwards, the experimental setup and the obtained results are discussed. Lastly, discussions and conclusions finalize the chapter.

#### 5.4.1 Related work on instance segmentation in 3D point cloud

Among the proposed deep learning models for point cloud analysis, only a few researchers have addressed the challenging issue of 3D instance segmentation. To better compare and position the proposed method, a brief survey is provided for recent deep learning models, all related to instance segmentation in a 3D point cloud.

*FrustumNet* [184] consists of a hybrid framework involving two stages. The first stage detects the objects bounding boxes in a 2D image. The second stage processes the 3D point cloud in a 3D search space, partially bound by the initially computed 2D bounding boxes. The *3D-SIS* model [185] also first processes the 2D images rendered from the point cloud through a 2D convolutional network (ConvNet). Afterwards, the learned features are back-projected on the voxelized point cloud data, where the extracted 2D features and the geometric information are combined to obtain the object proposal and per-voxel mask prediction. The dependency on the 2D image(s) of both preceding models limits their applications for 3D point cloud analysis.

In another approach, *GSPN* [186] is a deep learning framework that follows an analysis-by-synthesis strategy and instead of directly finding the object bounding boxes in a point cloud, it utilizes a conditional variational auto-encoder (CVAE). However, *GSPN* training requires an individual two-stage training of the CVAE part and the region-based networks (which performs the detection, localization and mask generation on the object proposals).

Alternative approaches to construct object proposals can be found in the work of *SGPN* [187] and *MASC* [188]. These methods perform clustering on the processed points for segmenting the object instances. *SGPN* [187] uses a similarity matrix between the features of each pair of points in the embedded feature space, to indicate whether the given pair of points belong to the same object instance or not. Although computing the pair-wise distance for small point clouds is feasible, it is impractical for large point clouds and especially for IOS data, where down-sampling would significantly affect the detection/segmentation perfor-

mance. MASC [188] voxelizes the point cloud for processing the volumetric data by a 3D U-Net model. Similar to SGPN, MASC uses a clustering mechanism to find similarities between each pair of points, by comparing their extracted features in several hidden layers of a trained U-Net. Unfortunately, as mentioned before, the voxelization of a large fine-detailed point cloud significantly limits the performance of such approaches.

In a different approach, *VoxelNet* [189] first divides a point cloud into equally spaced 3D voxels and then transforms a group of points within each voxel into a feature space. Subsequently, it uses a Region Proposal Network (RPN) to generate 3D box detections. The preliminary division of the 3D input space into voxels facilitates the processing of sparse point clouds, such as those collected by Lidar sensors, since the 3D convolutions can be applied to the constructed volumetric space. However, such voxelization techniques for dense point clouds (like those obtained from a dental scan) may lead to inhomogeneous feature extraction for neighboring points, which have been assigned to two adjacent voxels. This phenomenon can happen for a large number of points in a dense point cloud. To mitigate this effect, decreasing the number of bins would reduce the relevant population of border points in each voxel. However, this reduction come with the expense of losing spatial accuracy.

Instead of converting a 3D point cloud to a regular grid, *VoteNet* [190] directly votes for a virtual center of objects from the point clouds. It generates a group of high-quality 3D object proposals by aggregating vote features and predicting offset vectors to the corresponding object centers for seed points. This is followed by a clustering module to generate object proposals. *PointRCNN* [191] directly segments a 3D point cloud to obtain the foreground points. Afterwards, a bin-based 3D bounding-box generation is performed only around the foreground points to produce high-quality 3D boxes. The authors have shown that such an approach achieves state-of-the-art performance on the car detection task in a sparse LiDAR point cloud.

Summarizing the mentioned research work of the literature indicates that instance segmentation is more suitable for the considered teeth segmentation problem. Unfortunately, the available studies do not reveal a suitable computational model for processing the large dense 3D point cloud data, containing fine-detailed information. This results in the following novel design and contributions.

#### 5.4.2 Concepts and novelty of the proposed approach

The proposed model is similar to *PointRCNN* by using a backbone network as a feature extractor and using bins as a search space for the RPN. However, in contrast, we do not segment the input points into two classes of foreground and background. This is because the objects (i.e. teeth) are located closely on the dental arch, so the foreground points are the vast majority, and the technique would not be effective to reduce the search space. Instead, we use larger bins to search the entire 3D space. To compensate for such a coarse quantization step,



1. A new instance segmentation model, called Mask-MCNet, is presented. The proposed model is applied directly to an irregular 3D point cloud on its original spatial resolution and predicts the 3D bounding boxes of object instances along with their masks, indicating the segmented points of each instance.
2. To the best of our knowledge, this is the first solution that both detects and segments tooth instances in 3D IOS data by a deep learning model.
3. An extensive empirical evaluation is conducted, which shows that the proposed model significantly outperforms the state-of-the-art in IOS segmentation.

The remainder of this section starts with a detailed description of the individual processing modules of the Mask-MCNet (Section 5.4.3). Afterwards, the performed experiments and results are presented. Finally, a discussion on the model performance and conclusions are provided.

### 5.4.3 Mask-MCNet

Deep learning models for 3D point cloud analysis computationally differ from the ConvNet-based models, since they are applied on non-grid data (unstructured 3D points with varying distances in between). However, at a high level, the Mask-MCNet is similar to the ConvNet-based Mask R-CNN [192] because it includes three main parts: the backbone network, a Region Proposal Network (RPN), and three predictor heads for detection, localization by fine-tuning, and mask generation (see Figure 5.4). Each part is explained in detail below.

#### A. Backbone network

The *backbone* is a deep network based on the Multi-layer Perceptron (MLP) architecture, which is applied to the entire or cropped (depending on hardware limitations) point cloud and acts as a feature extractor. Every input patch includes  $n$  points (varying across patches), where each point is represented by its  $(x, y, z)$  3D coordinates and might have other attributes such as color or a normal vector.

A point cloud does not explicitly convey the information from neighboring points. Therefore, in order to aggregate over local neighborhoods, most existing methods augment the input point cloud with the surface normal vectors, or resort to a neighbor searching mechanism (e.g., KNN [193] or ball query [169]). In the experimental section, we evaluate the impact of augmenting the input with the normal vectors. Hence, in case of using normal vectors, the input to the backbone model is an  $n \times 6$  matrix. The output of the backbone model is a high-dimensional feature vector per given input point (e.g. a matrix of  $n \times 256$ ) which contains rich geometrical information around each point. In this study, we choose to employ a PointCNN [169] for its fine-detail processing capacity and its small model size. To demonstrate the genericness of the proposed approach, we instantiate the proposed Mask-MCNet with two different backbone architectures. In an ablation

study, we also report on using PointNet [164] as an alternative choice for the backbone.

### B. Region Proposal Network (RPN)

Since the points in the point cloud are distributed solely on the surface of objects, the computed features from the backbone only contain local geometrical representations on a manifold in 3D space. However, for accurate localization of a 3D bounding box that encompasses an object, the network is required to be aware of several parts (or sides) of each object. Such awareness leads to reliable learning and consequently an accurate prediction of the center and size of each bounding box. Hence, voxelization of the data and employing a 3D ConvNet on the obtained volumetric data is a common approach. However, the shortcomings of such an approach have been mentioned already.

Therefore, as an alternative method, for distributing and transferring the computed geometrical information from the surface of objects to the entire 3D space (e.g. into void space inside of the objects as well as the centroid of a 3D bounding box), we employ a *Monte Carlo ConvNet* (MCCNet) [194]. The MCCNet is a multi-layer network of which each layer consists of several MLP-based sub-networks. Each sub-network consists of an MLP with two hidden layers whose functions resemble a set of convolution kernels. Each sub-network (called kernel) receives a set of features at the location of points in its spherical *receptive field* and computes an output feature vector at the center of its receptive field. Similar to the standard convolution kernel, by positioning the kernel on any points in the point cloud, the kernel maps the input feature set into that point. However, in contrast to a standard convolution, such a mapping is performed by applying the transfer function of an MLP-based kernel. In order to elaborate on the motivation behind employing the MCCNet in the framework of the proposed Mask-MCNet, in the following, we will briefly explain the principle of Monte Carlo convolution in the work of Hermosilla *et al.* [194].

### Monte Carlo Convolution Network (MCCNet)

This subsection describes the MCCNet in a more formal way. We assume that the feature function  $f$  is defined on the surface of an object, in which we have a set  $\mathcal{S}$  of discrete samples  $y_i \in \mathcal{S}$  (the obtained data points). The function  $f$  represents a mapping between the 3D position of each point and its representation in the embedded feature space ( $f : \mathbb{R}^3 \mapsto \mathbb{R}^N$ ). In our particular case,  $N=256$  is the dimension of the backbone features. By defining a convolution kernel  $g$  with a spherical receptive field (centered at 0 with a radius equal to unity), the discrete convolution operator can be written as:

$$(f * g)(x) = \sum_{i \in \mathcal{N}(x)} f(y_i) \cdot g(x - y_i), \quad (5.5)$$

where  $\mathcal{N}(x)$  is the set of neighborhood indexes in the receptive field of  $g$  which is centered around the point  $x$ . The convolution kernel  $g$  is suggested to be im-



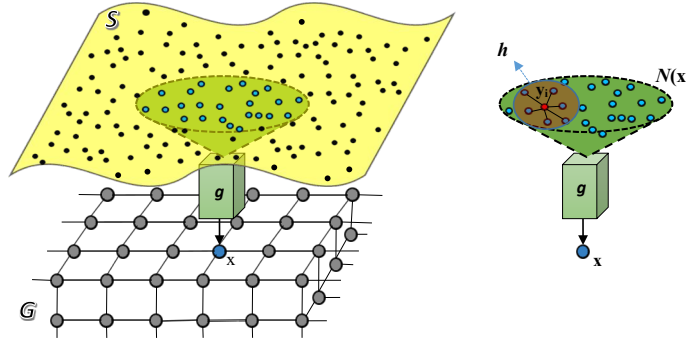
plemented by an MLP with 2 hidden layers [194]. Since the defined discrete convolution in Eq. (5.5) assumes that the sampled points are distributed uniformly on  $f$ , Hermosilla *et al.* [194] suggested to estimate and incorporate the probability density function (PDF) in the field of view of  $g$  to address a valid approximation of the convolution operator when the points are distributed non-uniformly on  $f$ . Therefore, Eq. (5.5) is modified to be written as:

$$(f * g)(x) \approx \frac{1}{|\mathcal{N}(x)|} \sum_{i \in \mathcal{N}(x)} \frac{f(y_i) \cdot g(\frac{x-y_i}{r})}{p(y_i|x)}, \quad (5.6)$$

where  $r$  is the distance of each point  $y_i$  in the receptive field of  $g$  from its center  $x$ . The expression  $|\mathcal{N}(x)|$  denotes the number of the points in the receptive field of the kernel, centered at point  $x$ . Here,  $p(y_i|x)$  is the PDF at point  $y_i$ . Since the PDF is unknown, the authors proposed using a kernel density estimation  $h$  for estimating the PDF at the location of  $y_i$ , as specified below:

$$p(y_i|x) \approx \frac{1}{|\mathcal{N}(x)| \cdot \sigma^3} \sum_{k \in \mathcal{N}(x)} \left\{ \prod_{d=1}^3 h\left(\frac{y_{i,d} - y_{k,d}}{\sigma}\right) \right\}, \quad (5.7)$$

where  $h$  is the *density estimation kernel*, a non-negative function of which the integral equals unity (e.g. a Gaussian) and  $\sigma$  is its bandwidth which determines the smoothness of estimation (e.g.  $\sigma = 0.25r$ ). It is worth mentioning that the derivatives of Eq. (5.6) with respect to the parameters of network  $g$  and employing the back-propagation technique for training the MCCNet are straightforward. For more information regarding the mechanism of MCCNet, we refer to the original paper [194].

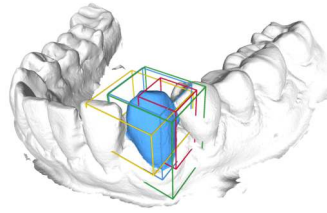


**Figure 5.5** (left) Mapping the backbone features from the location of set  $S$  (point cloud) into point  $x$  from a 3D grid domain  $G$  by a learning kernel  $g$ ; (right) estimating the PDF at each point  $y_i$  by kernel density estimation  $h$  (see Eq. (5.7)). The spherical receptive field of  $g$  is shown in a green color. In an ablation test, the Monte Carlo convolution is replaced by a max-pooling operator (i.e. the mapped feature vector on point  $x$  is  $f_x = \max(y_i)$  where  $i \in N(x)$ ).

As mentioned earlier, we aim to transfer and then distribute the extracted backbone features at the location of samples in the set  $\mathcal{S}$  into the entire 3D space at the location of nodes of a 3D grid ( $G$ ). The domain  $G$  spans the whole 3D space and is bounded by the bounding box of the input scan. Figure 5.5 illustrates such a mapping between  $\mathcal{S}$  and  $G$ .

For performing such a mapping, we propose using a set of Monte Carlo convolution kernels because of two important properties: (1) its capability of computing the convolution on an arbitrary output point ( $x$ ) within the kernel's receptive field, regardless of its presence within the set of input points ( $\mathcal{S}$ ); (2) its capability of handling the non-uniform distribution of points when computing the mapping. The first property makes it possible to transfer the computed backbone features on an arbitrary new domain such as the node of the regular 3D grid  $G$ , while the second property facilitates the processing of a non-uniform distribution of points on the surface of an object.

By aggregating the features of surface points on a regular 3D grid, only the nodes close to the surface will have a valid feature vector. The nodes near the object center are likely to have no feature vector at all (located in the void space). As a result, aggregating the shape context in the vicinity of object centers creates difficulties. Simply increasing the receptive field does not solve the problem because as the network captures a larger context, it also causes more inclusion of nearby objects and clutter. Hence, after mapping the features into domain  $G$  through the set of convolutional kernels ( $g$ 's) in the first layer of the MCCNet, the geometrical information is passed on and further processed in the deeper layer of MCCNet and is distributed to the neighbor nodes in domain  $G$  based on the pre-defined field of view of kernels in the hidden layers of MCCNet. Applying the MCCNet on the output of the backbone (e.g.  $n \times 256$  matrix), a high-dimensional feature vector (e.g. 256-dimensional vector) at the location of each node of the 3D grid domain is computed. By assuming  $m$  nodes for domain  $G$ , the network returns in total a feature matrix of size  $m \times 256$ .



**Figure 5.6** Example of typical 3D anchors that are used for tooth localization. The ground truth (colored in blue) has 3D IoU scores equal to 0.35, 0.42 and 0.60 with red, yellow and green anchors, respectively.

#### Object proposal (anchor) and Triangular Interpolation

To generate object proposals (i.e. 3D cubes encompassing teeth), we follow the idea of using *anchors* which is adopted from *Faster-RCNN* [195] that was originally proposed for 2D object localization in images. For the point cloud data, we

extend this idea by using anchors in 3D space. Here, each 3D anchor is indicated by a cube, which is represented with its central position  $[x_c, y_c, z_c]$  and its size  $[w, d, h]$ .

The orientation of 3D boxes is ignored in the proposed modeling approach. This is because some of the tooth crowns (e.g. premolars teeth) have symmetric and semi-cylindrical shapes. Considering the orientation for fitting a 3D box imposes a high degree of uncertainty to the learning module and contributes little towards point segmentation, which is the main goal of processing an intra-oral scan.

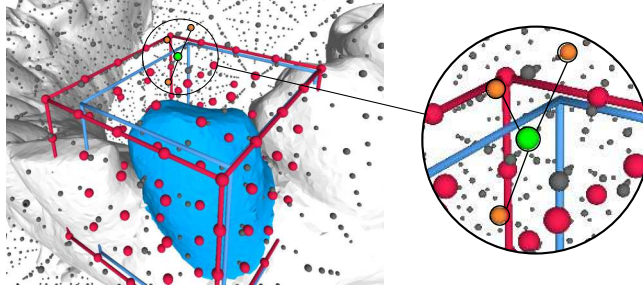
Figure 5.6 visualizes examples of such anchors with different sizes and aspect ratios that are localized at different positions in 3D space. Making no assumptions regarding the possible positions of objects (which leads to a more generic approach), the centers of the anchors should be located on the nodes of a regular 3D grid, which is domain  $G$  of where the MCCNet computes the geometrical information at its node. We will discuss the spatial resolution of grid  $G$  in a later section.

Since the 3D tooth size varies along the dental arch and between different subjects, we use multi-sized ( $k > 1$ ) anchors which are centered around each node of  $G$ . As each node on  $G$  indicates the center of one ( $k = 1$ ) or multi-sized ( $k > 1$ ) anchor(s), the total amount of anchors is  $k \times m$ . For the teeth segmentation problem, by considering various sizes of teeth (incisor and molar teeth), we position 4 different anchors (2 different sizes and 2 different aspect ratios).

Generally, the idea of using anchors is to change the problem of object localization into the anchor classification. To do so, the anchor which has a high overlap with an object should be classified as a positive class, otherwise, the anchor is classified as a negative class. Anchor classification and, later on, fine-tuning the position and size of positive anchors to precisely fit the 3D bounding box of the object, requires a fixed-length feature vector to be extracted from variable-sized anchors. Here, we use an idea similar to *ROI alignment*, which was introduced in Mask R-CNN [192], but in a 3D space of point cloud data. We re-sample a fixed number of points inside each anchor by applying the *triangular interpolation* by finding a set of three nearest neighbor nodes of the points domain  $G$  and weighting their feature vectors based on their distances to the new node in 3D space. Figure 5.7 shows an example of performing ROI alignment for an anchor and the triangular interpolation. In our experiment, we use a 3D grid of  $s \times s \times s$  nodes (e.g.  $s = 5$ ) for interpolating the features inside each anchor.

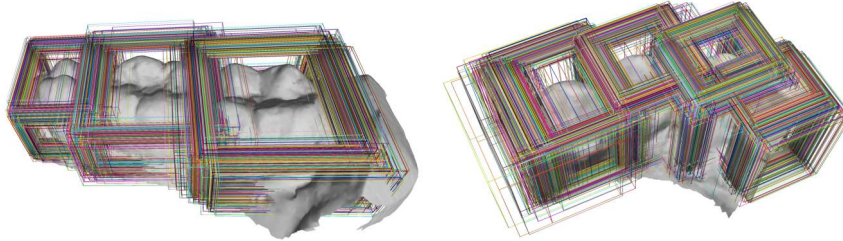
### C. Predictor networks

The predictor networks consist of three parallel branches for *detection*, *localization* by fine-tuning, and *mask generation*. The detection and localization branches both consist of a fully-connected MLP network that receives the interpolated feature set inside each anchor and performs a regression task. The detection branch aims to predict the IoU score of each anchor (in the unity interval), which in-



**Figure 5.7** Example of 3D ROI alignment. By applying the triangular interpolation in the domain  $G$  (gray nodes), a set of features is computed at the locations of red nodes (with  $5^3$  nodes) on a grid inside the red anchor. Hence, a fixed-length feature set is assigned to the anchor. Here, the blue box is the ground truth and the orange nodes are examples of using three nearest neighbor nodes of  $G$  to the green node for interpolating the feature vector on the green node.

indicates their maximum overlap with any object. Later, by applying a threshold to the predicted IoU scores, the anchors are classified into positive and negative classes. The assigned class indicates if an anchor acceptably encompasses an object instance or not (in total,  $k \times m$  anchors). Figure 5.8 shows two examples of input 3D patches and their overlaid positive detected anchors.



**Figure 5.8** Examples of positive detections for two 3D patches.

The localization branch aims to predict the adjustment values of the location and size of each positive anchor, to fit tightly to the true bounding box of the corresponding object. Since the 3D position of each anchor is encoded by its centroid and size, the localization branch aims to predict 6 values for each positive anchor at its output. Instead of estimating the absolute values of such parameters directly, the network only predicts the difference (i.e. residual vectors) between the center and size of the anchor with the center and size of the corresponding object. Learning such residual vectors in the form of difference values is easier and imposes less complexity on the learning model because its computation is performed locally in a canonical coordinate system.

Since the detection and localization branches work in parallel, the feature matrix that is supplied to these two branches has a size of  $k \times m \times s^3$  elements. As

mentioned earlier, in the training phase, an anchor is labeled positive if it has a high IoU score with any single tooth instance above a threshold (e.g. 0.4), and it is labeled negative if the IoU is lower than a certain threshold (e.g. 0.2). Since the number of positive and negative anchors is highly imbalanced, about 50% of each training batch is selected from the positive and 50% from the negative anchors. The marginal anchors ( $0.2 < IoU < 0.4$ ) are not utilized when training the model.

The mask-generation branch aims to segment the points that are located inside the 3D bounding box of each positive anchor. The outcome of the mask generator has the form of a binary mask that indicates whether each point inside the bounding box belongs to the corresponding tooth instance or not. To perform such a binary classification, the points inside a positive anchor along with their features from the backbone network are supplied to the mask-generation branch. Since the number of points in such a classification is highly imbalanced, we have trained the mask-generation branch on points sampled equally from both classes. In our architecture, the mask generator has similar computational layers to the backbone architecture, but it consists of only three layers. Figure 5.9 shows the details of Mask-MCNet architecture.

#### D. Loss function

Mask-MCNet performs multi-task learning, including the estimation of 3D bounding-box overlap (i.e. IoU score), center-offset and size-offset estimation, and mask generation. To perform these, the loss function of the model consists of an equal contribution of three terms. The first term ( $\mathcal{L}_{\text{det}}$ ) is a mean-squared error for estimating the IoU scores of each anchor at the output of the detection branch. The second term ( $\mathcal{L}_{\text{loc}}$ ) is also a mean-squared error at the linear output layer of the localization branch. Finally, the third term ( $\mathcal{L}_{\text{mask}}$ ) is a binary cross-entropy loss for the classification of all points in each positive anchor at the output softmax layer of the mask branch. The localization and mask losses are involved only if the examined anchor is labeled positive. Thus, the total loss function can be written as:

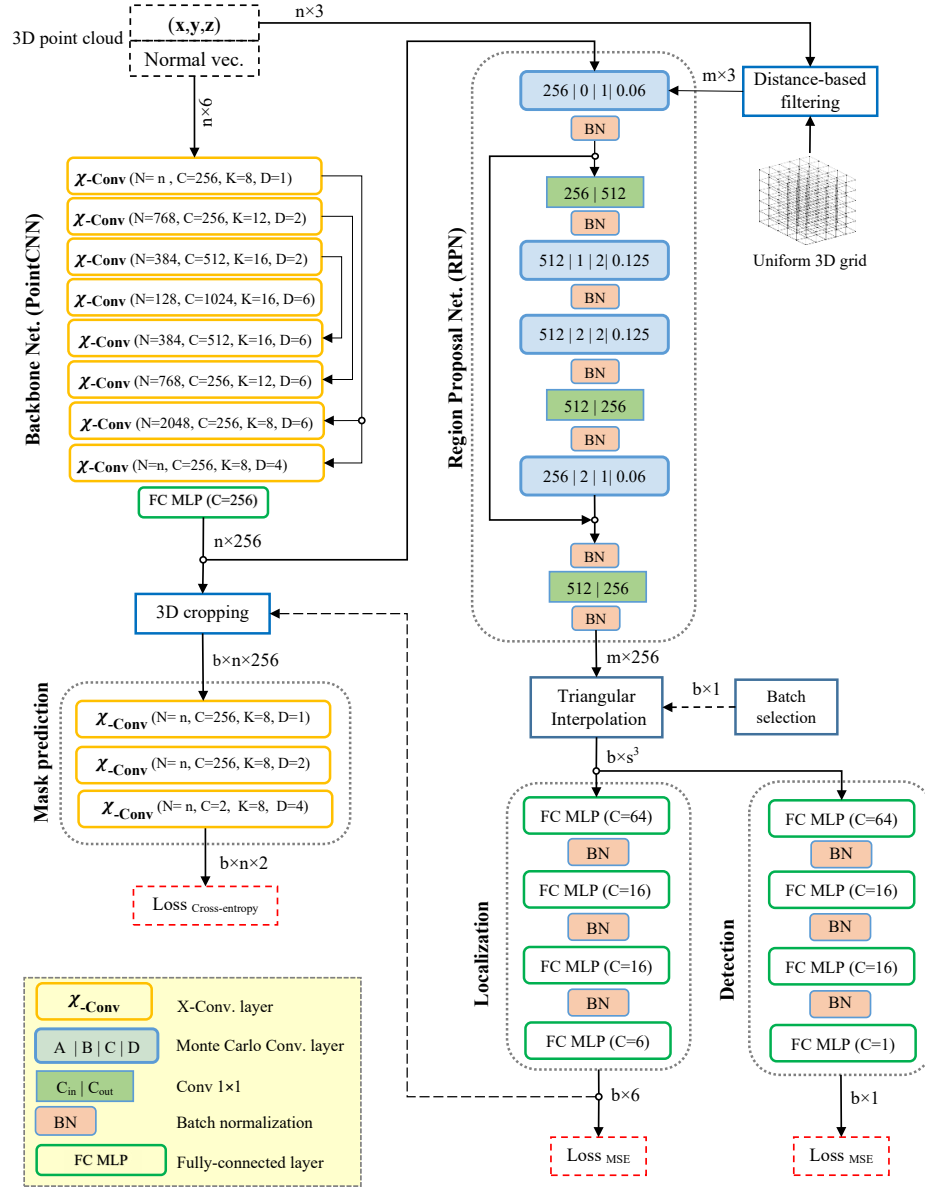
$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}}^{\{p,n\}} + \mathcal{L}_{\text{loc}}^{\{p\}} + \mathcal{L}_{\text{mask}}^{\{p\}}, \quad (5.8)$$

where the superscript  $\{p,n\}$  indicates that the term is calculated for both positive and negative anchors in the training batch.

#### E. Implementation Details

*E1-Training:* The entire Mask-MCNet is trained end-to-end by using gradient descent optimization and the Adam [46] learning adaptation technique for 1,000 epochs with a batch size of 32 (equally balanced between positive and negative anchors). The pre-processing of the input intra-oral scan only consists of normalizing the whole point cloud to have zero mean and unit variance. The input to the Mask-MCNet is a randomly cropped patch of the point cloud. Each patch contains 2-4 tooth instances. As explained earlier, the centers of the anchors are positioned on a regular 3D grid with a spatial resolution of 0.03 in a normalized coordinate system. As to impose sufficient overlap between anchors and both

#### 5.4. Model 2: Instance segmentation in 3D point cloud scans



**Figure 5.9** Mask-MCNet architecture (PointCNN as backbone).

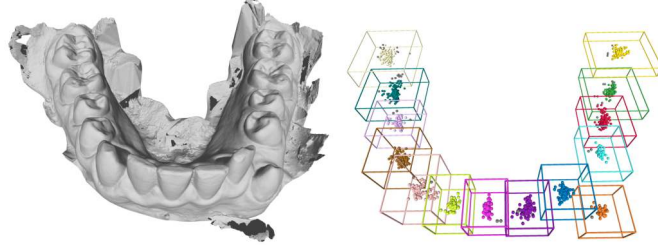
small and large objects (e.g. incisor and molar teeth, respectively), two types ( $k=2$ ) of anchors are employed (with a size of  $[0.3, 0.3, 0.2]$  and  $[0.15, 0.15, 0.2]$ ).

## 5. TEETH INSTANCE SEGMENTATION IN 3D POINT CLOUD DATA

---

For computational efficiency, reducing the number of anchors is desirable. Such a reduction can be considered in two ways. Firstly, by choosing a minimal number of types of 3D anchors that differ by their aspect ratios. This minimal number of types depends on the variation of object (tooth) sizes. Therefore, the box sizes are adapted to the tooth-instance sizes. Secondly, by reducing the number of nodes that are in the central position of the anchors, the total number of anchors required to be examined by the model for the presence of a tooth is reduced.

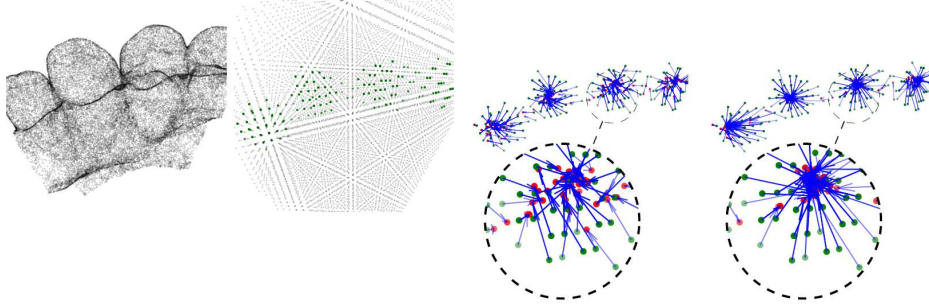
For the sake of obtaining a high recall in tooth detection, the resolution of grid  $G$  cannot be reduced too much. Instead, we can remove the nodes that are very unlikely to be close to the center of a tooth. To do so, we remove nodes of the grid, based on their distance to the closest point in the point cloud. Hence, the nodes and consequently the anchors that have a distance higher than a certain threshold are suppressed. Such a suppression mainly removes the points in void space, close to the center of the dental arch and significantly decreases the total computational time of the model.



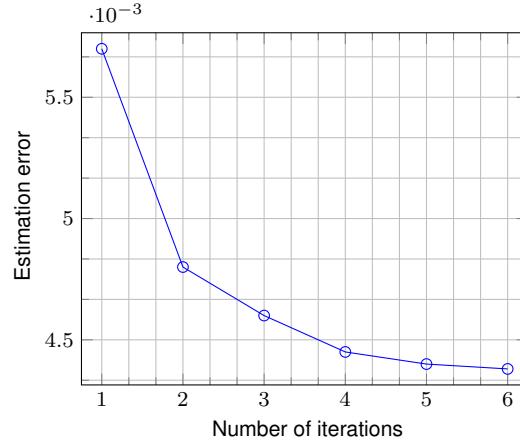
**Figure 5.10** DBSCAN clustering of all positive anchors (detections) for assigning a single 3D bounding box to each tooth instance.

*E2-Inference:* By applying the model to all patches (about 5 patches per jaw) that are extracted from the input point cloud and then aggregating the predictions, we can perform inference on the whole point cloud in its original resolution. Since for each tooth, several overlapping detections are obtained, we need to aggregate the detections to generate exactly one detection per tooth. Furthermore, handling the overlapping patches also requires using such an aggregation step. Since each detection has a predicted IoU score, we can use a non-maximum suppression algorithm, which is common practice in object detection models. However, to compensate for the effect of false-positive detection, instead of using a simple non-maximum suppression, we employ the DBSCAN clustering algorithm [196] to group the detected 3D bounding boxes, based on their centroid position and their size. Such clustering is preferred because it can detect outliers (i.e. false-positive detections) and based on our experiments the DBSCAN algorithm works better than other clustering methods. Figure 5.10 shows an example of a clustering result for a test intra-oral scan.

*Recursive regression:* The localization branch in the proposed Mask-MCNet (see Figure 5.9) aims to predict the offset values of the centroid and the size of



**Figure 5.11** Iterative regression of anchors centroids; (left) input 3D patch; (2nd column) 3D grid domain ( $G$ ) and positively detected centroids; (3rd column) predicted displacement vector of positive anchors; (right) predicted displacement after 2nd iteration.



**Figure 5.12** Changes of average estimation error (Euclidean distance) of the anchor centroids using an iterative regression.

each positive anchor. Since such an inference is based on given interpolated features inside each anchor, the predicted values for the anchors that are not highly overlapping with a tooth are more prone to errors. For compensating the source of such an error in the inference phase and improving the predicted offset values, we employ a simple *recursive* scheme that applies multiple executions of the regressor on the relocated anchor, according to the last predicted values. Such a recursive scheme works as a *negative feedback* in the model.

Applying this mechanism, the predicted offset values are used for relocating the bounding box of the anchor and then by re-applying the triangular interpolation and re-estimating the feature set at the location of the updated point set (with  $s^3$  points), the model predicts the offset values again. This combination of relocating the bounding box and updating the feature set corresponding to the box can be considered as one iteration in this recursive method.



Since in a following iteration, the updated bounding box is more likely to be closer to the actual bounding box of a tooth, the predicted values will become more precise. Figure 5.11 visualizes the predicted offset values for the positive anchors for two iterations. As can be observed, in the second iteration, the replacement vectors are more concentrated on a single central point of each tooth instance in the input patch.

Moreover, Figure 5.12 shows the changes in average estimation error (Euclidean distance) of anchor centroids across several iterations at inference time. The significant reduction in the distance error is clearly visible. The values are measured for the scans, normalized with zero mean and unit standard deviation. In our experiments, the regression branch was executed for two iterations that slightly improved the 3D box detection.

**Semantic label-assignment** The Mask-MCNet assigns a unique label to all the points that belong to each tooth. The assigned labels only distinguish each tooth from other tooth instances on the dental arch and lack any semantics. For clinical purposes and consistency of the tooth labeling assignments, we use a post-processing stage for translating (via a look-up table) the instance labels predicted by the Mask-MCNet into the FDI standard labels. By measuring the average central positions and sizes of the FDI labels within the training data, a combinatorial search algorithm identifies correspondences for the most likely label assignment, which satisfies the predefined constraint (prior measurements on training data) in the context of a constraint satisfaction problem (CSP) [183].

### 5.4.4 Experimental results

#### A. Data

This study is based on two datasets that have been collected from two different types of scanners. The first dataset called *Dataset I*, is used for both training and testing the models by using the cross-validation technique. The second dataset, called *Dataset II*, is only applied for evaluating the robustness of Mask-MCNet across different scanner types.

#### Dataset I

This dataset consists of 120 optical scans of dentitions from 60 adult subjects, each containing one upper and one lower jaw scan. The optical scans were recorded from *dental impressions* by a D500 optical desktop scanner (3Shape AS, Copenhagen, Denmark), which uses stereo-vision cameras and a three free-axes motion system for 3D reconstruction. The scanner has high spatial accuracy with a tolerance smaller than 20  $\mu\text{m}$  and obtains about 180k points per scanned jaw on average (varying in a range interval of [100k, 310k]). The dataset includes both healthy dentitions and a variety of abnormalities in dentition among subjects.

### Dataset II

This dataset consists of 48 optical scans of 24 adult subjects. The scans are captured by a Trios Move intra-oral scanner (3Shape AS) that uses confocal laser scanning microscopy and structured light projection. As mentioned earlier, this dataset is only used for evaluating the trained model across different scanner types.

All the optical scans were manually annotated by using Meshmixer 3.4 (Autodesk Inc, San Rafael CA, USA) and their respective points were categorized, according to the FDI standard into one of the 32 classes by a dental professional and reviewed and adjusted by another dental expert. Annotation of each optical scan took about 45 minutes on average, which shows that it forms an intensive laborious task for a human.

### B. Experimental setup

The performance of the Mask-MCNet in comparison with state-of-the-art systems is evaluated by fivefold cross-validation. The average Jaccard Index (also known as mIoU) of all teeth instances is measured. On top of the measured mIoU, by treating each class individually as a binary segmentation (one-versus-all) problem and then by averaging on all measured precision and recall scores, we report the mean average precision (mAP) and mean average recall (mAR) for evaluating the multi-class tooth segmentation performance.

Although the instance segmentation performance is coupled with the result of 3D bounding-box detection in the proposed Mask-MCNet, for evaluating the intermediate steps, we also report the performance of 3D bounding-box detection. To do so, which is common in object detection problems, we measure the average precision value for the recall values of 0 and 1, when the predicted bounding box overlaps with the ground truth by applying a threshold of 0.25 (which was proposed in [184]) and 0.5 on 3D IoU.

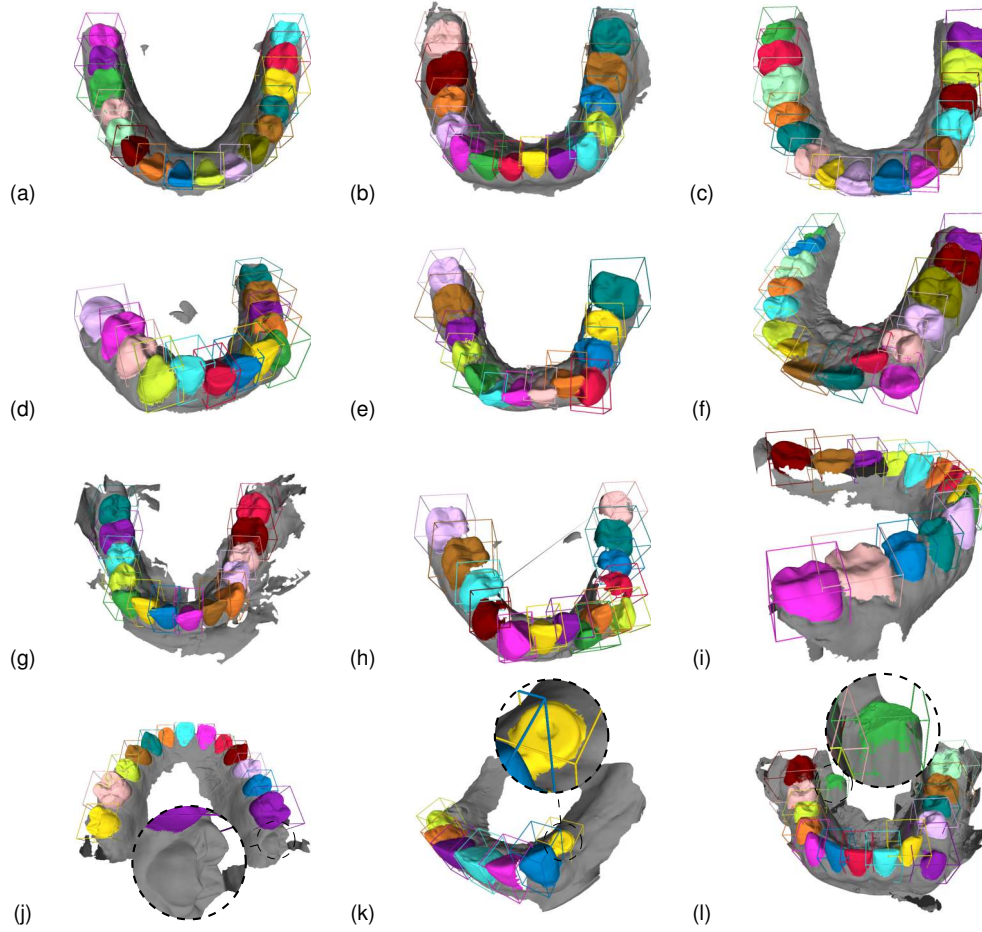
### C. Main Results

We have evaluated the proposed Mask-MCNet in comparison with the competitive models for 3D point cloud semantic segmentation on Dataset I. The performance of each method was tested by fivefold cross-validation. The obtained results of instance tooth segmentation are shown in Table 5.2. The results show that the Mask-MCNet outperforms state-of-the-art models by achieving 0.98 mIoU on the instance tooth segmentation. Figure 5.13 visualizes the segmentation result for a number of test scans with various abnormalities in dentition and in the scanning artifacts.

The measured execution times of Mask-MCNet are relatively high because of two reasons. Firstly, in contrast with compared semantic segmentation models, the Mask-MCNet is able to process the input scans at their original spatial resolution by using a patch-processing technique. This results in a dense prediction without employing a down-sampling at the expense of a longer computational

## 5. TEETH INSTANCE SEGMENTATION IN 3D POINT CLOUD DATA

time. Secondly, in our implementation, executing the triangular interpolation on a multi-threading CPU slows down the inference speed. However, the computation time of Mask-MCNet is still a small fraction of the time needed by a human expert for annotating an intra-oral scan.



**Figure 5.13** Examples of tooth instance segmentation on test data by Mask-MCNet; (a-c) normal dentition (d-f) subjects with different abnormality in dentition; (g) artifacts in scanning (h-i) typical missing data; (j-l) failure cases.

#### 5.4. Model 2: Instance segmentation in 3D point cloud scans

**Table 5.2** Performance of tooth segmentation on Dataset I, based on different metrics. The mean IoU (mIoU), mean average precision (mAP), mean average recall (mAR), and the execution time are reported for several competitive methods.

Method	Segmentation type		Metric			Exec.time *
	Semantic	Instance	mIoU	mAP	mAR	(sec.)
PointNet [164]	✓	-	0.76	0.73	0.65	<b>0.19</b>
PointGrid [170]	✓	-	0.80	0.75	0.70	0.88
MCCNet [194]	✓	-	0.89	0.88	0.84	1.01
PointCNN [169]	✓	-	0.88	0.87	0.83	0.66
PointCNN++ [197]	✓	-	0.94	0.93	0.90	6.86
MASC [188]	-	✓	0.93	0.92	0.89	1.31
VoxelNet [189]	-	✓	0.94	0.94	0.91	0.54
PointRCNN [191]	-	✓	0.97	0.96	0.97	0.23
Mask-MCNet (ours)	-	✓	<b>0.98</b>	<b>0.98</b>	<b>0.97</b>	14.6

\* NVIDIA Titan-X GPU

For evaluating the robustness of Mask-MCNet across different scanner types, the trained model on Dataset I is tested on Dataset II. The results in Table 5.3 show only a 1% drop in the mIoU score, which indicates acceptable robustness of the proposed model across different scanner types.

**Table 5.3** Results of tooth instance detection and segmentation on the Dataset II.

Model	Mask segmentation			Bounding-box localization		
	mIoU	mAP	mAR	mIoU	$mAP_{0.25}$	$mAP_{0.5}$
MASC [188]	0.90	0.89	0.88	0.48	0.92	0.78
VoxelNet [189]	0.92	0.92	0.88	0.50	0.90	0.80
PointRCNN [191]	0.95	0.95	0.92	0.52	0.97	0.83
Mask-MCNet	<b>0.97</b>	<b>0.96</b>	<b>0.94</b>	<b>0.53</b>	<b>0.99</b>	<b>0.84</b>

#### D. Ablation Experiments

In ablation experiments, we evaluate multiple basic instantiations, which allow to demonstrate the robustness of the model and analyze the effects of core aspects in the proposed Mask-MCNet. We have examined the performance of Mask-MCNet across:

- Different backbone architectures,
- Different surface normal-vector estimation,
- Coupling mechanisms between backbone and Monte Carlo network,
- The granularity (i.e. spatial resolution) of the grid domain, where the anchors are distributed.

### a. Backbone architecture

As mentioned earlier, we have chosen PointCNN as the backbone in the framework of Mask-MCNet because of its lower number of training parameters, compared with other deep networks for point cloud analysis. Here, by replacing the PointCNN with PointNet [164] as an alternative choice for the backbone, the performance of Mask-MCNet in both tooth instance segmentation and 3D bounding-box detection is measured. Table 5.4 shows the performance of Mask-MCNet by employing each of these two backbones. The results show that the extracted features from the PointCNN lead to slightly higher accuracy in both segmentation and detection tasks. This observation is in agreement with what we expected because the PointCNN extracts a richer set of geometrical features by incorporating KNN points in its representation at  $\chi$ -Conv layers [169].

**Table 5.4** Ablation test results on choosing the backbone architecture. Mask-level and box-level AP for two models are reported.

Backbone Arch.	Mask segmentation			Bounding-box localization		
	mIoU	mAP	mAR	mIoU	$mAP_{0.25}$	$mAP_{0.5}$
PointCNN [169]	<b>0.98</b>	<b>0.98</b>	<b>0.97</b>	0.64	1.0	0.94
PointNet [164]	0.97	0.97	0.94	0.65	1.0	0.72

### b. Data augmentation with local surface geometry

A point cloud does not explicitly convey information from neighboring points. Therefore, to aggregate over local neighborhoods, most existing methods augment the input with the surface normal vectors, or resort to a neighbor-searching mechanism [193], or the ball query [169]. Since some scanner software additionally exports the surface mesh data, computing the normal vectors per point (mesh vertices) is trivial. In this case, a normal vector per query point can be simply computed by averaging over all normal vectors of faces connecting to the point. However, this approach is prone to noise.

An alternative approach is to use an approximation method such as an analysis of the local covariance matrix, which is computed from the neighbors of the query point. To compute the local covariance matrix, we used all points within a fixed distance from the center of a sphere on the query point in the 3D space. This approach is less vulnerable to the number of points (resolution) than using KNN. The computed local covariance matrix of size  $3 \times 3$  can be either vectorized to a size of  $1 \times 9$  and directly used as an attribute per point [198], or it can be analyzed by Eigen-decomposition analysis (i.e. PCA) [199]. Table 5.5 shows the impact of using the local surface geometry on tooth instance localization and segmentation. The results show that adding neighboring point information can slightly improve the results. The PointCNN is employed as the backbone, and since it internally computes the features on KNN points, it can make such an augmentation less effective.

#### 5.4. Model 2: Instance segmentation in 3D point cloud scans

**Table 5.5** Ablation test with/without data augmentation by using local surface geometry. The normal vectors are computed based on either mesh data or PlanePCA method. Alternatively, the local covariance matrix can be vectorized and used as an attribute for each input point.

Data augmentation	Mask segmentation			Bounding-box localization		
	mIoU	mAP	mAR	mIoU	$mAP_{0.25}$	$mAP_{0.5}$
by local surface geometry						
—	0.97	0.97	0.96	0.62	1.0	0.93
Mesh surface normal	0.98	0.98	0.97	0.64	1.0	0.94
PlanePCA normal [199]	0.98	0.97	0.97	0.64	1.0	0.94
Local covariance [198]	<b>0.98</b>	<b>0.98</b>	<b>0.97</b>	<b>0.65</b>	1.0	<b>0.95</b>

#### c. Coupling mechanism

By the *coupling* mechanism, the model transfers the computed features from the 3D point cloud into a predefined 3D grid space ( $G$  domain). More specifically, the mechanism of coupling the backbone with the Monte Carlo network indicates the way that the extracted backbone features at the location of points in the input point cloud are transferred to the nodes of the grid domain  $G$ . Here, we evaluate two different coupling mechanisms for performing the aforementioned transportation.

In the first mechanism, as explained earlier, each MLP-based kernel in the first layer of the Mask-MCNet, maps the backbone feature vectors of all points within its receptive field into its center, where the node of  $G$  is located. Thus, transferring the geometrical information between these two domains is performed by a set of learning convolutional kernels.

In the second mechanism, simply a max-pooling operator can aggregate (i.e. pool) the backbone feature vectors in a predefined spherical receptive field and assigns the maximum of that feature set to a node of  $G$ , where it is located at the center of its receptive field.

For comparing the performance of these two mechanisms, we have kept the radii of their receptive fields identical. Table 5.6 shows the performance of Mask-MCNet in tooth instance segmentation and detection with these two coupling mechanisms. The results show that employing the learning kernel has a slightly higher performance than max-pooling for transferring the information between these two spatial domains.

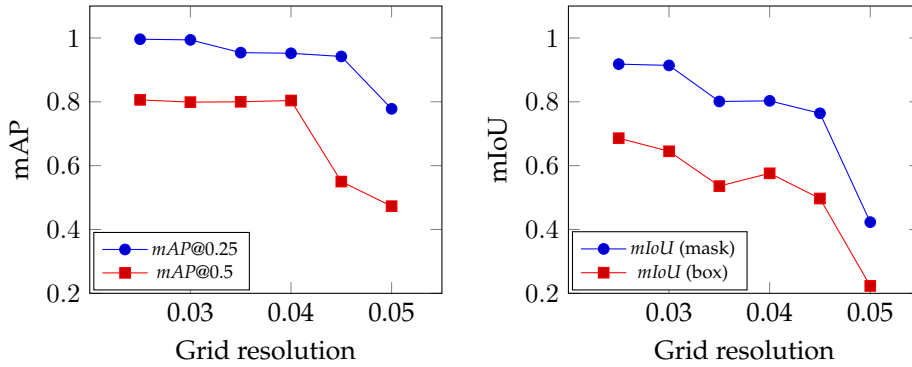
**Table 5.6** Ablation test on the type of coupling mechanism between backbone and Monte Carlo networks. The best performing mechanism is printed in bold.

Coupling mechanism	Mask segmentation			Bounding-box localization		
	mIoU	mAP	mAR	mIoU	$mAP_{0.25}$	$mAP_{0.5}$
1. MLP Conv. filter	<b>0.98</b>	<b>0.98</b>	<b>0.97</b>	0.71	1.0	0.95
2. Max-pooling	0.97	0.96	0.96	0.64	0.99	0.94

#### d. Granularity of anchor domain

As mentioned earlier, the nodes of grid domain  $G$  indicate the central position of a single ( $k = 1$ ) or multiple anchors ( $k > 1$ ). The spatial resolution of  $G$  affects the performance of tooth detection and localization. On one hand, employing a dense high-resolution grid helps for posing a sufficient number of anchors around each tooth (especially around small teeth such as incisors). This increases the chance of positioning a set of highly overlapping anchors with each tooth and consequently detecting several positive anchors for each tooth instance. On the other hand, employing a low-resolution, hence sparser grid, is more likely to miss some of the teeth, because the number of detected positive anchors may not be sufficient for forming a cluster at a later stage. However, employing a very high-resolution grid requires the processing of more anchors and hence a higher computation time.

Figure 5.14 shows the performance of the model, both for the detection and segmentation of tooth instances against changing the granularity of the  $G$  domain. The left plot in Figure 5.14 illustrates that the rate of tooth-instance detection (vertical axis) drops when decreasing the spatial resolution of  $G$  (horizontal axis). Since the 3D coordinates of points in the input point cloud are normalized to unity standard deviation, the unit of spatial resolution is reported here by its normalized value. Figure 5.14 (right) plots the mIoU scores as a measure of tooth-instance segmentation performance (i.e. mask generation). It also plots the mIoU scores between the 3D bounding boxes of the ground truth and the prediction. Because of a trade-off between the granularity of  $G$  and the computational cost of the model, in all our experiments, we adjust the grid resolution to be equal to 0.03 normalized value in the unity-deviation scaled data.



**Figure 5.14** Impact of granularity of grid domain ( $G$ ) in Mask-MCNet architecture on the tooth-instance detection and segmentation tasks. The  $y$ -axis denotes the spatial resolution of  $G$  in the normalized 3D space (where the data has a mean of zero and a standard deviation equal to unity). (left) 3D bounding-box detection results at 3D IoU thresholds equal to 0.25 (mAP@0.25) and 0.5 (mAP@0.5). (right) Tooth-instance segmentation (in blue) and localization (in red) performances.

### 5.4.5 Discussion

Accurate tooth-instance segmentation is an important step towards automated computational dentistry with many clinical applications in implantology and orthodontics. In this study, we have presented Mask-MCNet, a new end-to-end deep learning framework for tooth-instance segmentation in a 3D point cloud of an intra-oral scan.

*A. Processing data in its original resolution:* In contrast to existing deep learning models, the Mask-MCNet does not employ a voxelization or down-sampling step for processing the input 3D point cloud. Instead, by first localizing the 3D bounding box of teeth in extracted patches and then segmenting the points that belong to each tooth instance, the model can process a large point cloud in patches, where each patch covers about 2-3 teeth. Hence, the large point cloud is processed at its native high resolution, thereby preserving the finely detailed geometrical information, which is crucial for accurate teeth segmentation.

*B. Efficient feature transportation in the 3D space:* In the proposed architecture of Mask-MCNet, the Monte Carlo ConvNet transfers the information from the point cloud where it is spread over the surface of objects into the entire 3D space (e.g. the void space inside of objects). This property facilitates the inference of the center and size of objects (i.e. teeth). Furthermore, the employed Monte Carlo ConvNet in the Mask-MCNet can handle the processing of non-uniformly distributed samples. This feature leads to an efficient search of object proposals which is important for the scalability of the method, such that it is applicable for processing intra-oral scans with large point clouds (more than 180k points).

*C. Highly accurate and fast segmentation:* Our experiments have shown that the proposed model achieves a 98% mIoU on the test data, thereby outperforming the state-of-the-art networks in tooth-instance segmentation. This level of performance is close to the human level and obtained in only a few seconds of processing time, whereas for a human it would form a lengthy and labour-intensive task.

## 5.5 Conclusions

The focus of this chapter has been on designing a deep learning model for tooth segmentation in the point cloud data of intra-oral scanners. Accurate semantic tooth segmentation in IOS data is a fundamental and challenging task in computational dentistry. It is fundamental because success in building an accurate CAD model that performs at the human level in segmentation of the scans can automate many laborious procedures for accurate treatment planning and provide useful clinical information to support better treatment. Precise tooth segmentation in IOS data can support an automated clinical workflow in the implantology and orthodontic fields. This problem is challenging because for designing an accurate deep learning model, several issues should be addressed such as working with unstructured point sets, incorporating finely detailed shape information in a large 3D point cloud with hundreds of thousands of points, abnormalities in



dentition, and missing data due to the interventional scanning of the oral cavity with the handheld intra-oral scanners.

We have presented two different approaches to the problem and analyzed their performances on real-world clinical datasets.

- The first approach defines tooth segmentation as a semantic segmentation task for a deep learning model. Although this can be considered as a standard problem in machine learning, designing a CAD system with a limited computational and memory budget that performs close to the human level requires overcoming several issues.

We have proposed two techniques namely non-uniform re-sampling and adversarial training to address the shortcomings of patch processing. Non-uniform re-sampling allows use to process the point cloud with different levels of granularity. As a consequence, the network processes the point cloud in a few iterations and each time the input point cloud is presented to the network has locally dense and globally sparse samples. This technique preserves both the fine details information and the global context that are required for an accurate inference on points semantic labels. Furthermore, the adversarial training by incorporating statistics of teeth labels regularizes the distribution of predicted labels as an inductive bias for the segmentation network.

- In the second approach, to facilitate the optimization and exclude the adversarial training, we have defined the tooth segmentation problem as object-instance localization and segmentation tasks. We have presented Mask-MCNet based on recent advances in deep learning models for point cloud analysis. The Mask-MCNet is trained on the 3D patches of the point cloud and simultaneously localizes and segments the tooth instances in the patches. As a result of this segmentation, a unique label is assigned to each tooth instance. In the post-processing stage, for converting the instance labels into FDI standard labels, we have introduced a combinatorial search for label assignment. Through our extensive experiments, we have shown that the Mask-MCNet outperforms state-of-the-art deep learning models in IOS segmentation. Furthermore, several ablation studies have shown the impact of each proposal in designing the Mask-MCNet.

To the best of our knowledge, this is the first study on designing a deep learning model for automating the tooth-segmentation task in the point cloud scans. Since the proposed model achieves highly accurate segmentation of the IOS at the human level, employing the proposed solution in clinical practice is likely to be established in the near future.

# CHAPTER 6

## Conclusions

This thesis has presented deep learning-based methods for automated analysis of four different cases in the medical imaging domain. In this final chapter, the contributions and conclusions presented in the individual chapters are summarized and answers are provided for the research questions posed in Chapter 1. More specifically, this chapter starts in Section 6.1 with wrapping up the conclusions and findings from the individual chapters. In Section 6.2, the research questions from the opening chapters are reconsidered and addressed individually. Section 6.3 provides an outlook on future developments and emerging research issues.

### 6.1 Conclusion of the individual chapters

**Chapter 2.** This chapter presents two different deep learning models for cancer detection in mass spectrometry (MS) data. MS provides the molecular profile of an examined tissue, which represents a rich and informative measure that can be used for cancer diagnosis. However, understanding and interpreting a mass spectrum image with several thousands of channels is difficult for a pathologist and therefore requires developing appropriate computational algorithms for data abstraction or automated diagnosis.

It is empirically shown that learning long-range dependencies across mass spectra improves the discriminating features for cancer detection. For learning such dependencies, dilated convolutional kernels in a 1D ConvNet were employed. Alternatively, it was also shown that recurrent neural networks (RNNs), equipped with Long-short term memory (LSTM), outperforms other competitive methods by a moderate 1.87% and 1.45% higher accuracy on two clinical datasets, but with an impressive 6 times faster training time.

**Chapter 3.** This chapter presents solutions to three important problems in automated histopathology, including (1) stain-color normalization, (2) metastases

detection and their classification, and (3) the impact of compression-aware neural network training and inference. For each of these problems, individual deep learning-based solutions are proposed, which advance the borders of this field by outperforming state-art-the-art methods. The following findings related to these problems have been collected.

Stain-color variations can be critical when the color profile of the test sample deviates too much from the training set. It was shown that the proposed generative neural models are able to reduce the color variations between samples and this led to a higher likelihood between the training and test sample distributions. Such adaptation of sample distributions is considered as *domain adaptation*, which is known to reduce the gap between the color profiles of training and test samples and consequently increases classifier performance in the inference on unseen examples. At the time of performing this research, the work has resulted into the first GAN-based approach for color normalization, which inspired several follow-up studies.

Additionally, cancer detection and segmentation as another important problem in computational pathology has been studied. This study has proposed a cancer detection and classification framework, based on recent advances in deep learning architectures and graphical modeling. We have shown that modeling dependencies between image patches of histopathological slides by using conditional random fields, combined with updating the predictions of a neural network on image patches, can increase the overall performance in metastases detection and categorization. This is mainly due to broadening the decision-making of the model over all patches and not only the individual prediction. Here, expanding the field of view of the model is performed through a fully-connected graph that is constructed and optimized over all patches in the image.

The proposed technique has proven to be effective on achieving a rather high performance in pathology metastases detection when tested on a large-scale dataset. One of the main reasons of this high-performance score is the full exploitation of visual contextual information in WSIs. This result has obtained the second place in the international CAMELYON17 competition on a large-scale dataset, including samples from several laboratories.

Lastly, we have studied the impact of a state-of-the-art image compression technique (namely JPEG 2000) on the performance of cancer detection by a neural network. The compression aspect is hardly explored for this domain and is therefore highly relevant for designing a CAD system. We have conducted three series of experiments, referring to different scenarios in training or inferring with a CAD model across various compression ratios. The empirical studies imply that training on high-quality images and testing on compressed data results in decreasing the performance in cancer classification significantly when the compression factor exceeds 24. This low tolerance of the model against compression is mainly explained by absence of distortion artifacts in the training set, so that the parameters of the model are not adapted to such effects. However, when the training does include the same amount of compression artifacts that exist in the test data, it shows a constantly high performance across various

compression ratios. Lastly, we have shown that compressing the data with a factor of 48, the model is considerably tolerant to increase the compression ratios to higher values above 48, as well as decreasing the compression ratios below 48 with a reasonable range.

**Chapter 4.** This chapter has introduced two different deep learning techniques for detecting needles in 3D US data, which achieve very high precision at a low false-negative rate. Comparison with the conventional technique shows that deep learning-based solutions significantly outperform the conventional approach with a large margin. The largest part of improvement is explained by using learning filters, instead of fixed Gabor filters. Furthermore, the high precision is achieved by exploiting dedicated convolutional networks for needle segmentation in 3D US volumes. Furthermore, novel modeling of 3D US context information is introduced, using 2.5D data of a multi-view thick-sliced fully-convolutional network.

The proposed patch classification and semantic segmentation systems are evaluated on several ex-vivo datasets and outperform the classification of the state-of-the-art handcrafted features, achieving 78 and 80%  $F_1$ -scores in the chicken breast data, respectively. Furthermore, our proposed needle segmentation method based on 2.5D US information achieves a 84%  $F_1$ -score in the porcine leg datasets that are acquired with a low-resolution phased-array transducer. These results show strong semantic modeling of the needle context in challenging situations, where the intensity of the needle is inconsistent and even partly invisible.

**Chapter 5.** This chapter presents deep learning methods for tooth segmentation in the point cloud data of intra-oral scanners. Accurate semantic tooth segmentation in IOS data is a fundamental and challenging task in computational dentistry. Precise tooth segmentation in IOS data can support an automated clinical workflow in the implantology and orthodontic fields. We have presented two different approaches to the problem and analyzed their performances on real-world clinical datasets.

In the first approach, we have proposed two techniques namely non-uniform re-sampling and adversarial training to address the shortcomings of patch processing in 3D point cloud of IOS data. Non-uniform re-sampling allows to process the point cloud with different levels of granularity. This technique preserves both the fine-detail information and the global context that are required for an accurate inference on points with semantic labels. Furthermore, the adversarial training by incorporating statistics of teeth labels regularizes the distribution of predicted labels as an inductive bias for the segmentation network.

In an alternative approach, to facilitate the optimization and exclude the complications of adversarial training, we have defined the tooth segmentation problem as object-instance localization and segmentation tasks. To this end, the Mask-MCNet is presented, based on recent advances in deep learning models for point cloud analysis. The Mask-MCNet is trained on the 3D patches of the point cloud and simultaneously localizes and segments the tooth instances in the patches.

## 6. CONCLUSIONS

---

Through our extensive experiments, it has been shown that the Mask-MCNet outperforms state-of-the-art deep learning models in IOS segmentation.

To the best of our knowledge, this is the first study on designing a deep learning model for automating the tooth-segmentation task in the point cloud scans. Since the proposed model achieves highly accurate segmentation of the IOS at the human level, employing the proposed solution in clinical practice is likely to be established in the near future.

### 6.2 Discussion on the research questions

The proposed methods and solutions are discussed with respect to the research questions, formulated in Chapter 1.

#### RQ1. Metastases detection in 1D mass spectrometry data

**RQ1a.** *What is the influence of the receptive fields of convolution kernels in ConvNets to enhance the expressiveness of the learned features and add to the overall performance of cancer detection?*

The mass spectrometry signal can be read as a histogram, indicating the intensity values over several thousands mass-to-charge ( $m/z$ ) bins. The 1D convolution kernels applied across the  $m/z$  bins to extract discriminating features, to perform a binary classification between cancer and normal samples. In contrast to natural images, the MS data is a sparse signal with several information spikes distributed over all bins.

Increasing the field of view of the ConvNet, equipped with the 1D kernels, requires either a very deep architecture, or a too large pooling size, which both have their own issues. For these reason, an alternative was discussed that exploiting dilated kernels with various dilation factors. The dilated kernels increase the field of view in each layer of the network by learning long-distance patterns in the signal while keeping the number of learning parameters unchanged. It was shown that this approach can achieve a higher performance by learning better expressive features across the  $m/z$  bins and results in higher performance in cancer detection.

**RQ1b.** *For learning the existing dependencies across the mass spectrum, which architecture of the Recurrent Neural Networks (RNNs) operates best and can the RNNs outperform the ConvNets with 1D kernels for cancer detection?*

In Section 2.3, we have performed an architecture search over multiple factors in a baseline RNN architecture. The key property of the RNN is that the extracted features from the data theoretically incorporate the entire range of the input sequence, hence it can learn the dependencies among  $m/z$  bins both locally and globally. This extended learning capability is enabled by using a 2-layer bidirectional LSTM module, where each layer contains 100 cells. It has been shown

that this obtains the best performance in terms of balanced accuracy for cancer detection in a Lung cancer dataset. As expected, the RNN is capable of learning long-range dependencies across  $m/z$  bins, so that it outperforms the ConvNet architecture on both Lung and Bladder datasets by achieving about 0.86 and 0.84  $F_1$  scores, respectively.

## **RQ2. Deep learning-based 2D histopathological image analysis**

**RQ2a.** *Is it possible to model the process of stain-colorization in pathology with a generative neural model? And how can we make this framework a generic solution for various types of histopathological examinations?*

The stain-color normalization task can be translated into a colorization procedure where the gray-scale images are colorized with a generative model to produce realistically colored images. In this process, the model is capable of learning color profiles existing in the training data, which then can be guiding the conversion of coloring a source image to resemble a target image. This process can indeed be formulated by a generative neural network, and results in the proposed unsupervised GAN-based model in Section 3.2.2.

The proposed stain-color normalization method implicitly learns to extract different image structures that have the same chromatic characteristics. The proposed method lacks any threshold, ground truth or any other assumptions about the shape and color of structures appearing in histopathology images. Taking fewer assumptions about the image contents enables the method to become more general and applicable to different histopathology images, corresponding with distinct tissue types.

**RQ2b.** *How can we create a hybrid algorithm that combines a probabilistic model and a neural network to ensure a high level of fidelity in the color conversion process?*

In Section 3.2.4, we have introduced the Neural-GMM, a model that combines a Gaussian mixture model (GMM) and ConvNets into one hybrid learning framework. This hybrid framework enables an unsupervised learning approach to histopathological image colorization. Neural-GMM simultaneously optimizes the parameters of the GMM and ConvNets to perform two tasks. First, it segments the histopathological images, and second, it fits Gaussian distributions to the pixel colors of each segment. As a result, unlike GAN-based methods, this model has the ability to modify pixel colors while preserving the original image contents in term of structure. This ensures a high level of fidelity in the color conversion process, maintaining the integrity of tissue structures in the histopathological images. The proposed Neural-GMM method outperforms existing GAN-based methods in terms of color consistency in the stain-color normalization across several pathology laboratories.

**RQ2c.** *How can we effectively combine and integrate data across features that have been learned through individual patch-level analysis? Is it feasible to achieve this fusion by employing a probabilistic graphical model?*

## 6. CONCLUSIONS

---

Patch-level analysis enables the training of neural networks on histopathological slides of large sizes (i.e. WSIs). However, the decision-making capability of the network is constrained by the content contained within each individual patch. In Section 3.3.2, we have proposed to integrate the features and predictions at the patch level within a graphical framework. In this approach, a fully-connected graph is constructed, where each node represents the neural features extracted from each individual patch of the whole slide images (WSIs). This graph allows to combine feature descriptions at the patch level to the whole slide level. We have demonstrated that by applying conditional random fields (CRFs) on this graph to perform data fusion across patches, there is a notable enhancement in the overall performance. The proposed data fusion technique achieved a high kappa score equal to 0.876 in patient-level metastasis classification on the breast lymph data and won the 2nd place in the Camelyon international challenge from the IEEE ISBI Conference 2018.

**RQ2d.** *How does the utilization of a standardized data compression technique impact the training and inference performance of a deep learning model? What compression ratio is recommended to achieve optimal performance for histopathology image analysis?*

In Section 3.4, we have introduced different scenarios, involving training and inference on both low and highly compressed images. Our objective was to assess the impact of image degradation on the classification performance of a neural network. The results indicate that when the network is trained on high-quality uncompressed histopathology images, it maintains its performance on compressed images up to a compression factor of 24, as measured by the  $F_1$  score and the AUC. However, for compression factor of 32 or higher, there is a noticeable decline in the measured  $F_1$  scores. Therefore, our recommendation is using a compression factor less than 32 when a JPEG2000 technique is employed, which is an acknowledged standard for professional image processing applications.

Additionally, we have observed that training the network on compressed images leads to significant improvement, achieving an  $F_1$  score of 0.93 with a compression factor of 164, compared to  $F_1=0.58$  when the model is trained on uncompressed images. Training the network on the compressed images is recommended because the network can learn compression artifacts, which are included in the data.

### **RQ3. Deep learning-based needle localization in 3D US scans**

**RQ3a.** *Is it possible to use 2D convolution kernels as a proxy for 3D kernels in ConvNets for processing 3D US volumes when there is only a small dataset available?*

In Section 4.2.2, we have proposed the first deep learning-based method for efficient processing of volumetric ultrasound (US) data, by introducing 2D orthogonal convolutional kernels as a substitute for 3D kernels in Convolutional Neural Networks (ConvNets). The results demonstrate that the proposed ConvNet architecture is capable of detecting and localizing needles within US data, by processing orthogonal multiple views of US patches. This approach has

shown a 25% improvement in  $F_1$  score for needle detection, compared to the conventional 2D Gabor-based feature classification method.

Whereas the Gabor-based features are fixed and can only jointly represent a needle, the proposed deep learning network is capable of directly learning the filter parameters and consequently represents the discriminating features for needle detection. This explains why there is a substantial growth in the obtained performance of up to 25% in  $F_1$  score. In comparison with 3D kernels, the proposed 2D orthogonal approach involves a much lower number of learning parameters, so that the network is less prone to over-fitting issues. Additionally, the 3D convolution operator is not suited for edge-device implementation.

**RQ3b.** *Can the segmentation of needles in a 3D ultrasound (US) volume be formulated as an image-slice segmentation task? Does this approach yield better results compared to the patch classification technique, in terms of relevant evaluation metrics?*

In Section 4.2.3, we have introduced a novel approach for needle segmentation in 3D ultrasound (US) volumes, as an alternative to US patch classification. The proposed method involves decomposing the 3D volume into 2D cross-sections to label the needle parts and subsequently reconstructing the 3D needle labels from multiple views. This approach offers significant benefits, including a reduced number of parameters in the convolution kernels compared to full 3D kernels. Consequently, the network requires fewer training samples and operates at a higher speed.

The semantic segmentation technique utilizing 2.5D US information achieves an impressive  $F_1$  score of 84% when applied to porcine leg datasets, captured with a lower-resolution phased-array transducer. These findings demonstrate the robustness of the semantic modeling approach in capturing needle context, even in challenging scenarios where the needle intensity is inconsistent or even partially invisible. In comparison with patch-level analysis, the semantic segmentation approach processes a larger content including several slices for segmenting the needle in a fully convolutional architecture (FCN). Moreover, it combines multiple levels of resolutions in its FCN architecture which is not exploited in patch-level analysis.

#### **RQ4. Deep learning-based tooth instance segmentation in 3D point cloud**

**RQ4a.** *Can we create a non-uniform re-sampling method that effectively preserves both the local intricate details and overall structure of dentition in intraoral scanner (IOS) point clouds? Additionally, how can we utilize an adversarial neural framework to impose the prior distribution of teeth positions on the dental arch?*

Section 5.3 has introduced a novel approach for analyzing point clouds at their original spatial resolution, enabling predictions for all points. This approach incorporates a non-uniform re-sampling technique along with a corresponding loss weighting mechanism, leveraging both foveation and Monte Carlo sampling. By employing this re-sampling technique, we effectively capture both local, fine-detail information and the sparse global structure of the data, which is crucial for



## 6. CONCLUSIONS

---

accurate point-level predictions. The proposed re-sampling technique allows to include both local fine details and global structures of IOS data in the sampled input point cloud data that is processed by input stages of the neural network. Hence, the network is capable of processing multiple granularity levels of the input point cloud geometry.

Furthermore, the conducted research demonstrates that incorporating the distribution of labels through adversarial training enhances the semantic segmentation of teeth. This improvement is attributed to the strong correlation between the semantic label of each tooth and its relative position in relation to other teeth within the dental arch.

**RQ4b.** *Can we devise a novel tooth instance segmentation method that addresses the degradation issues, commonly associated with conventional discretization methods such as occupancy volume representation?*

In Section 5.4, we have introduced a novel deep learning framework called Mask-MCNet for tooth instance segmentation in 3D point clouds, obtained from intra-oral scans. Unlike existing models, the Mask-MCNet does not rely on voxelization or down-sampling to handle the large point cloud data. Instead, it operates on the native high resolution of the IoS scan, preserving the intricate geometric details that are crucial for accurate teeth segmentation.

The experimental results demonstrate that the proposed model achieves an impressive 98% mIoU (mean Intersection over Union) on the test data, thereby surpassing the performance of state-of-the-art networks in tooth instance segmentation. This level of accuracy is comparable to human-level performance, and the model achieves it within a few seconds of processing time, which significantly reduces the time and labor required, compared to manual human segmentation efforts.

### 6.3 Utilization and outlook

The rapid growth and advancements of artificial intelligence are accelerating, which enables the automation of more sophisticated and complicated tasks in nearly every domain. Medical image analysis can substantially benefit from these advancements. With the current trend of progress of AI in computer vision, pattern recognition, and neural reasoning, we are not far away from the point that many medical procedures and even surgeries performed by automated machines, have a higher accuracy than human-level accuracy. The new generation of CADe and CADx systems can perform various diagnosis and prognosis tasks. In the following, an outlook is presented on AI enabling the four studied medical image systems.

#### A. Mass spectrometry analysis:

The integration of AI with mass spectrometry could involve the development of predictive models that can accurately classify different types of cancer based on

their unique mass spectrometry profiles, even in the early stage of disease. These models could learn from large-scale datasets, including data from various cancer types and subtypes, to improve accuracy and generalization. Another revolutionary advancement is the integration of AI with real-time mass spectrometry systems, enabling rapid and automated analysis of cancer samples during surgical procedures. This can provide instant feedback to surgeons or surgical robots, aiding in real-time decision-making and enhancing the precision of tumor resection.

Chapter 2 has introduced proposed neural architectures that achieve state-of-the-art (SOTA) performance on MS data, by considering either dilation factors in ConvNet or selecting different components of an RNN. While these architectural advancements are realized through manual processes, Neural Architecture Search (NAS) offers a promising avenue for further enhancing MS analysis in the future. NAS is a field of deep learning research that aims to automate the design of neural network architectures. Instead of relying on manual design, NAS employs search algorithms, often guided by reinforcement learning or evolutionary algorithms, to automatically discover optimal or high-performing architectures for specific tasks. By exploring a broad search space, NAS algorithms have the potential to unveil novel architectures that surpass manually designed architectures. This exploration allows for the discovery of complex patterns and interactions within the data, ultimately leading to improved model performance. Furthermore, NAS enables customization of architectures based on specific requirements and constraints. By incorporating task-specific objectives, such as latency constraints or model size limitations, NAS algorithms can uncover architectures that are tailor-made for specific deployment scenarios.

## **B. Computational histopathology analysis:**

Recent progress in the field of computational pathology has facilitated the identification and quantification of various features by pathologists, including abnormal cell morphology, tissue structures, and specific biomarkers. AI technology has emerged as a valuable tool in automating time-consuming tasks, enhancing accuracy, and expediting diagnoses. In particular, deep learning models hold immense potential for pathologists since they augment the diagnostic capabilities. By analyzing extensive datasets of histopathology images, these algorithms can learn intricate patterns and support pathologists in making more precise diagnoses. By combining the expertise of pathologists with AI-based data analysis, the accuracy and efficiency of diagnoses can be significantly improved.

In addition to automating conventional pathology practices, deep learning algorithms enable the analysis of complex multimodal data, which poses challenges for conventional methods and even pathologists. By leveraging patient data such as histopathology images, clinical records, and genetic information, AI algorithms can develop predictive models for various diseases. These models can aid in early detection, risk assessment, and prognosis prediction, thereby facilitating timely interventions and personalized treatment plans.

## 6. CONCLUSIONS

---

The future of efficient WSI compression lies in the development of neural compression techniques. Neural compression is a data compression approach that harnesses neural networks to achieve effective compression of different types of data, including images, videos, and audio. Unlike conventional methods such as JPEG2000 that rely on conventional compression algorithms, neural compression utilizes deep learning models to understand the underlying patterns and redundancies within the data, resulting in a more compact encoding. This approach enables higher compression ratios while preserving superior visual quality for WSIs.

It is important to note that neural compression methods are still an active area of research and development. Although they demonstrate promising outcomes, there are challenges associated with computational complexity, training requirements, and deployment considerations that need to be addressed. Nevertheless, the advancements in neural compression hold tremendous potential for achieving enhanced compression efficiency and improved visual quality across various data compression applications.

### C. Ultrasound 3D analysis:

Image guidance plays a vital role in minimally-invasive interventions, utilizing ultrasound (US) imaging as a key imaging modality for live (low-latency) monitoring guidance. The research outlined in Chapter 4 demonstrates the effectiveness of supervised deep learning models in accurately localizing needles within 3D ultrasound scans, even in cases where the needle is partially inserted only. While this study has focused on small-scale US scans, it is expected that further improvements can be achieved with access to a larger dataset. Additionally, considering the labor-intensive nature of voxel-level manual annotation for 3D data, exploring alternative learning frameworks such as self-supervised learning methods or Multiple Instance Learning (MIL) techniques, which rely on scan-level labels only, could be an interesting avenue for future research.

Another promising line of research is temporal analysis that can play a significant role in improving instrument detection in ultrasound volumetric data. Incorporating the temporal analysis for instrument detection enhancement can be performed in several ways:

(1) *Motion tracking*: By processing temporal changes in the position and movement of instruments over US video frames, it becomes easier to differentiate instruments from the surrounding anatomical structures. Temporal analysis can facilitate the tracking of instruments, even in the presence of noise or artifacts, by establishing consistent motion patterns over time.

(2) *Doppler analysis*: Temporal analysis allows for the examination of blood flow patterns and velocities over time, using Doppler ultrasound. This information can aid in distinguishing blood vessels from instruments. By analyzing the changes in Doppler signals over successive frames, it is possible to identify instruments that do not exhibit the expected blood flow patterns.

(3) *Temporal filtering*: Applying temporal filters to ultrasound volumetric data can help to suppress noise and enhance the visibility of instruments. By taking advantage of the temporal coherence of the instrument signal and the temporal differences between instruments and surrounding tissues, temporal filtering can improve the detection accuracy.

(4) *Temporal averaging*: By averaging multiple frames of ultrasound data, the signal-to-noise ratio can be enhanced, making it easier to identify instruments in spatial domain. Temporal averaging can help in reducing the impact of speckle noise and improve the visualization of subtle instrument details.

In conclusion, temporal analysis provides valuable insights into the dynamic behavior of instruments and the consistency of their appearances within ultrasound volumetric data. By exploiting the temporal characteristics and patterns, instrument detection algorithms can achieve higher accuracy, robustness, and visualization quality that can be considered in future research.

#### **D. Computational dentistry:**

Emerging advancements in imaging technology, specifically handheld intra-oral scanners (IOS), offer real-time 3D scanning of the oral cavity. The accurate segmentation of teeth from IOS data can greatly assist in automating clinical workflows in implantology and orthodontics, where precise 3D geometry profiles of teeth and gums are crucial for successful outcomes.

In fields that demand precision and accuracy within a confined space, robot-assisted technology can provide dental surgeons with unparalleled precision, resulting in optimal functional and aesthetic results for patients. The dental industry has increasingly embraced robotic technology, leading to modernization in dentistry. Robot-assisted surgery is poised to surpass freehand dental implant procedures, benefiting both practitioners and patients alike. A fundamental aspect of implementing robot-assisted technology is 3D perception, which involves tooth instance segmentation.

A significant challenge in training deep learning-based segmentation models for IOS data lies in the limited availability of 3D annotated data. Currently, manual annotation tools are used to expedite the annotation process in 3D. However, in the near future, other annotation techniques that incorporate various input prompts [200] may support the creation of large-scale annotated 3D data, directly enhancing the segmentation results.

#### **General outlook:**

The future of AI-enabled medical imaging holds great promise and potential for revolutionizing healthcare. AI algorithms will continue to improve in accuracy and astonishing speed, enabling more precise and efficient diagnosis across various imaging modalities. AI systems will assist radiologists by quickly analyzing medical images, highlighting abnormalities, and providing quantitative measurements, leading to faster and more accurate diagnostics. AI algorithms will

## 6. CONCLUSIONS

---

continue to advance in their ability to analyze complex medical images. They will aid in the detection and characterization of subtle abnormalities that might be challenging for human observers. AI systems will also integrate multi-modal data and provide comprehensive insights to improve diagnostic accuracy and confidence.

AI algorithms will be integrated into imaging equipment such as mass spectrometry and ultrasound imaging, providing real-time decision support during image acquisition. These systems will offer guidance to technicians and radiologists, ensuring optimal image quality, reducing errors, and minimizing the need for repeat scans. Such integration will assist in the early detection of diseases by identifying subtle patterns and biomarkers in medical images that may indicate the presence of disease even before symptoms become manifest. This early detection will enable timely interventions, improving patient outcomes and tremendously reducing healthcare costs because interventions become more simplified and minimally invasive.

AI systems will further optimize workflow and increase efficiency in medical imaging departments. By automating tasks, streamlining image analysis, and prioritizing urgent cases, AI will help reduce waiting times, improve resource allocation, and enhance overall productivity. While this vision represents exciting possibilities, it is important to note that the successful integration of AI in medical imaging will require careful validation, intensive collaboration between clinicians and technologists, and continuous monitoring to ensure patient safety, privacy and the ethical use of technology.

# Acronyms

<b>1D</b>	One-dimensional
<b>2D</b>	Two-dimensional
<b>3D</b>	Three-dimensional
<b>AUC</b>	Area Under the Curve
<b>CAD</b>	Computer-Aided Diagnosis
<b>ConvNets</b>	Convolutional Neural Networks
<b>CSP</b>	Constraint Satisfaction Problem
<b>CVAE</b>	Conditional Variational Auto-Encoder
<b>FFPE</b>	Formalin-Fixed Paraffin-Embedded
<b>GPU</b>	Graphics Processing Unit
<b>H&amp;E</b>	Hematoxylin and Eosin
<b>IOS</b>	Intra-Oral Scans
<b>IR</b>	Interquartile Range
<b>m/z</b>	Mass-to-charge ratio
<b>MALDI</b>	Matrix-Assisted Laser Desorption/Ionization
<b>MLP</b>	Multilayer Perceptron
<b>MRI</b>	Magnetic Resonance Imaging
<b>MSI</b>	Mass Spectrometry Imaging
<b>NAS</b>	Neural Architecture Search
<b>NLP</b>	Natural Language Processing
<b>PDF</b>	Probability Density Function

## 6. CONCLUSIONS

---

**RNNs** Recurrent Neural Networks

**ROI** Region-of-Interest

**RPN** Region Proposal Network

**SGD** Stochastic Gradient Descent

**ToF** Time-of-Flight

**UHD** Ultra-High-Definition

**US** Ultrasound

## Bibliography

- [1] J. Zhang, J. Rector, J. Q. Lin, J. H. Young, M. Sans, N. Katta, N. Giese, W. Yu *et al.*, “Nondestructive tissue analysis for ex vivo and in vivo cancer diagnosis using a handheld mass spectrometry system,” *Science translational medicine*, vol. 9, no. 406, p. eaan3968, 2017.
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [3] R. Longuespée, R. Casadonte, M. Kriegsmann, C. Pottier, G. Picard de Muller, P. Delvenne, J. Kriegsmann, and E. De Pauw, “Maldi mass spectrometry imaging: A cutting-edge tool for fundamental and clinical histopathology,” *PROTEOMICS—Clinical Applications*, vol. 10, no. 7, pp. 701–719, 2016.
- [4] O. J. Gustafsson, J. S. Eddes, S. Meding, S. R. McColl, M. K. Oehler, and P. Hoffmann, “Matrix-assisted laser desorption/ionization imaging protocol for in situ characterization of tryptic peptide identity and distribution in formalin-fixed tissue,” *Rapid Communications in Mass Spectrometry*, vol. 27, no. 6, pp. 655–670, 2013.
- [5] J. Kriegsmann, M. Kriegsmann, and R. Casadonte, “Maldi tof imaging mass spectrometry in clinical pathology: a valuable tool for cancer diagnostics,” *International journal of oncology*, vol. 46, no. 3, pp. 893–906, 2015.
- [6] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, “Histopathological image analysis: A review,” *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.
- [7] S. Banerjee, “Empowering clinical diagnostics with mass spectrometry,” *ACS omega*, vol. 5, no. 5, pp. 2041–2048, 2020.
- [8] T. Alexandrov, “Maldi imaging mass spectrometry: statistical data analysis and current computational challenges,” *BMC bioinformatics*, vol. 13, no. 16, pp. 1–13, 2012.
- [9] C. Yang, Z. He, and W. Yu, “Comparison of public peak detection algorithms for maldi mass spectrometry data analysis,” *BMC bioinformatics*, vol. 10, no. 1, pp. 1–13, 2009.
- [10] T. Boskamp, D. Lachmund, J. Oetjen, Y. C. Hernandez, D. Trede, P. Maass, R. Casadonte, J. Kriegsmann *et al.*, “A new classification method for maldi imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples,” *Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics*, vol. 1865, no. 7, pp. 916–926, 2017.



## BIBLIOGRAPHY

---

- [11] P. W. Siy, R. A. Moffitt, R. M. Parry, Y. Chen, Y. Liu, M. C. Sullards, A. H. Merrill, and M. D. Wang, "Matrix factorization techniques for analysis of imaging mass spectrometry data," in *2008 8th IEEE International Conference on BioInformatics and BioEngineering*. IEEE, 2008, pp. 1–6.
- [12] L. A. Klerk, A. Broersen, I. W. Fletcher, R. van Liere, and R. M. Heeren, "Extended data analysis strategies for high resolution imaging ms: New methods to deal with extremely large image hyperspectral datasets," *International Journal of Mass Spectrometry*, vol. 260, no. 2-3, pp. 222–236, 2007.
- [13] E. A. Jones, A. van Remoortere, R. J. van Zeijl, P. C. Hogendoorn, J. V. Bovée, A. M. Deelder, and L. A. McDonnell, "Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma," *PloS one*, vol. 6, no. 9, p. e24913, 2011.
- [14] W. M. Abdelmoula, B. Balluff, S. Englert, J. Dijkstra, M. J. Reinders, A. Walch, L. A. McDonnell, and B. P. Lelieveldt, "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 43, pp. 12 244–12 249, 2016.
- [15] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. Heeren, and F. A. Hamprecht, "Concise representation of mass spectrometry images by probabilistic latent semantic analysis," *Analytical chemistry*, vol. 80, no. 24, pp. 9649–9658, 2008.
- [16] S.-O. Deininger, M. Becker, and D. Suckau, "Tutorial: multivariate statistical treatment of imaging data for clinical biomarker discovery," in *Mass Spectrometry Imaging*. Springer, 2010, pp. 385–403.
- [17] D. Bonnel, R. Longuespee, J. Franck, M. Roudbaraki, P. Gosset, R. Day, M. Salzert, and I. Fournier, "Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in maldi-msi: application to prostate cancer," *Analytical and bioanalytical chemistry*, vol. 401, no. 1, pp. 149–165, 2011.
- [18] M. Hanselmann, U. Kothe, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. Heeren, and F. A. Hamprecht, "Toward digital staining using imaging mass spectrometry and random forests," *Journal of proteome research*, vol. 8, no. 7, pp. 3558–3567, 2009.
- [19] S. A. Thomas, A. M. Race, R. T. Steven, I. S. Gilmore, and J. Bunch, "Dimensionality reduction of mass spectrometry imaging data using autoencoders," in *2016 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2016, pp. 1–7.
- [20] P. Inglese, J. S. McKenzie, A. Mroz, J. Kinross, K. Veselkov, E. Holmes, Z. Takats, J. K. Nicholson *et al.*, "Deep learning and 3d-desi imaging reveal the hidden metabolic heterogeneity of cancer," *Chemical science*, vol. 8, no. 5, pp. 3500–3511, 2017.
- [21] J. Zhang, J. Liu, Y. Luo, Q. Fu, J. Bi, S. Qiu, Y. Cao, and X. Ding, "Chemical substance classification using long short-term memory recurrent neural network," in *2017 IEEE 17th International Conference on Communication Technology (ICCT)*. IEEE, 2017, pp. 1994–1997.

- [22] J. Behrmann, C. Etmann, T. Boskamp, R. Casadonte, J. Kriegsmann, and P. Maaß, "Deep learning for tumor classification in imaging mass spectrometry," *Bioinformatics*, vol. 34, no. 7, pp. 1215–1223, 2018.
- [23] L. Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial Intelligence and Statistics*. PMLR, 2009, pp. 384–391.
- [24] N. H. Tran, X. Zhang, L. Xin, B. Shan, and M. Li, "De novo peptide sequencing by deep learning," *Proceedings of the National Academy of Sciences*, vol. 114, no. 31, pp. 8247–8252, 2017.
- [25] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [26] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.
- [27] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.
- [28] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical systems and signal processing*, vol. 151, p. 107398, 2021.
- [29] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.
- [30] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in psychology*. Elsevier, 1997, vol. 121, pp. 471–495.
- [31] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.
- [32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [33] F. A. Gers and E. Schmidhuber, "Lstm recurrent networks learn simple context-free and context-sensitive languages," *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1333–1340, 2001.
- [34] R. K. Brouwer, "A method for training recurrent neural networks for classification by building basins of attraction," *Neural networks*, vol. 8, no. 4, pp. 597–603, 1995.
- [35] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.
- [36] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
- [37] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 1442–1450.

## BIBLIOGRAPHY

---

- [38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [39] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and accurate entity recognition with iterated dilated convolutions," *arXiv preprint arXiv:1702.02098*, 2017.
- [40] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 525–542.
- [41] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026*, 2013.
- [42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
- [43] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [45] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *20th international conference on pattern recognition*. IEEE, 2010, pp. 3121–3124.
- [46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE Int. Conf. on computer vision*, 2015, pp. 1026–1034.
- [48] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.
- [49] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.
- [50] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagation for classification," *International Journal of Computer Theory and Engineering*, vol. 3, no. 1, pp. 1793–8201, 2011.
- [51] N. M. Nawli, W. H. Atomi, and M. Z. Rehman, "The effect of data pre-processing on optimized training of artificial neural networks," *Procedia Technology*, vol. 11, pp. 32–39, 2013.
- [52] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2015.

- 
- [53] A. Sethi, L. Sha, A. R. Vahadane, R. J. Deaton, N. Kumar, V. Macias, and P. H. Gann, "Empirical comparison of color normalization methods for epithelial-stromal classification in h and e images," *Journal of pathology informatics*, vol. 7, 2016.
  - [54] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. De Souza, A. Baidoshvili, G. Litjens, B. Van Ginneken, I. Nagtegaal *et al.*, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 160–163.
  - [55] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.
  - [56] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke, "Colour normalisation in digital histopathology images," in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, vol. 100. Citeseer, 2009, pp. 100–111.
  - [57] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE international symposium on biomedical imaging: from nano to macro*. IEEE, 2009, pp. 1107–1110.
  - [58] M. Niethammer, D. Borland, J. Marron, J. Woosley, and N. E. Thomas, "Appearance normalization of histology slides," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2010, pp. 58–66.
  - [59] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, A. N. Young, and M. D. Wang, "Automatic batch-invariant color segmentation of histological cancer images," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011, pp. 657–660.
  - [60] A. Basavanthally and A. Madabhushi, "Em-based segmentation-driven color standardization of digitized histopathology," in *Medical Imaging 2013: Digital Pathology*, vol. 8676. SPIE, 2013, pp. 152–163.
  - [61] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.
  - [62] P. A. Bautista, N. Hashimoto, and Y. Yagi, "Color standardization in whole slide imaging using a color calibration slide," *Journal of pathology informatics*, vol. 5, 2014.
  - [63] X. Li and K. N. Plataniotis, "A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 7, pp. 1862–1873, 2015.
  - [64] —, "Circular mixture modeling of color distribution for blind stain separation in pathology images," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 150–161, 2015.
  - [65] L. Sha, D. Schonfeld, and A. Sethi, "Color normalization of histology slides using graph regularized sparse nmf," in *Medical Imaging 2017: Digital Pathology*, vol. 10140. International Society for Optics and Photonics, 2017, p. 1014010.

## BIBLIOGRAPHY

---

- [66] N. Alsubaie, N. Trahearn, S. E. A. Raza, D. Snead, and N. M. Rajpoot, "Stain deconvolution using statistical analysis of multi-resolution stain colour representation," *PloS one*, vol. 12, no. 1, p. e0169875, 2017.
- [67] Y.-Y. Wang, S.-C. Chang, L.-W. Wu, S.-T. Tsai, and Y.-N. Sun, "A color-based approach for automated segmentation in tumor tissue classification," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE*, 2007, pp. 6576–6579.
- [68] A. C. Ruifrok, D. A. Johnston *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.
- [69] A. BenTaieb and G. Hamarneh, "Adversarial stain transfer for histopathology image analysis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 792–802, 2017.
- [70] H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using gan for histopathological images," *arXiv preprint arXiv:1710.08543*, 2017.
- [71] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [72] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.
- [73] G. Bueno, O. Déniz, J. Salido, M. Milagro Fernández, N. Váñez, and M. García-Rojo, "Colour model analysis for histopathology image processing," in *Color medical image analysis*. Springer, 2013, pp. 165–180.
- [74] L. Du, H. Lang, Y.-L. Tian, C. C. Tan, J. Wu, and H. Ling, "Covert video classification by codebook growing pattern," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 11–18.
- [75] G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 959–973, 2003.
- [76] R. Moshavegh, B. E. Bejnordi, A. Mehnert, K. Sujathan, P. Malm, and E. Bengtsson, "Automated segmentation of free-lying cell nuclei in pap smears for malignancy-associated change analysis," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE*, 2012, pp. 5372–5375.
- [77] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
- [78] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumar, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv preprint arXiv:1611.02648*, 2016.
- [79] A. Van den Oord and B. Schrauwen, "Factoring variations in natural images with deep gaussian mixture models," *Advances in neural information processing systems*, vol. 27, 2014.

- [80] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International conference on learning representations*, 2018.
- [81] J. A. Van der Laak, M. M. Pahlplatz, A. G. Hanselaar, and P. C. de Wilde, "Hue-saturation-density (hsd) model for stain recognition in digital images from transmitted light microscopy," *Cytometry: The Journal of the International Society for Analytical Cytology*, vol. 39, no. 4, pp. 275–284, 2000.
- [82] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.
- [83] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter *et al.*, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.
- [84] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. van der Laak, and P. H. de With, "Histopathology stain-color normalization using deep generative models," in *Medical Imaging with Deep Learning (MIDL)*, 2018.
- [85] O. Geessink, P. Bándi, G. Litjens, and J. van der Laak. (2017) Camelyon17: Grand challenge on cancer metastasis detection and classification in lymph nodes. [Online]. Available: <https://camelyon17.grand-challenge.org>
- [86] L. Geert, T. Kooi, B. E. Bejnordi, A. Arindra, A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak *et al.*, "A survey on deep learning in medical image analysis," ser. arXiv:1702.05747v2, 2017.
- [87] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.
- [88] G. Litjens, C. I. Sanchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.
- [89] B. Kong, X. Wang, Z. Li, Q. Song, and S. Zhang, "Cancer metastasis detection via spatially structured deep network," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 236–248.
- [90] P. Krahenbuhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, 2011.
- [91] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [92] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [93] P. Baqué, T. Bagautdinov, F. Fleuret, and P. Fua, "Principled parallel mean-field inference for discrete random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5848–5857.

## BIBLIOGRAPHY

---

- [94] P. Bandi, O. Geessink, Q. Manson, M. van Dijk, M. Balkenhol, M. Hermesen, B. E. Bejnordi, B. Lee *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE Transactions on Medical Imaging*, 2018.
- [95] B. Ehteshami Bejnordi and J. van der Laak. (2016) Camelyon16: Grand challenge on cancer metastasis detection in lymph nodes. [Online]. Available: <https://camelyon16.grand-challenge.org>
- [96] F. G. Zanjani, S. Zinger *et al.*, "Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces," in *Medical Imaging 2018: Digital Pathology*, vol. 10581. International Society for Optics and Photonics, 2018.
- [97] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermesen *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.
- [98] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" *arXiv preprint arXiv:1511.06348*, 2015.
- [99] S. A. El-Seoud, H. F. El-Sofany, M. Abdelfattah, and R. Mohamed, "Big data and cloud computing: Trends and challenges." *International Journal of Interactive Mobile Technologies*, vol. 11, no. 2, 2017.
- [100] T. Kalinski, R. Zwönitzer, F. Grabellus, S.-Y. Sheu, S. Sel, H. Hofmann, and A. Roessner, "Lossless compression of jpeg2000 whole slide images is not required for diagnostic virtual microscopy," *American journal of clinical pathology*, vol. 136, no. 6, pp. 889–895, 2011.
- [101] M. D. Herrmann, D. A. Clunie, A. Fedorov, S. W. Doyle, S. Pieper, V. Klepeis, L. P. Le, G. L. Mutter *et al.*, "Implementing the dicom standard for digital pathology," *Journal of pathology informatics*, vol. 9, 2018.
- [102] L. Platiša, L. Van Brantegem, A. Kumcu, R. Ducatelle, and W. Philips, "Influence of study design on digital pathology image quality evaluation: the need to define a clinical task," *Journal of Medical Imaging*, vol. 4, no. 2, p. 021108, 2017.
- [103] E. A. Krupinski, J. P. Johnson, S. Jaw, A. R. Graham, and R. S. Weinstein, "Compressing pathology whole-slide images using a human and model observer evaluation," *Journal of pathology informatics*, vol. 3, 2012.
- [104] L. Pantanowitz, C. Liu, Y. Huang, H. Guo, and G. K. Rohde, "Impact of altering various image parameters on human epidermal growth factor receptor 2 image analysis data quality," *Journal of Pathology Informatics*, vol. 8, 2017.
- [105] A. Marcelo, P. Fontelo, M. Farolan, and H. Cualing, "Effect of image compression on telepathology: a randomized clinical trial," *Archives of pathology & laboratory medicine*, vol. 124, no. 11, pp. 1653–1656, 2000.
- [106] J. P. Johnson, E. A. Krupinski, M. Yan, H. Roehrig, A. R. Graham, and R. S. Weinstein, "Using a visual discrimination model for the detection of compression artifacts in virtual pathology images," *IEEE transactions on medical imaging*, vol. 30, no. 2, pp. 306–314, 2010.

- 
- [107] C. López, M. Lejeune, P. Escrivà, R. Bosch, M. T. Salvadó, L. E. Pons, J. Baucells, X. Cugat *et al.*, "Effects of image compression on automatic count of immunohistochemically stained nuclei in digital images," *Journal of the American Medical Informatics Association*, vol. 15, no. 6, pp. 794–798, 2008.
- [108] A. Sharma, P. Bautista, and Y. Yagi, "Balancing image quality and compression factor for special stains whole slide images," *Analytical Cellular Pathology*, vol. 35, no. 2, pp. 101–106, 2012.
- [109] A. Brüggmann, M. Eld, G. Lelkaitis, S. Nielsen, M. Grunkin, J. D. Hansen, N. T. Foged, and M. Vyberg, "Digital image analysis of membrane connectivity is a robust measure of her2 immunostains," *Breast cancer research and treatment*, vol. 132, no. 1, pp. 41–49, 2012.
- [110] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The jpeg2000 still image coding system: an overview," *IEEE transactions on consumer electronics*, vol. 46, no. 4, pp. 1103–1127, 2000.
- [111] F. Auli-Llinas and J. Serra-Sagrasta, "Jpeg2000 quality scalability without quality layers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 7, pp. 923–936, 2008.
- [112] G. Fan and W.-K. Cham, "Model-based edge reconstruction for low bit-rate wavelet-compressed images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 120–132, 2000.
- [113] D. Santa-Cruz, R. Grosbois, and T. Ebrahimi, "Jpeg 2000 performance evaluation and assessment," *Signal Processing: Image Communication*, vol. 17, no. 1, pp. 113–130, 2002.
- [114] A. Kanakatte, R. Subramanya, A. Delampady, R. Nayak, B. Purushothaman, and J. Gubbi, "Cloud solution for histopathological image analysis using region of interest based compression," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 1202–1205.
- [115] H. Helin, T. Tolonen, O. Ylinen, P. Tolonen, J. Näpänkangas, and J. Isola, "Optimized jpeg 2000 compression for efficient storage of histopathological whole-slide images," *Journal of pathology informatics*, vol. 9, 2018.
- [116] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *2016 eighth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2016, pp. 1–6.
- [117] M. Dejean-Servières, K. Desnos, K. Abdelouahab, W. Hamidouche, L. Morin, and M. Pelcat, "Study of the impact of standard image compression techniques on performance of image classification with a convolutional neural network," Ph.D. dissertation, INSA Rennes; Univ Rennes; IETR; Institut Pascal, 2017.
- [118] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [119] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.



## BIBLIOGRAPHY

---

- [120] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in neural information processing systems*, vol. 25, 2012.
- [121] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.
- [122] E. Kldiashvili, "Telemedicine for pathology," *Studies in Health technology and Informatics*, vol. 131, pp. 227–244, 2008.
- [123] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
- [124] A. J. de Groof, M. R. Struyvenberg, K. N. Fockens, J. van der Putten, F. van der Sommen, T. G. Boers, S. Zinger, R. Bisschops *et al.*, "Deep learning algorithm detection of barrett's neoplasia with high accuracy during live endoscopic procedures: a pilot study (with video)," *Gastrointestinal endoscopy*, vol. 91, no. 6, pp. 1242–1250, 2020.
- [125] A. Pourtaherian, H. J. Scholten, L. Kusters, S. Zinger, N. Mihajlovic, A. F. Kolen, F. Zuo, G. C. Ng *et al.*, "Medical instrument detection in 3-dimensional ultrasound data volumes," *IEEE transactions on medical imaging*, vol. 36, no. 8, pp. 1664–1675, 2017.
- [126] M. Barva, M. Uhercik, J.-M. Mari, J. Kybic, J.-R. Duhamel, H. Liebgott, V. Hlavác, and C. Cachard, "Parallel integral projection transform for straight electrode localization in 3-d ultrasound images," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 55, no. 7, pp. 1559–1569, 2008.
- [127] M. Uherčík, J. Kybic, Y. Zhao, C. Cachard, and H. Liebgott, "Line filtering for surgical tool localization in 3d ultrasound images," *Computers in biology and medicine*, vol. 43, no. 12, pp. 2036–2045, 2013.
- [128] P. Beigi, R. Rohling, S. E. Salcudean, and G. C. Ng, "Spectral analysis of the tremor motion for needle detection in curvilinear ultrasound via spatiotemporal linear sampling," *International journal of computer assisted radiology and surgery*, vol. 11, no. 6, pp. 1183–1192, 2016.
- [129] C. Mwikirize, J. L. Noshier, and I. Hacıhaliloglu, "Signal attenuation maps for needle enhancement and localization in 2d ultrasound," *International journal of computer assisted radiology and surgery*, vol. 13, no. 3, pp. 363–374, 2018.
- [130] A. Pourtaherian, N. Mihajlovic, S. Zinger, H. H. Korsten, P. H. de With, J. Huang, and G. C. Ng, "Automated in-plane visualization of steep needles from 3d ultrasound data volumes," in *2016 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2016, pp. 1–4.
- [131] A. Pourtaherian, F. Ghazvinian Zanjani, S. Zinger, N. Mihajlovic, G. Ng, H. Korsten *et al.*, "Improving needle detection in 3d ultrasound using orthogonal-plane convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 610–618.
- [132] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

- [133] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 640–651, 2016.
- [134] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
- [135] V. Sundaresan, C. P. Bridge, C. Ioannou, and J. A. Noble, "Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 671–674.
- [136] X. Yang, L. Yu, S. Li, X. Wang, N. Wang, J. Qin, D. Ni, and P.-A. Heng, "Towards automatic semantic segmentation in volumetric ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 711–719.
- [137] A. Prason, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2013, pp. 246–253.
- [138] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 1, pp. 23–38, 1998.
- [139] T. Tieleman, G. Hinton *et al.*, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.
- [140] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [141] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [142] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [143] C. Papalazarou, P. H. De With, and P. Rongen, "Sparse-plus-dense-ransac for estimation of multiple complex curvilinear models in 2d and 3d," *Pattern Recognition*, vol. 46, no. 3, pp. 925–935, 2013.
- [144] H. Yang, C. Shan, A. F. Kolen, and P. H. de With, "Improving catheter segmentation & localization in 3d cardiac ultrasound using direction-fused fcn," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1122–1126.
- [145] H. Yang, C. Shan, T. Tan, A. F. Kolen *et al.*, "Transferring from ex-vivo to in-vivo: Instrument localization in 3d cardiac ultrasound using pyramid-unet with hybrid loss," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 263–271.
- [146] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

## BIBLIOGRAPHY

---

- [147] A. Pourtaherian, S. Zinger, H. H. Korsten, N. Mihajlovic *et al.*, "Benchmarking of state-of-the-art needle detection algorithms in 3d ultrasound data volumes," in *Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9415. International Society for Optics and Photonics, 2015, p. 94152B.
- [148] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [149] P. Medina-Sotomayor, A. Pascual-Moscardó, and I. Camps, "Relationship between resolution and accuracy of four intraoral scanners in complete-arch impressions," *Journal of clinical and experimental dentistry*, vol. 10, no. 4, p. e361, 2018.
- [150] F. Ghazvinian Zanjani, D. Anssari Moin, B. Verheij, F. Claessen, T. Cherici, T. Tan, and P. de With, "Deep learning approach to semantic segmentation in 3D point cloud intra-oral scans of teeth," in *Proc. 2nd Int. Conf. Medical Imaging with Deep Learning (MIDL)*, vol. 102. PMLR, 2019, pp. 557–571.
- [151] T. Kondo, S. Ong, and K. W. Foong, "Tooth segmentation of dental study models using range images," *IEEE Transactions on medical imaging*, vol. 23, no. 3, pp. 350–362, 2004.
- [152] N. Wongwaen and C. Sinthanayothin, "Computerized algorithm for 3D teeth segmentation," in *International Conference on Electronics and Information Engineering (ICEIE)*. IEEE, 2010, pp. 277–280.
- [153] T. Yuan, W. Liao, N. Dai, X. Cheng, and Q. Yu, "Single-tooth modeling for 3D dental model," *Journal of Biomedical Imaging*, vol. 2010, p. 9, 2010.
- [154] Y. Kumar, R. Janardan, B. Larson, and J. Moon, "Improved segmentation of teeth in dental models," *Computer-Aided Design and Applications*, vol. 8, no. 2, pp. 211–224, 2011.
- [155] M. Yaqi and L. Zhongke, "Computer aided orthodontics treatment by virtual segmentation and adjustment," in *International Conference on Image Analysis and Signal Processing (IASP)*. IEEE, 2010, pp. 336–339.
- [156] S. M. Yamany and A. M. El-Bialy, "Efficient free-form surface representation with application in orthodontics," in *Three-Dimensional Image Capture and Applications II*, vol. 3640. International Society for Optics and Photonics, 1999, pp. 115–125.
- [157] M. Zhao, L. Ma, W. Tan, and D. Nie, "Interactive tooth segmentation of dental models," in *27th Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS)*. IEEE, 2006, pp. 654–657.
- [158] Z. Li, X. Ning, and Z. Wang, "A fast segmentation method for stl teeth model," in *International Conference on Complex Medical Engineering (ICME)*. IEEE, 2007, pp. 163–166.
- [159] M. Grzegorzec, M. Trierscheid, D. Papoutsis, and D. Paulus, "A multi-stage approach for 3D teeth segmentation from dentition surfaces," in *International Conference on Image and Signal Processing*. Springer, 2010, pp. 521–530.
- [160] T. Kronfeld, D. Brunner, and G. Brunnett, "Snake-based segmentation of teeth from virtual dental casts," *Computer-Aided Design and Applications*, vol. 7, no. 2, pp. 221–233, 2010.

- 
- [161] B. Zou, S. Liu, S. Liao, X. Ding, and Y. Liang, "Interactive tooth partition of dental mesh base on tooth-target harmonic field," *Computers in biology and medicine*, vol. 56, pp. 132–144, 2015.
- [162] K. Guo, D. Zou, and X. Chen, "3d mesh labeling via deep convolutional neural networks," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 1, p. 3, 2015.
- [163] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong, "3d deep shape descriptor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2319–2328.
- [164] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, vol. 1, no. 2, p. 4, 2017.
- [165] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.
- [166] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view cnns for object classification on 3d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5648–5656.
- [167] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3d shape segmentation with projective convolutional networks," in *Proc. CVPR*, vol. 1, no. 2, 2017, p. 8.
- [168] S. Ravanbakhsh, J. Schneider, and B. Póczos, "Deep learning with sets and point clouds," *arXiv preprint arXiv:1611.04500*, 2016.
- [169] Y. Li, R. Bu, M. Sun, and B. Chen, "Pointcnn," *arXiv preprint arXiv:1801.07791*, 2018.
- [170] T. Le and Y. Duan, "Pointgrid: A deep network for 3d shape understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9204–9214.
- [171] S. Chen, D. Tian, C. Feng, A. Vetro, and J. Kovačević, "Fast resampling of three-dimensional point clouds via graphs," *IEEE Transactions on Signal Processing*, vol. 66, no. 3, pp. 666–681, 2018.
- [172] H. Huang, S. Wu, M. Gong, D. Cohen-Or, U. Ascher, and H. R. Zhang, "Edge-aware point set resampling," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 1, p. 9, 2013.
- [173] M. Skrodzki, J. Jansen, and K. Polthier, "Directional density measure to intrinsically estimate and counteract non-uniformity in point clouds," *Computer Aided Geometric Design*, 2018.
- [174] M. Ghafoorian, C. Nugteren, N. Baka, O. Booi, and M. Hofmann, "El-gan: Embedding loss driven generative adversarial networks for lane detection," *arXiv preprint arXiv:1806.05525*, 2018.
- [175] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.
- [176] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing, "Scan: Structure correcting adversarial network for organ segmentation in chest x-rays," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 263–273.

- [177] Y. Huo, Z. Xu, S. Bao, C. Bermudez, A. J. Plassard, J. Liu, Y. Yao, A. Assad *et al.*, "Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks," in *Medical Imaging 2018: Image Processing*, vol. 10574. International Society for Optics and Photonics, 2018, p. 1057409.
- [178] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein, "Adversarial networks for the detection of aggressive prostate cancer," *arXiv preprint arXiv:1702.08014*, 2017.
- [179] P. Moeskops, M. Veta, M. W. Lafarge, K. A. Eppenhof, and J. P. Pluim, "Adversarial training and dilated convolutions for brain mri segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 56–64.
- [180] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "Segan: Adversarial network with multi-scale l1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3-4, pp. 383–392, 2018.
- [181] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, and D. Comaniciu, "Automatic liver segmentation using an adversarial image-to-image network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 507–515.
- [182] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," *arXiv preprint arXiv:1705.07115*, vol. 3, 2017.
- [183] V. Kumar, "Algorithms for constraint-satisfaction problems: A survey," *AI magazine*, vol. 13, no. 1, pp. 32–32, 1992.
- [184] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 918–927.
- [185] J. Hou, A. Dai, and M. Nießner, "3D-SIS: 3D semantic instance segmentation of RGB-D scans," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4421–4430.
- [186] L. Yi, W. Zhao, H. Wang, M. Sung, and L. Guibas, "GSPN: Generative shape proposal network for 3D instance segmentation in point cloud," *arXiv preprint arXiv:1812.03320*, 2018.
- [187] W. Wang, R. Yu, Q. Huang, and U. Neumann, "Sgpn: Similarity group proposal network for 3d point cloud instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2569–2578.
- [188] C. Liu and Y. Furukawa, "Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation," *arXiv preprint arXiv:1902.04478*, 2019.
- [189] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.
- [190] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9277–9286.

- 
- [191] S. Shi, X. Wang, and H. Li, "Pointtrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.
  - [192] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 2961–2969.
  - [193] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in neural information processing systems*, 2017, pp. 5099–5108.
  - [194] P. Hermosilla, P.-P. Ritschel, Tobias and Vázquez, À. Vinacua, and T. Ropinski, "Monte Carlo convolution for learning on non-uniformly sampled point clouds," in *SIGGRAPH Asia 2018 Technical Papers*. ACM, 2018, p. 235.
  - [195] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems (NIPS)*, 2015, pp. 91–99.
  - [196] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.
  - [197] F. G. Zanjani, A. Panteli, S. Zinger, F. van der Sommen, T. Tan, B. Balluff, D. N. Vos, S. R. Ellis *et al.*, "Cancer detection in mass spectrometry imaging data by recurrent neural networks," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 674–678.
  - [198] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 206–215.
  - [199] K. Jordan and P. Mordohai, "A quantitative evaluation of surface normal estimation in point clouds," in *International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 4220–4226.
  - [200] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead *et al.*, "Segment anything," *arXiv:2304.02643*, 2023.



# Acknowledgment

I am profoundly grateful for the guidance, support, and companionship of numerous individuals whose contributions have made this thesis possible. I would like to extend my heartfelt appreciation to the following individuals:

First and foremost, I extend a special thank you to my primary advisor, Prof. Peter de With, for providing me with the opportunity to join his esteemed Video Coding and Architecture (VCA) research group. His unwavering support, invaluable guidance, and dedication to excellence have been instrumental in shaping my research, including the completion of this thesis. Many late evenings were spent tirelessly improving my work, for which I am truly indebted.

I would also like to express my sincere gratitude to my co-advisor, Dr. Sveta Zinger, who showed faith in me by inviting me to join TUE as a PhD candidate. Her unwavering encouragement during moments of doubt and challenges in project management has been invaluable.

Throughout my PhD journey, I had the privilege of collaborating with esteemed experts at Promaton Ltd, a leading company in AI-enabled computational dentistry. I am deeply thankful to David and Bass for entrusting me with the project and providing funding. My heartfelt appreciation goes to Frank, Sarah, and Teo for their warm welcome and continuous support.

I wish to extend my thanks to all the members of the promotion committee for their time and effort in reviewing this thesis. Particularly, I am grateful to Prof. Peter Schelkens from Vrije Universiteit Brussel for his constructive feedback and supervision on image compression during our 3D histopathology ITEA project. I also extend my gratitude to Prof. Jeroen van der Laak from Radboud University of Nijmegen for his invaluable feedback and guidance on our joint publications on digital pathology. I am honored to have distinguished professors, including Prof. Arnold Smeulders from the University of Amsterdam, Prof. Massimo Mischi, and Prof. Josien Pluim from Eindhoven University of Technology as part of my PhD committee.

I would like to acknowledge my former colleagues at the VCA group for creating an energizing and friendly environment. Special thanks to Arash for his collaboration on Ultrasound Image analysis and significant contributions to this thesis. My appreciation also goes to Marieke and Anja for their assistance in various project tasks and maintaining supplies. Thanks to all my former colleagues, especially Fons, Marco, Amir, Joost, Ariyan, Cheng, Joy, Ronald, Hani, Willem, Hongxu, Egor, Gijs, Francesca, Anweshan, Roger, Panos, Dennis, Herman, Patrick, Raffaele, Clint, Luis, and Jos.

Lastly, my heartfelt gratitude to my family—Ali and Marjan—for their immeasurable support, continuous care, and encouragement to pursue my dreams.



## ACKNOWLEDGMENT

---

Without the contributions of all these incredible individuals, this thesis would not have been possible. I am truly grateful for their impact on my academic and personal growth.

# Publication list

## Journal articles

- [J-1] **F. G. Zanjani**, A. Pourtaherian, S. Zinger, D. A. Moin, F. Claessen, T. Cherici, S. Parinussa, and P. H. N. de With, "Mask-MCNet: tooth instance segmentation in 3D point clouds of intra-oral scans," *Neurocomputing*, vol. 453, pp. 286–298, Elsevier, 2021.
- [J-2] J. Xing, Z. Li, B. Wang, Y. Qi, B. Yu, **F. G. Zanjani**, A. Zheng, R. Duits, T. Tan, "Lesion segmentation in ultrasound using semi-pixel-wise cycle generative adversarial nets," *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, 2020.
- [J-3] **F. G. Zanjani**, S. Zinger, B. Piepers, S. Mahmoudpour, P. Schelkens, and P. H. N. de With, "Impact of JPEG 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images," *Journal of Medical Imaging*, vol. 6, no. 2, pp. 027501, 2019.
- [J-4] A. Pourtaherian, **F. G. Zanjani**, S. Zinger, N. Mihajlovic, G. C. Ng, H. H. M. Korsten, and P. H. N. de With, "Robust and semantic needle detection in 3D ultrasound using orthogonal-plane convolutional neural networks," *Int. J. Computer Assisted Radiology and Surgery (IJCARS)*, vol. 13, no. 9, pp. 1311–1323, 2018.
- [J-5] P. Bandi, O. Geessink, Q. Manson, M. Dijk, M. Balkenhol, M. Hermesen, B. E. Benjordi, B. Lee, K. Paeng, A. Zhong, Q. Li, **F. G. Zanjani**, S. Zinger et al. "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE transactions on medical imaging (TMI)*, vol. 38, no. 2, pp. 550–560, 2018.
- [J-6] T. Tan, Z. Li, H. Liu, **F. G. Zanjani**, Q. Ouyang, Y. Tang, Z. Hu, Q. Li, "Optimize transfer learning for lung diseases in bronchoscopy using a new concept: sequential fine-tuning," *IEEE journal of translational engineering in health and medicine*, vol. 6, pp. 1–8, 2018.

## International conference proceedings

- [C-1] **F. G. Zanjani**, D. A. Moin, F. Claessen, T. Cherici, S. Parinussa, A. Pourtaherian, S. Zinger, and P. H. N. de With, "Mask-MCNet: Instance segmentation in 3D point cloud of intra-oral scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 128–136, Springer, 2019.

- [C-2] **F. G. Zanjani**, D. A. Moin, B. Verheij, F. Claessen, T. Cherici, T. Tan, S. Zinger, and P. H. N. de With, "Deep learning approach to semantic segmentation in 3D point cloud intra-oral scans of teeth," in *International Conference on Medical Imaging with Deep Learning (MIDL)*, pp. 557–571, PMLR, 2019.
- [C-3] **F. G. Zanjani**, A. Panteli, S. Zinger, F. van der Sommen, T. Tan, B. Balluff, N. Vos, S. Ellis, RMA Heeran, M. Lucas, H. Marquering, and P. H. N. de With, "Cancer detection in mass spectrometry imaging data by recurrent neural networks," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, pp. 674–678, IEEE, 2019.
- [C-4] J. van Kersbergen, **F. G. Zanjani**, S. Zinger, F. van der Sommen, B. Balluff, N. Vos, S. Ellis, RMA Heeran, M. Lucas, H. Marquering, and P. H. N. de With, "Cancer detection in mass spectrometry imaging data by dilated convolutional neural networks," in *Medical Imaging 2019: Digital Pathology*, vol. 10956, pp. 94–101, SPIE, 2019.
- [C-5] **F. G. Zanjani**, S. Zinger, and P. H. N. de With, "Deep convolutional gaussian mixture model for stain-color normalization of histopathological images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 274–282, Springer, 2018.
- [C-6] **F. G. Zanjani**, S. Zinger, B. E. Bejnordi, J. AWM van der Laak, and P. H. N. de With, "Histopathology stain-color normalization using deep generative models," in *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.
- [C-7] **F. G. Zanjani**, S. Zinger, B. E. Bejnordi, J. AWM van der Laak, and P. H. N. de With, "Stain normalization of histopathology images using generative adversarial networks," in *IEEE 15th International symposium on biomedical imaging (ISBI)*, pp. 573–577, 2018.
- [C-8] A. Pourtaherian, N. Mihajlovic, **F. G. Zanjani**, S. Zinger, G. C. Ng, H. H. M. Korsten, and P. H. N. de With, "Localization of partially visible needles in 3D ultrasound using dilated convolutional neural networks," *Proc. IEEE Int. Ultrasonics Symp. (IUS)*, Kobe, Japan, 2018.
- [C-9] **F. G. Zanjani**, S. Zinger, and P. H. N. de With, "Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces," in *SPIE Medical imaging 2018: Digital pathology*, vol. 10581, pp. 105810I, 2018.
- [C-10] **F. G. Zanjani**, A. Pourtaherian, X. Tang, S. Zinger, N. Mihajlovic, G. C. Ng, H. H. M. Korsten, and P. H. N. de With, "Coherent needle detection in ultrasound volumes using 3D conditional random fields," in *Proc. SPIE Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling*, Houston, TX, USA, p. 105760W, 2018.
- [C-11] A. Pourtaherian, **F. G. Zanjani**, S. Zinger, N. Mihajlovic, G. C. Ng, H. H. M. Korsten, and P. H. N. de With, "Improving needle detection in 3D ultrasound using orthogonal-plane convolutional networks," in *Proc. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Québec City, Québec, Canada, pp. 610–618, 2017.

### International patent applications

- [P-1] **F. G. Zanjani**, T. Cherici, F. Claessen, "Object detection and instance segmentation of 3d point clouds based on deep learning," Patent Application WO2021009258A1, Jan. 1, 2021.

- [P-2] F. T. C. Claessen, D. A. Moin, T. Cherici, **F. G. Zanjani**, "Automated semantic segmentation of non-euclidean 3d data sets using deep learning," Patent Application 17415465, March. 3, 2022.
- [P-3] F. T. C. Claessen, D. A. Moin, T. Cherici, **F. G. Zanjani**, " Method implemented in computer, computer program system and product," Patent Application BR112021011722A2, August. 31, 2020.



## Curriculum vitae



Farhad Ghazvinian Zanjani obtained his BSc. degree in Electrical Engineering from University of Karaj, Iran in 2004. He obtained his MSc. degree in Biomedical Engineering, specialization in image processing from Amirkabir University of Technology, Tehran, Iran in 2007. He worked for seven years as a computer vision engineer and later as technical lead in a start-up company in Tehran by designing and developing various algorithms for 3D perception. He then moved to the Netherlands to pursue his academic career, where he obtained (with honor) a MSc. degree in Computer Science, specialization in Machine Learning at the Radboud University, Nijmegen in

2016. He then continued his research as a PhD candidate in the Video Coding and Architecture - Signal Processing Systems group at the Electrical Engineering Faculty of the Eindhoven University of Technology (TU/e). His PhD research focused on designing several deep learning methods for medical image analysis in various directions, including the computational pathology, ultrasound image-guided interventions, and 3D point cloud analysis. This research has resulted in 3 patent applications and about 17 publications in international peer-reviewed scientific journals and top international conferences in medical imaging with deep learning. Several of the techniques developed during this period have been successfully transferred to several Business Groups, including the Philips Digital Pathology, Philips Ultrasound, and the Promaton Company for future AI product development efforts. Currently, he is a senior scientific researcher at Qualcomm AI Research, developing deep learning methods for 3D perception problems by using both wireless signals and vision sensory data.

