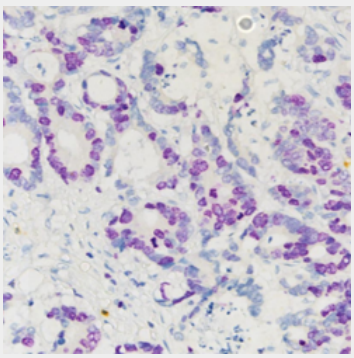
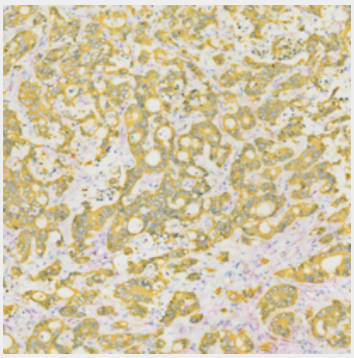
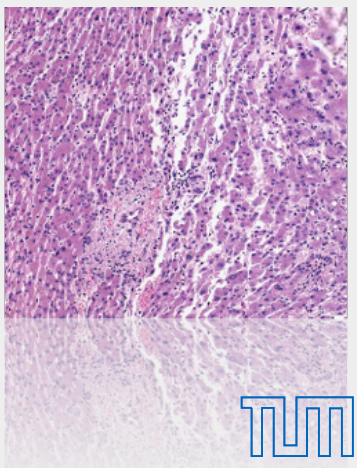
Computer Aided Medical Procedures
Prof. Dr. Nassir Navab







Fakultät für Informatik
Technische Universität München

Roche

Dissertation

Deep Learning Solutions for Cancer Drug Development in Digital Pathology

Amal Lahiani

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin

Deep Learning Solutions for Cancer Drug Development in Digital Pathology

Amal Lahiani

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r): Prof. Dr. Stephan Jonas

Prüfer der Dissertation: 1. Prof. Dr. Nassir Navab

2. Prof. Dr. Nasir M. Rajpoot
University of Warwick, United Kingdom

3. Prof. Dr. Wilko Weichert

Die Dissertation wurde am 18.02.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 06.08.2020 angenommen.

Amal Lahiani
Deep Learning Solutions for Cancer Drug Development in Digital Pathology
Dissertation

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin
Boltzmannstraße 3
85748 and Garching bei München

Abstract

In the domain of oncology, pathology tasks such as patient diagnosis and cancer drug development, have been revolutionized since the growing availability and quality of high resolution whole slide scanners. The transition from conventional glass slide microscopic assessment to digital pathology has been beneficial and promising for both pathologists and researchers due to the ability to collect, process and analyse much larger datasets with more stains (clinical markers) using more objective, accurate and consistent methods. Furthermore, with the recent advances in computational power, deep learning solutions have been developed allowing to explore previously unachievable predictive algorithms and to enhance the value and insights that can be generated from pathology tissue images. However, different challenges need to be taken into consideration in order to make the applications more efficient and to increase confidence and adoption in the very conservative field of human pathology. These challenges include, among others, application generalization over multiple stains, limitations in terms of tissue staining materials and procedures, size of histologic whole slide images (WSIs) and intra and inter-pathologist variability. In this thesis, we explore new methodologies, techniques and deep learning solutions to the aforementioned challenges in the context of two different applications: stain generalization and stain virtualization applied to digital images of Colorectal Carcinoma metastases in liver tissue from biopsy and surgical specimen.

Slide annotation is a key step in pathology routines and in cancer biomarker research aiming to quantify pattern changes in microscopic WSIs of tumor biopsies. The process of manual annotation can be tedious and subjective, especially in the context of drug development research where a correlative analysis of multiple biomarkers is required. In the first part of this dissertation, we elaborate on a supervised learning based method that allows to segment different functional compartments in histology images of various stainings including Hematoxylin and Eosin (H&E) staining and multiple immunohistochemistry (IHC) stainings. We show the effect of the proposed solution on increasing the stain generalization performance and dealing with dataset variability by comparing to state-of-the-art methods and using different visualization techniques.

In the second part we propose stain virtualization solutions allowing to virtually generate an IHC staining from different input stainings using unsupervised learning methods. Stain virtualization presents many advantages to the pathology drug development workflow, such as reduction of lab workload, reduction in tissue and costs and multiplexing of different biomarkers on the same tissue and with the same coordinate system. In this context, we additionally propose solutions to one of the main challenges of high resolution style transfer in general and to WSI processing in particular, i.e. the tiling artifact caused by the necessity of tilewise processing. We describe different experiments and visualizations aiming to prove the hypothesis and we validate the application mathematically and histologically.

Zusammenfassung

Im Bereich der Onkologie sind die Aufgaben der Pathologie, wie z.B. die Patientendiagnose und die Entwicklung von Krebsmedikamenten, seit der zunehmenden Verfügbarkeit und Qualität von hochauflösenden Whole Slide Scannern revolutioniert worden. Der Übergang von der konventionellen gläsernen Objektträger-Mikroskopie zur digitalen Pathologie war sowohl für Pathologen als auch für Forscher vorteilhaft und vielversprechend, da sie nun in der Lage sind, viel größere Datensätze mit mehr Färbungen (klinische Markern) mit objektiveren, genaueren und konsistenteren Methoden zu sammeln, zu verarbeiten und zu analysieren. Darüber hinaus wurden mit den jüngsten Fortschritten bei der Rechenleistung Deep Learning entwickelt, die es ermöglichen, bisher unerreichbare prädiktive Algorithmen zu erforschen und den Wert und die Erkenntnisse zu verbessern, die aus Bildern von pathologischem Gewebe generiert werden können. Es müssen jedoch verschiedene Herausforderungen berücksichtigt werden, um die Anwendungen effizienter zu gestalten und das Vertrauen und die Akzeptanz im sehr konservativen Bereich der Humanpathologie zu erhöhen. Zu diesen Herausforderungen gehören unter anderem die Verallgemeinerung der Anwendung von unterschiedlichen Färbungen, Einschränkungen hinsichtlich der Materialien und Verfahren der Gewebefärbung, die Größe der histologischen Ganzkörperdarstellung (WSIs) und die Variabilität innerhalb und zwischen den Pathologen. In dieser Arbeit erforschen wir neue Methoden, Techniken und Deep Learning für die oben genannten Herausforderungen im Kontext zweier verschiedener Anwendungen: Generalisierung und Virtualisierung von Gewebsfärbungen, die auf digitale Bilder von kolorektalen Karzinom-Metastasen in Lebergewebe aus Biopsie- und Operationspräparaten angewendet werden.

Die Annotation von Objektträgern ist ein wichtiger Schritt in der Pathologieroutine und in der Krebs-Biomarker-Forschung, um Musterveränderungen in mikroskopischen WSIs von Tumorbiopsien zu quantifizieren. Der Prozess der manuellen Annotation kann langwierig und subjektiv sein, insbesondere im Kontext der Forschung in der Medikamentenentwicklung, wo eine korrelative Analyse mehrerer Biomarker erforderlich ist. Im ersten Teil dieser Dissertation arbeiten wir an einer lernbasierten Methode, die es erlaubt, verschiedene funktionelle Kompartimente in histologischen Bildern verschiedener Färbungen, einschließlich Hämatoxylin- und Eosin (H&E)-Färbungen und multipler Immunhistochemie (IHC)-Färbungen, zu segmentieren. Wir zeigen den Effekt der vorgeschlagenen Lösung auf die Leistungssteigerung der Färbegeneralisierung und den Umgang mit der Variabilität des Datensatzes durch den Vergleich mit modernsten Methoden und die Verwendung verschiedener Visualisierungstechniken.

Im zweiten Teil schlagen wir Lösungen zur Virtualisierung der Färbung vor, die es erlauben, eine IHC-Färbung aus verschiedenen Eingangsfärbungen mit Hilfe von unüberwachten Lernmethoden virtuell zu erzeugen. Die Färbe-Virtualisierung bietet viele Vorteile für den Arbeitsablauf bei der Entwicklung von Medikamenten in der Pathologie, wie z.B. die Redu-

zierung des Arbeitsaufwands im Labor, die Reduzierung von Gewebe und Kosten und die Kombination von Färbungen verschiedener Biomarker auf demselben Gewebe und mit demselben Koordinatensystem. In diesem Zusammenhang schlagen wir zusätzlich Lösungen für eine der größten Herausforderungen der hochauflösenden Stilübertragung (Style Transfer) im Allgemeinen und der WSI-Verarbeitung im Besonderen vor, nämlich das durch die Notwendigkeit der kachelweisen Verarbeitung verursachte Kachel-Artefakt. Wir beschreiben verschiedene Experimente und Visualisierungen, um die Hypothese zu beweisen und wir validieren die Anwendung mathematisch und histologisch.

Acknowledgments

I would like to start with a quote that I profoundly like said by John F. Kennedy: "We must find time to stop and thank the people who make a difference in our lives." For this reason, at this point, I would like to stop and express my gratitude to all the amazing people I have been blessed to have with me during my PhD journey. This work would not have been possible without the precious support and encouragement of all of them.

First of all, I wish to express my sincere thanks to my advisor, Prof. Nassir Navab for giving me the chance to have this enriching experience, for granting me the freedom to explore different research areas and for being always available for guidance and advice. I would like to thank Dr. Oliver Grimm for offering me the opportunity to enjoy working in the "Early Biomarker Development Oncology" department at Roche Innovation Center Munich and for his continuous trust and assistance. Additionally, I am thankful to Prof. Wilko Weichert for accepting to be my second supervisor, for being always supportive and for his enthusiasm and continuous encouragement.

I would like to express my special thanks and heartiest gratitude to Eldad Klaiman who positively marked my PhD journey with all its happy and challenging moments. No words can express how grateful I am to him for his valuable guidance, cheerful enthusiasm and continuous supervision. He was never tired of helping and advising me and I fully enjoyed working and having long and productive discussions with him. I will never forget what he always tells me "what could possibly go wrong, try it !". Further, I wish to thank Dr. Shadi Albarqouni for his valuable supervision and his continuous advice and encouragement. I am grateful for all the fruitful discussions I had with him during the PhD and I appreciate his help and support. I would also like to thank Dr. Maximilian Baust for his help, advice and support in the beginning of my PhD.

I wish to extend my thanks to many amazing people from Roche. Special thanks go to Dr. Fabian Gaire and Dr. Bruno Gomes for their support. I would also like to thank Dr. Irina Klaman and Dr. Marta Canamero for their valuable pathological contribution. I am grateful to Sabine Moosmann, Marc Zaetschky, Dr. Konstanty Korksi, Natalie Zwing, Katy Wilson, Dr. Claudia Ferreira, Georges Marchal, Dr. Natascha Rieder, Dr. Suzana Vega Harring, Dr. Christin Ruoff, Dr. Juha Lindner, Dr. Astrid Heller, Birgit Hartmann, Dr. Franclim Ribeiro, Gabriele Dietmann, Quincy Wong, Dr. Fabian Schmich, Veronique Kayser, Dr. Khen Sagiv, Jacob Gildenblat and many more. Thank you all for making my PhD journey more exciting and pleasant.

I would also like to thank my friends and colleagues from the chair of Computer Aided Medical Procedures (CAMP) at TUM: Mai Bui, Gerome Vivar, Anees Kazi, Roger Soberanis Mukul, Christoph Baur, Hasan Sarhan, Abhijit Guha Roy, Dr. Tingying Peng, Azade Farshad, Agnieszka Tomczak, Ashkan Khakzar, Tariq Bdair and many more. Thank you for the insightful discussions and the nice moments we shared together. Special thanks go to Martina Hilla. Moreover, I wish to thank the Graduate School of Bioengineering (GSB) for their help and support. In particular, I would like to thank Dr. Anja Drescher for her advice, continuous support and enthusiasm. I am also grateful to many amazing people that I met at the GSB: Marwan Muhammad, John LaMaster, Dhritimann Das and Kaushik Basak Chowdhury.

Last but not least, I would like to express my deepest and heartiest gratitude to the most amazing and wonderful people in the world, my Dad Mongi Lahiani and my Mom Hela Trigui, for their enduring support, love and advice. Thank you for always believing in me, encouraging me and keeping my morale high during the relaxed and the challenging periods, thank you for being the best parents in the world. This achievement would not have been possible without you. Further, I am very thankful to my brother Yessine Lahiani for his continuous support and encouragement. Thank you for being such an amazing brother. My special thanks go to my dear cousins and friends Asma Ben Said, Yosra Drissi, Maha Rekik, Omaima Kammoun, Maissa Sghaier and Jessica Luzio. Finally, I would like to dedicate this work to the memory of my grandfather who left us in the last months of the PhD.

Contents

List of Authored and Co-authored Publications

2020

[1] **Amal Lahiani**, Irina Klaman, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. "Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency". *IEEE Journal of Biomedical and Health Informatics (JBHI), 2020*.

2019

[2] **Amal Lahiani**, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. "Perceptual embedding consistency for seamless reconstruction of tilewise style transfer". *Medical Image Computing and Computer Assisted Intervention (MICCAI), 2019*.

[3] **Amal Lahiani**, Jacob Gildenblat, Irina Klaman, Shadi Albarqouni, Nassir Navab, and Eldad Klaiman. "Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach.". *European Congress of Digital Pathology (ECDP), 2019*.

[4] **Amal Lahiani**, Jacob Gildenblat, Irina Klaman, Nassir Navab, and Eldad Klaiman. "Generalising multistain immunohistochemistry tissue segmentation using end-to-end colour deconvolution deep neural networks". *IET Image Processing, 2019)*.

2018

[5] **Amal Lahiani**, Eldad Klaiman, and Oliver Grimm. "Enabling histopathological annotations on immunofluorescent images through virtualization of hematoxylin and eosin.". *Journal of Pathology Informatics (JPI), 2018*.

Part I

Introduction and Background

1 Introduction

> *For AI to add the most value and for patients and physicians to embrace it, it needs to support, not supplant, the patient-physician relationship... AI will be most effective when it enhances physicians' ability to focus their full attention on the patient by shifting the physicians' responsibilities away from transactional tasks toward personalized care that lies at the heart of human healing.*
>
> **— Steven Lin**
> (MD, a clinical assistant professor of medicine and vice chief for technology innovation in Stanford University's division of primary care and population health)

1.1 Motivation and Objective

In oncology as in any other medical field, the patient is in the center of clinical and research efforts. With the recent technological advances and the success of digital data analysis, histopathology started to benefit from the new digital solutions. Generating insights from histological data and translating them into patient diagnosis and drug development innovations became the main motivations of using computerized image analysis in digital pathology. Effective collaboration between researchers and pathologists has been enabled by the success of Whole Slide Imaging which allowed to digitize traditional histological glass slides and helped to make patient diagnosis more efficient, personalized, fast and accurate. As the traditional approaches in pathology rely mainly on subjective human visual assessment of glass slides under a microscope by an expert pathologist they are therefore limited in many ways. For example, human ability to interpret these images limits the number of stains per image as well as complex quantification methods. The transition from conventional glass slide visual evaluation to digital pathology resulted in large amounts of digital high resolution Whole Slide Images (WSIs) which are possible to share and remotely visualize. While this capacity to generate huge amounts of digital histological data opened up promising opportunities, it did also come with the need to develop image analysis solutions in order to extract as much useful and relevant information as possible. This histological data revolution is what motivated the research in the area of digital pathology in order to bring computer vision and machine learning solutions to the field, assist pathologists in decision making and enhance drug and biomarker development research. So instead of being restricted to traditional diagnosis methods based on time-consuming visual microscopic glass slide assessment, today's

pathologists are rather clinical experts with the capacity to use computer based systems, interpret new digital solution outputs and perform their task more consistently, rapidly and accurately. This combination of clinical and technological knowledge is pushing histopathology in the direction of personalized patient care where each treated patient becomes the center of all healthcare contributor interest. However, the aforementioned change in pathological practice and patient situation would not have been possible without the talent and contribution of interdisciplinary researchers and scientists who made significant efforts in linking the clinical applications to computer science solutions.

This new trend has been growing and expanding gradually and continuously since two decades and it does nowadays include a broad range of pathology applications such as tissue compartment segmentation, cell or tumor subtype classification, biomarker quantification, tumor grading, outcome prediction, registration and more recently stain virtualization. That being said, all this was not challenge free and pathology digitization has been facing continuous and complex issues starting from pathologist reluctance and lack of digital system trust in the very beginning to more technical challenges such as data storage and management, staining and scanning artifacts, color variations and high resolution WSI size. Since the introduction of computer science to histopathology practice, digital pathology scientists have been working continuously in order to smartly address the aforementioned challenges. However, the full potential of the research area is still far from being reached which is making the field more and more attractive to researchers.

In this context of digital pathology expansion and machine learning success, the main motivation of this work is to leverage the concept of learning from data in order to generate valuable insights from histological images and optimize histology routines and workflows. We addressed several digital pathology challenges in the contexts of tissue compartment segmentation and stain virtualization. We believe that transferring and adapting computer vision knowledge to histopathology and combining it with pathologist advanced digital training would provide pathologists with valuable information which is otherwise unattainable or difficult to reach on time and make patient care more accessible and optimized for patient benefit.

1.2 Contributions and Outline

This thesis is dedicated to developing deep learning solutions for stain generalization and virtualization in digital pathology. In the following, we present the structure of this dissertation.

Part I: Introduction and Background. In the first part of this dissertation we introduce the motivation and the context of our work. In the first chapter we present the objective of adopting digital solutions including machine learning based methods in histopathology. In the second chapter we present the field of digital pathology including its history and temporal evolution, its general workflow routines and its main applications and challenges.

Part II: Deep Learning and Digital Pathology. With the great progress of computational resources, deep learning models have been successfully applied for multiple tasks in natural

image processing and more recently in histopathological image analysis. The first chapter of this part contains the essentials and relevant technical details of deep learning in general including supervised and unsupervised learning. The second chapter contains the particularities of deep learning applied to digital pathology including the most frequent challenges, the most common practices and the importance of method validation.

Part III: Generalization of Multistain Immunohistochemistry Tissue Segmentation. A key challenge in cancer immunotherapy biomarker research is quantification of pattern changes in microscopic WSIs of tumor biopsies. Drug development requires correlative analysis of various biomarkers. To enable that, tissue slides are traditionally manually annotated by expert pathologists. As manual annotation of tissue slides is a tedious and error-prone task, part III of this dissertation presents a method to automatically segment digitized slide images with multiple stainings into different tissue compartments in order to improve accuracy and consistency while reducing workload and cost. We address the task in the context of drug development where multiple stains exist and look into solutions for generalizations over these image populations. We also apply visualization techniques to help understand the network decisions and gain more trust from pathologists.

Part IV: Unsupervised Stain Virtualization. Histopathological evaluation and analysis of tissue samples is a key practice in patient diagnosis and drug development, especially in oncology. Historically, Hematoxylin and Eosin (H&E) has been used by pathologists as a gold standard staining in classical histopathology. However, in many cases, various target specific stains, including in-situ hybridization (ISH) and immunohistochemistry (IHC), are needed in order to highlight and identify specific structures and/or targets in the tissue. As tissue is scarce and staining procedures are tedious and laborious, part IV of this dissertation introduces unsupervised deep learning approaches aiming to virtually generate images of stained tissue thus saving lab efforts, time and valuable tissue. Two combinations of input/output stainings with different validation and evaluation setups are presented in this part.

Part V: Conclusions and Outlook. In the last part of this dissertation we sum up and discuss the findings of our contributions and present our vision about the possible future directions in the field of digital pathology.

2 Digital Pathology

Pathology is the study and diagnosis of diseases including cancer and has long been associated with medical development and patient care and treatment. More specifically histopathology is a medical subdivision based on microscopic examination of biological tissues under a microscope and aiming to diagnose diseases, guide therapy and push drug development research in the domain of oncology [6]. Based on visual examination and assessment of expert pathologists, conventional histological methods are expensive, time consuming and prone to intra and inter-pathologist agreement. While the introduction of medical image computing has been smooth and successful over the last decades in several medical fields (e.g. radiology), the transition to digital pathology has been slower [7] due to image quality and resolution requirements, scanning and storage limitations and the cost of digital pathology systems [8]. However, recent and significant advances in computational power and Whole Slide Imaging allowed for an evolution of pathology solutions from traditional photomicroscopy to innovative digital imaging, hence improving the quality of patient care and the speed of medical research progresses. Whole Slide Imaging consists of digitizing conventional glass slides using high resolution scanners and utilizing software solutions to view and analyze the digital slides [7]. Since the introduction and success of WSI solutions, histology image data have been continuously growing at a rapid pace and several clinical and research oriented applications have been emerging and progressing allowing to improve the quality of patient care and to foster medical discovery.

2.1 History and Evolution

The term histology was introduced in 1819 by Karl Mayer as a combination of two Greek words: histos for tissues and logos for study [9]. However, the history of histology stretches back many centuries and goes hand in hand with the advent of microscopy. The first published work about microscopy "Micrographia" by Robert Hooke dates back from 1665 [10]. In 1673, a dutch scientist called Anthony van Leeuwenhoek made improvements to the microscope allowing to improve the magnification and used a dye for the first time to study a biological structure [9]. He therefore became known as the "Father of Microbiology" [11, 12]. Then it was not until the late 1800s and early 1900s that significant developments in the use of dyes and stains started and allowed to visualize cellular structures. In 1858 Joseph von Gerlach used the carmine dyes to differentiate nucleus and cytoplasm [13] and in 1875-1878 Wissowzky introduced the H&E staining [9]. At the same time, major advances were made in light microscopy after the elaboration of the diffraction limit theory [11]. In 1911 the first practical fluorescence microscope was developed by Oskar Heimstädt [14]. In 1986 the concept of telepathology was introduced by Ronald Weinstein [15, 16] and allowed to use remote robotic microscopy to transmit microscope images between remote locations. Then in the late 1990s the first WSI scanners were introduced [7] but they were very limited and slow.

In the 2000s more efficient WSI scanners were developed allowing to scan entire slides at high resolution within a reasonable time and with reasonable costs. This success made Wole Slide Imaging more widespread and helped the evolution and progress of digital pathology during the past two decades [8, 12, 17, 18, 19, 20, 21]. Nowadays modern WSI scanners with the capacity to scan high resolution entire slides within a few minutes became commercially available.

2.2 Tissue Preparation and General Workflow

When a biopsy or a surgical specimen is taken from a patient, it goes through a sequence of tissue preparation steps [22, 23, 24] including fixation, dehydration, clearing, infiltration, embedding, sectioning and staining in order to be ready for pathologist examination or scanning (Fig. 2.1).

Fixation

Fixation is a crucial step in tissue preparation allowing to permanently preserve the tissue structure, leave it as close as possible to its natural living state and prevent autolysis and bacterial attacks. Currently, several fixatives such as precipitant and cross-linking are being used in histopathology. Precipitant fixatives remove lipid and reduce the solubility of proteins while the cross-linking fixatives create links between adjacent protein chains. One of the most commonly used fixatives are formalin solutions which result in chemical and physical changes protecting the tissue against the next processing steps by slowly penetrating the tissue causing chemical and physical changes preserving and protecting the tissue. This step can take between 6 and 24 hours. After fixation, the specimens to be processed are trimmed, fit into suitable labelled cassettes and stored in formalin until ready to proceed.

Tissue Processing

Tissue processing include a series of steps that are needed in order to allow thin sections to be cut properly. For this, it is possible either to freeze the tissue and keep it frozen while cutting the sections or to infiltrate the tissue specimen with a liquid agent that can be converted into a solid having the appropriate conditions needed for the sectioning. The most common treatments make use of different reagents to process the fixed tissue and embed it in a stable medium like paraffin wax. In this case, the tissue processing steps include dehydration, clearing, infiltration and embedding.

- **Dehydration:** Since they are immiscible with paraffin wax, the water and the fixative contained in the specimen have to be removed before the specimen can be infiltrated with wax. Dehydration generally involves the immersion of the specimen in a series of ethanol solutions with progressively increasing alcoholic concentrations in order to prevent excessive tissue shrinking. After dehydration all the water and fixating solution are replaced by the alcohol.

- **Clearing:** Even after dehydrating the tissue, it is still not ready for wax infiltration because paraffin wax is also immiscible with alcohol. Therefore a solvent agent such

as Xylene is used to displace the ethanol. Clearing also helps to remove the substantial amount of fat from the tissue which otherwise presents a barrier to wax infiltration.

- **Wax infiltration:** After dehydration and clearing, the tissue becomes ready for histological wax infiltration. The most commonly used reagent is the molten paraffin wax. Paraffin wax melts at a temperature between 55°C and 63°C [24]. At these temperatures, the main morphological characteristics are preserved and the tissue structures are not altered. Paraffin displaces the clearing agent and the specimen becomes ready for embedding.

- **Embedding:** This step consists of removing the tissue samples from the cassettes and filling a suitable metal mold with the molten wax and the specimen. The specimen should be carefully oriented in the mold because its placement will define the plane of section. The labelled cassette is then placed on top of the mold and the whole set is placed on a cold plate to solidify and provide a support for sectioning.

- **Sectioning:** The embedded tissue is cut into thin slices of 2.5-5μM using a microtome. The cut sections are then floated in warm water to remove the wrinkles, mounted on glass slides and dried.

Tissue Staining

After processing and sectioning the tissue, a reverse tissue processing step is used to remove the paraffin wax and allow water soluble coloring dyes to penetrate the sections. Different dyes are used to stain different tissue structure. Each dye binds to a particular structure in the tissue resulting in a specific color. Then, the stained sections are cover-slipped with a piece of plastic or glass in order to protect the tissue and get a better visual quality of microscopic examination. The staining can be chromogenic or fluorescent. Chromogens are chemical compounds that give a specific chromogenic dye (color) after a reaction with a biological structure while fluorophores correspond to chemical compounds which emit specific wavelengths when excited by exposure to a particular wavelength of light. The standard staining protocol used in histopathology is the (H&E) staining as it gives a good overview of the tissue and the cellular components and it clearly shows different types of structures. Hematoxylin is a basic dye staining acidic components in shades of purple and blue and eosin is an acidic contrasting counterstain coloring the basic components (mostly proteins) in shades of pink and red [25]. In a normal H&E staining, cell nuclei are colored in blue and purple whereas cell cytoplasm and most connective tissue are colored in degrees of pink. In many cases, H&E staining only is not sufficient to get enough information about the disease and to analyse and understand the tumor microenvironment. In these cases additional IHC stainings are used in order to resolve uncertainties about the disease and ensure a better diagnosis. IHC stainings are based on the principle of antibodies binding to specific targets (antigens) which can be visualized with enzyme labels (dyes) tagged to antibodies and catalyzing a color-producing reaction [25].

After staining, the tissue is ready for microscopic examination by expert pathologists and for scanning (Fig. 2.1). Depending on the staining type, brightfield or fluorescent scanners can be used in order to produce digital WSIs at the required magnification. These digital images are then ready for storage, visualization and advanced analysis.

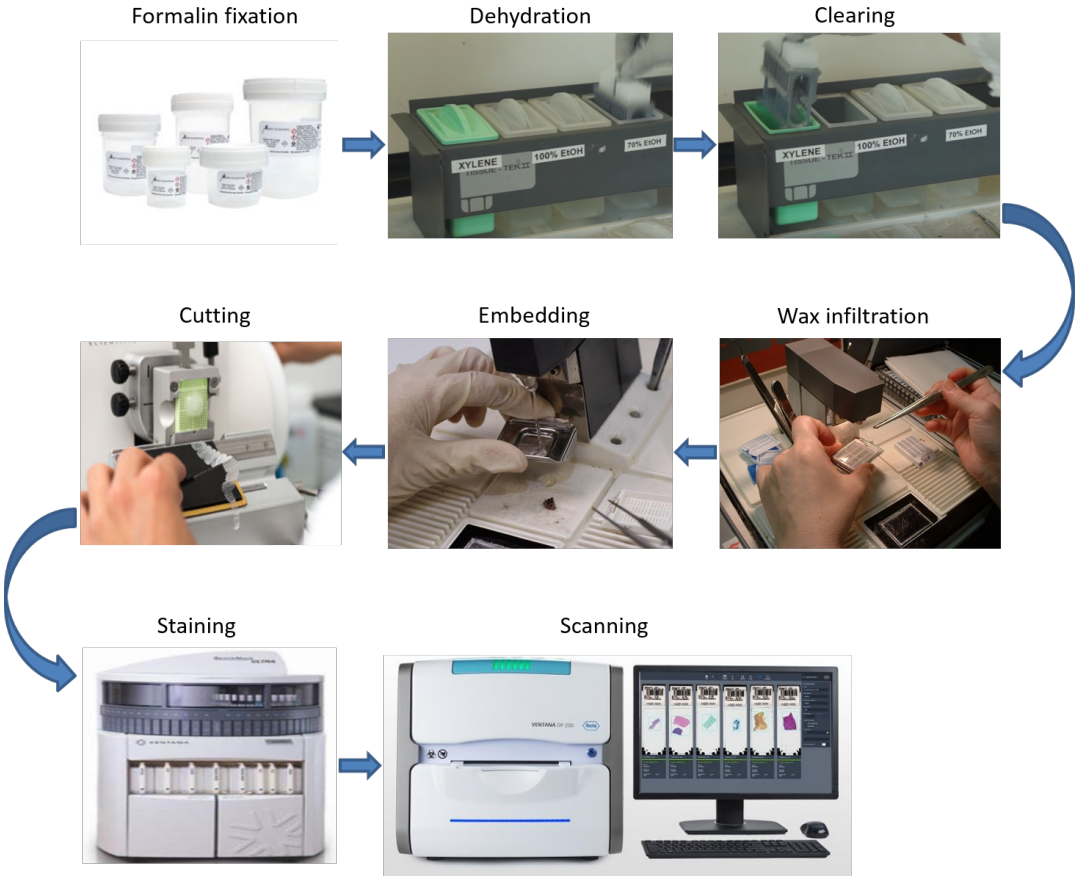


Fig. 2.1 Tissue preparation and general workflow in histology. Tissue preparation includes fixation (e.g. with Formalin), dehydration (e.g. with ethanol solutions), clearing (e.g. with Xylene), infiltration (e.g. with Parrafin), embedding, sectioning and staining in order to be ready for pathologist examination or scanning.

2.3 Applications

The success of Whole Slide Imaging and the advances in computing power allowed to enhance the research in the direction of image processing solutions for digital pathology. Over the past two decades, several applications have been developed and improved in order to assist the pathologists in this new era of digital pathology. The following list includes the most common digital pathology applications developed for clinical and research purposes.

- **Segmentation and detection:** Segmentation refers to the task of delineating specific objects (e.g. cell nuclei) [26, 27, 28, 29, 30] or regions (e.g. tumor) [31, 32] while detection aims to identify the center of the object of interest (e.g. mitosis detection) [33]. We detail the content of our scientific contribution about multi-stain tissue segmentation in part III of this dissertation.

- **Tissue classification, disease grading and outcome prediction:** In many cases, tissue architecture and morphological nuclei patterns are predictive of the histological subtype of the tissue and the grade and outcome of the disease [34, 35, 36]. Therefore, digital

pathology community has been interested in automating these classification tasks [37, 38, 39] using WSIs as inputs. For this, different machine learning approaches have been widely used and tested. These approaches are based either on interpretable handcrafted features biologically defined by clinicians and mathematically modeled by researchers and scientists or on extracted features learned by training a deep learning network or on a combination of both feature categories.

- **Biomarker quantification:** Using IHC stains allows to assess the presence or absence of specific targets and to evaluate the expression of the studied proteins [21]. However, this task is highly sensitive and visual interpretation of biomarker expression in IHC stainings can be subjective, poorly reproducible and prone to error. Several recent studies have shown significant variability rates between different observers in the evaluation of biomarker expressions [40, 41]. In [42] an error rate of 20% has been reported in the assessment of human epidermal growth factor receptor-2 (HER2) in invasive breast cancer. Slide digitization made it possible to automate the quantification of immunohistochemical stains (e.g. HER2, estrogen receptor (ER) and progesterone receptor (PR)) using image analysis based methods. When properly developed and validated, these methods would allow to quantitatively characterize the spatial distribution of the studied targets, standardize IHC scoring and increase the reproducibility and objectivity of the results. In many cases, the IHC staining includes two different biomarkers visualized with two colors in the digital image. In this case, color deconvolution algorithms are used in order to separate between the different stain channels and be able to quantify the expression of the different biomarkers in the IHC staining.

- **Registration:** In drug development and histological diagnosis, it is very common that multiple stain types are needed. As IHC staining and brightfield microscopy are limited in the number of simultaneous different separable stains that can be made on a single section due to antibody reactions and the nature of the color space, one of the possible solutions is to stain consecutive sections of the same patient block with the required stainings. In this case, the different stained sections need to be registered [43, 44] in order to bring them to the same coordinate system and help clinicians evaluate paired images and understand the colocalization of the different studied targets. This ability to combine information from different staining modalities and to study the behaviour of different biomarkers within the same tissue areas helps experts to gain more understanding about the tumor microenvironment and the immune response.

- **Stain virtualisation:** This application is relatively new to digital pathology and gained increased interest from the computer vision community with the success and progress of style transfer algorithms. It aims to simulate realistic, clinically meaningful and interpretable stained tissue images thus saving time, lab efforts and valuable tissue resources [45, 46]. Virtual staining could also generate in-silico multiplexing of different stains on the same tissue segment, enabling analysis of different markers in the same whole slide coordinate space and without the need to perform intermediate sensitive washing steps or deal with slide registration. We detail the content of our contributions about stain virtualization in digital pathology in part IV of this dissertation.

2.4 Challenges

Even though Whole Slide Imaging and image analysis tools are helping clinicians and researchers improve patient diagnosis and drug development by leveraging technology to perform tedious and time consuming tasks, many technical and computational challenges still need to be addressed and overcome in order to ensure the successful application of computer assisted analysis in clinical routines. The following list includes some of the most common and substantial challenges in digital pathology.

- **Digital slide storage and management:** While the success of Whole Slide Imaging as a solution to scan entire slides at high magnifications allowed the transition from conventional histology to digital pathology solutions, it has also resulted in enormous amounts of data which can reach a few terabytes per week in the case of large pathology practice [21]. Image analysis generates additional volumes of data that need to be stored together with the WSIs and their metadata. This large volume of rapidly increasing data can not be handled with normal workstations or small sized clusters. Therefore, ensuring scalable big data storage solutions is one of the first and major requirements needed for digital pathology. Additionally, these solutions need to meet high security and technical standards in order to ensure long term robust and reliable storage, to protect patient data and to ensure efficient loading and fast viewing and query response [21].

- **Stain color variations:** One of the main challenges in computerized analysis of digital slides is color variability between images of the same stain type [47]. This variability can originate from several factors including experiment conditions, different lab protocols, reagent variability, different staining machines and imaging pipeline variability (e.g. microscope scanner lighting and focus). In order to ensure high reproducibility and low error rates of digital pathology analysis results, image analysis algorithms need to be robust and resilient to these color variations.

- **Image artifacts:** In some cases, the staining quality is affected by different types of artifacts acquired during slide preparation or image acquisition [48]. Common artifacts include tissue folds, out of focus regions, shadows, pen marks and blood drops. In order to avoid errors and ensure the accuracy of computerized analysis results, these artifacts need to be detected and excluded from the analysis.

- **Planar histology images:** Most current WSI scanners generate 2D planar slide images without the depth information which is commonly available on most microscopes [47]. This could make some digital pathology tasks (e.g. mitosis detection) more challenging if the depth information is useful for the task.

Part II

Deep Learning and Digital Pathology

3 Learning

> *Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.*
>
> **— Stephen Hawking**
> (Famous theoretical physicist, cosmologist, and author)

The enormous advances in deep learning networks in recent years allowed to solve several complex problems that were very difficult or impossible to solve in the past. Deep neural networks are a subset of machine learning algorithms designed to learn from large amounts of data by inherently learning high level relevant features for the considered task using a cascade of layered and interconnected computational units. Training deep neural networks is based on backpropagation of the gradient which allows to gradually optimize the parameters of the network in order to minimize a loss function defined beforehand by the network's developer. Backpropagation has been introduced in 1986 [49] and is a smart and efficient way of calculating the partial derivatives of the loss with respect to the learnable parameters. After gradient calculation, the network's parameters are gradually updated in the direction of the optimal solution (i.e. minimizing the loss function) using a step size called learning rate. Currently, all training methods are based on different variants of gradient descent optimization algorithm [50] (e.g. stochastic gradient descent [51], momentum [52] and Adam [53]). Depending on the nature of the problem and the available data, different learning approaches have been proposed. These approaches include supervised, semi-supervised and unsupervised learning. In this thesis we used supervised learning for tissue compartment segmentation and unsupervised learning for stain virtualisation. This is why we introduce these two concepts in the first couple of sections of this chapter. We additionally present two important subsets of deep learning algorithms which are relevant to the content of this dissertation in the third and fourth sections.

3.1 Supervised Learning

Supervised learning is a sub-branch of machine learning algorithms where the algorithm is trained based on a training dataset consisting of inputs paired with their correct outputs known as ground truths [54]. This labeled training dataset helps the designed model optimize its parameters by means of minimizing a previously defined cost function (Fig. 3.1).

When the training phase is successful, the learned model becomes ready to predict outcomes and generalise well to new unseen data. Supervised learning has been used to solve several

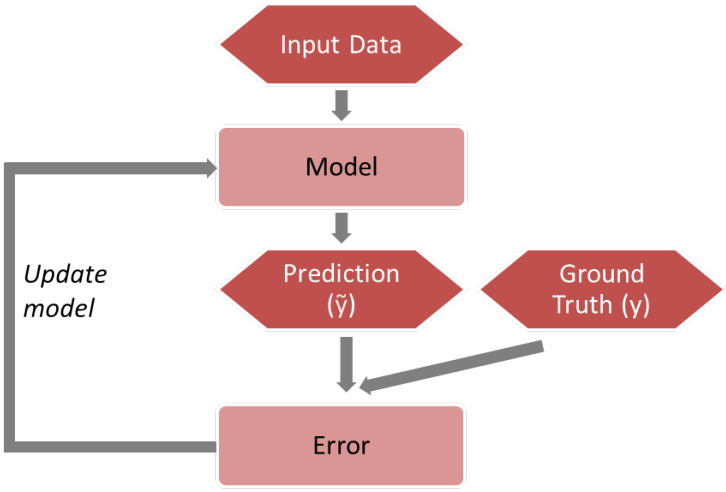


Fig. 3.1 Example of supervised learning.

types of problems including regression and classification. In the case of deep learning, large amounts of labeled training datasets are needed in order to train a good supervised learning based network.

In supervised learning we are given a dataset with N training samples (x_i, y_i) where x_i is the input and y_i is the ground truth label for each $i \in \{1, ..., N\}$. The objective is to learn a function $f_\theta(x) = \tilde{y}$ that minimizes the error on the training examples. In this case, θ groups all the learnable parameters, f is parametrized with θ and $\tilde{y}$ is the predicted output. Learning the function f translates then to finding an optimal θ^* minimizing a defined loss function $\mathcal{L}$:

$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i) \tag{3.1}$$

Depending on the problem, the requirements and the characteristics of the training dataset, different differentiable loss functions have been customized and used (e.g. Mean Squared Error Loss, Mean Absolute Error Loss and Cross-Entropy). For example, the Mean Squared Error Loss (also called ℓ_2 loss) is the default loss in case of regression problems and is a preferred and appropriate measure when the distribution of the target distribution is Gaussian. However, when this distribution has outliers, the Mean Absolute Error loss (also called ℓ_1 loss) can be more appropriate. Cross-Entropy is the default loss function used for binary and multi-class classification problems and it aims to fit the probability distribution of the predictions to the actual probability distribution. The cross entropy loss function is defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \log(\hat{p}_c^{(n)}), \tag{3.2}$$

where $\hat{p}_c^{(n)}$is the probability of sample n being in class c, $y_c^{(n)}$ is the label of sample n for class c when the label is given in one-hot encoding and C is the number of classes.

3.2 Unsupervised Learning

Unlike supervised learning, unsupervised learning methods are based on training the model based only on the training input data without any ground truth information (Fig. 3.2). In this category of methods, the model is supposed to extract relevant features and make sense of the data without having a "supervisor" continuously correcting the model [55, 56, 57].

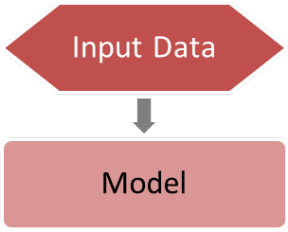


Fig. 3.2 Example of unsupervised learning.

In general supervised learning approaches give better performances. However, in many cases labeled data are difficult or expensive or impossible to get which makes the use of unsupervised learning based approaches necessary. Unsupervised learning has been used for instance in clustering and image generation problems. In the case of deep learning, autoencoders [58] are another example of unsupervised learning networks. Autoencoders are used to encode and compress a complex input into a compact feature representation by training a neural network to predict its own input. This compact representation can then be stored instead of the original input hence saving storage space or used for other tasks such as clustering or dimension reduction. Additionally, Generative Adversarial Networks (GANs) have been used for unsupervised deep learning image generation. More details about GANs will be given in the last section of this chapter.

3.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a subset of deep learning algorithms [59, 60, 61, 62, 63] based on the assumption that images are inherently structured. In the last decades CNNs have been widely used in several tasks such as semantic segmentation, classification and object detection. A CNN consists of stacked layers of different building blocks. In the following we briefly discuss the most important ones.

Convolutions Unlike regular neural networks where each layer is made up of neurons fully connected to all the neurons of the previous layer, CNNs have a different architecture based on using convolutions in order to exploit the structured nature of the images and significantly reduce the number of learnable parameters. A convolution consists in applying a filter (also called kernel) having a restricted number of weights (typically 3×3 pixels) to an input image

or the previous feature maps in order to produce a new feature map. For this a dot product between the filter and the input is computed then the filter is moved with a defined step (called stride) so that at the end the filter is applied to each filter-sized patch of the input image. The fact that the filters are spatially restricted allows to detect similar patterns in different locations of the image and hence to better understand the sophistication of the image. Convolutions are in general followed by non-linear activation functions [64] (e.g. ReLu [65]). Mathematically, the ReLu function is given by:

$$f(x) = \max(x, 0) \tag{3.3}$$

Pooling Pooling layers are used in-between convolution layers and allow to gradually reduce the spatial size of the feature maps [66] in order to reduce the computational power needed to process the data and to help extracting the dominant features. There are two types of pooling: Max Pooling and Average Pooling. Pooling layers downsample each feature map independently by returning the maximum or the average value within the pooling window size. Max Pooling contributes also to suppressing the noise as it helps to eliminate the noisy activations from the feature maps.

Batch Normalization When training a deep neural network the distribution of each layer's input changes which makes the training slower and very sensitive to parameter initialization and the learning rate. This phenomenon is called the covariate shift and it has been one of the serious challenges in the training of deep networks. During training batch normalization is used to normalize the activations by subtracting the batch mean and dividing by the batch standard deviation in order to increase the stability of the network [67]. During network's inference the learned statistics (mean and standard deviation) are used instead of the actual statistics.

Instance Normalization Instance normalization is very similar to batch normalization but it operates on the instance level instead of the batch level and is used in the same way during training and inference.This normalization has proven useful in image generation and style transfer tasks [68].

Dropout Overfitting the training data is one of the most frequent challenges of training deep networks. Dropout is a regularization method consisting in randomly dropping out neurons with a certain probability during training in order to reduce overfitting and increase generalization power [69]. This forces the neurons of each layer to contribute to the final output. Dropout is used only during training and turned off at testing time. Dropout layers are generally used towards the end of the network after the fully connected layers.

3.4 Generative Adversarial Networks

GANs are a subset of unsupervised deep learning methods based on simultaneously training two models: a generative model learning the distribution of input variables in order to

generate new synthetic samples in the input space and a discriminative model classifying the examples as real (from the training dataset) or fake (generated) [70, 71]. During training, the generator G and the discriminator D compete against each other to solve a minimax game: the generator learns to produce real-looking samples in order to fool the discriminator and the discriminator learns to discriminate between real and generated images (Fig. 3.3).

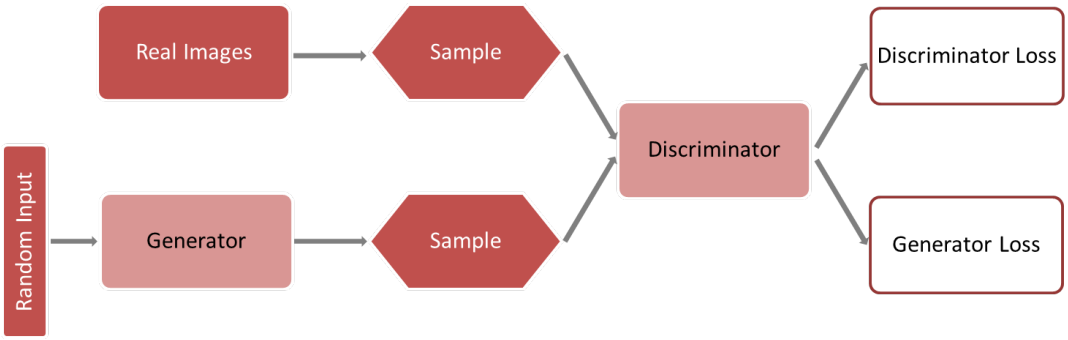


Fig. 3.3 GAN model. During training, the generator learns to produce real-looking samples in order to fool the discriminator and the discriminator learns to discriminate between real and generated images.

In the first GAN version, the following loss (called minimax loss) is being minimized by the generator and maximized by the discriminator:

$$\mathbb{E}_x[\log(D(x))] + \mathbb{E}_z[\log(1 - D(G(z)))], \tag{3.4}$$

where x is a real data sample, z is the noise input vector to the generator, $D(x)$ is the probability estimation of the discriminator for x to be real, $G(z)$ is the output of the generator and $D(G(z))$ is the probability estimation of the discriminator for $G(z)$ to be real. With this configuration, the discriminator is trained to maximize the minimax loss (Eq. 3.4) while the generator is trained to minimize $\log(1 - D(G(z)))$ (as it can not directly affect $\log(D(x))$).

GANs have been successfully used in several applications such as style transfer, image colorization and new data generation. However, there are several challenges related to GAN training that researchers have been trying to address. These challenges include:

- **Vanishing gradients:** When the real/fake discrimination task is very easy for the discriminator, the discriminator becomes quickly very good at its task and gradients become vanishingly small. In this case, the generator does not receive enough information to make progress. Some solutions have been already investigated to address this issue such as modifying the minimax loss such that the generator's objective is changed to maximizing $\mathbb{E}_z[\log(D(G(z)))]$. An alternative solution that has been proposed is using the Wasserstein loss [72] where the discriminator tries to maximize the difference between its outputs on real and fake instances instead of classifying them and the generator tries to maximise the discriminator's output on the fake instance.

- **Mode collapse:** One of the common issues in GAN training is that the generator manages to produce one plausible output (or a very small set of plausible outputs) that fools the discriminators then keeps generating this same output to all inputs. In the best

case scenario the discriminator learns to always classify this example as fake. However, it is possible that the discriminator gets stuck in a local minima and the generator keeps producing the same output. This issue is called mode collapse. Some solutions to mode collapse have been proposed such as Wasserstein loss and unrolled GANs [73] where the generator is updated based not only on the current discriminator's classification but on the outputs of future discriminator versions as well.

4 Deep Learning in Digital Pathology

With the significant advances in histology digital data generation and computational power, deep learning solutions have been increasingly applied and adapted to digital pathology applications including segmentation, cell or structure detection and classification, biomarker quantification, disease grading, tumor mutation or response to drug prediction and stain virtualization.

4.1 Challenges

Besides the general digital pathology challenges detailed in the previous chapter, the application of deep learning to digital pathology resulted in additional challenges that have been investigated and explored by researchers during the last couple of decades with different degrees of success. These deep learning specific challenges include:

- **Quantity and quality of annotated data:** Machine learning algorithms are based on the concept of learning from data. This means that the system implicitly learns to perform a specific task using a sample dataset without being explicitly programmed to do the task. While these algorithms allowed to solve very complex problems that are difficult or impossible to solve with modeling and explicit instructions, it has also risen a significant challenge that experts and scientists have to handle. Actually, the success of machine learning methods depends on the quantity and the quality of the sample data used to train the algorithm. For many tasks (e.g. segmentation, classification and output prediction), these data need to be well annotated so that the algorithm can successfully generalize to new and unseen data. Particularly in the case of deep learning networks, very large amounts of good quality annotated data have to be prepared and used in order to learn a good model [74, 75]. Collecting these high quality labeled data is one of the most challenging constraints in deep learning based applications especially in case of medical applications (e.g. digital pathology) where only expert clinicians are able to annotate the data.

- **Weak labeling:** Another common issue related to WSI labeling is that the labels can be assigned on the WSI level and not on the patch level [76, 77, 78]. This means that it is possible that the information about the presence or absence of a specific region of interest exists but the position of this region is not available. For example, when a WSI is labeled as malignant, this means that at least one of the patches contains malignant tissue. In this case, it can be that the slide contains healthy, benign and malignant tissue at the same time. These types of labels are called weak labels and are generally easier and cheaper to get than patch level labels. However, using them as patch level labels

with regular supervised learning approaches can result in confusing the network with noisy labels and hence in a poor performance.

- **Lack of interpretability:** Deep learning networks are very often described as "black boxes" and criticized for being poorly intuitive and hard to interpret. In medical applications (e.g. digital pathology), it is always very important for clinicians to understand how did the system get to its decision so that they can accept and trust the system. A lot of effort has been made in this direction in order to make deep learning networks easier to understand and interpret by humans [79, 80]. This includes visualizing the learned filters, highlighting the input pixels that were significant for the network's decision [81, 82], creating input data causing a high activation for a specific category [82, 83] and combining domain specific hand crafted features with deep learning features.

- **Class imbalance:** In medical applications such as digital pathology, it is common that the data is not equally balanced between the different studied classes [76]. For example in case of tumor subtype classification it can happen that one subtype is much more abundant than the other. Using deep learning networks with imbalanced training datasets can result in a network biased towards predicting the majority class and poorly recognizing the minority class. This issue can be very dangerous in clinical scenarios and the problem should be carefully addressed before deep learning based systems could be safely used for diagnosis. Many solutions to class imbalance have been already proposed such as smart sampling techniques [84, 85, 86, 87] and loss weighting strategies [88, 89, 90, 91].

- **Image size and tiling artifact:** WSIs are typically very large and can contain billions of pixels at high resolution. In the case of deep learning, tile based analysis is usually used in order to overcome the memory bound hardware limitations. In some specific applications, this tilewise processing can result in a tiling artifact after reconstruction of WSI outputs from the output tiles [1, 2, 3].

- **Choice of the appropriate magnification:** Digital histology images are generally stored in a multi-resolution pyramid structures providing distinct zoom levels. Choosing an appropriate magnification appropriate to the task in hand is a delicate task significantly affecting the performance of image analysis algorithms [92]. In the case of deep learning networks, there is a trade-off between the fine details available in high magnification and the image context (called receptive field) seen by the network before deciding on its output. If very high magnification is used, it is very likely that the network will not be able to see enough context in the receptive field area especially in the case of high level tasks such as epithelium segmentation. If very low magnification is used, it is possible that only a limited number of effective pixels in the receptive field area is relevant to the task in hand. This results in using memory resources inefficiently and sometimes in confusing the network with non relevant information. Theoretically, networks with large receptive fields at high magnification would result in a better performance. However, this comes with additional costs in terms of training and inference time and memory usage.

4.2 Common Practices

Depending on the task in hand and the problem's constraints, several practices have been commonly used by researchers and scientists as an attempt to address the challenges related to applying deep learning to digital pathology. These great efforts allowed to improve the performance of deep learning approaches, but there is still room for improvement. Some of the most common practices are:

- **Preprocessing:** Different preprocessing techniques have been considered in digital pathology applications. For example, several state of the art methods used **stain normalization** solutions [93, 94, 95, 96, 97] as a preprocessing step in order to deal with stain color variation and improve the generalisation of machine learning algorithms. Different stain normalization techniques have been developed such as color matching approaches based on the color distribution of a reference template image picked by an expert, stain separation approaches based on using a reference image and normalizing each channel independently, and learning approaches based on style transfer methods. Additionally, **color deconvolution** has been used in many tasks in order to separate between the different stains in the image. For example, as Hematoxylin is a stain for nuclei, it is common to separate between Hematoxylin and Eosin in H&E images and use only the hematoxylin channel in case of cell detection or cell segmentation tasks [98]. Moreover, several methods have been proposed to **detect and eliminate artifacts** (e.g. blur detection [99] and tissue fold detection [100]) in order to avoid negatively affecting image analysis performance. These methods could be used as preprocessing steps before performing any further analysis.

- **Efficient labeling:** Several solutions have been investigated in order to efficiently increase the number of labeled data. For example, active learning allows to automatically select the most useful and valuable unlabeled samples based on some criteria (e.g. uncertainty sampling [101], query by committee [102] and variance reduction [103]) in order to alleviate the annotation burden for pathologists and at the same time reach a good performance. A second efficient labeling technique that has been tested by some groups is labeling data based on pathologists' diagnosis behavior such as tracking their eye movements [104] and their mouse position [105].

- **Transfer learning:** As it is difficult and not always possible to get enough annotated data to train a good deep learning network, transfer learning has been used in digital pathology in order to transfer the knowledge from networks properly trained on other datasets (e.g. natural image datasets). The technique consists of using the learned parameters of a pre-trained network optimized for another similar task and fine-tuning them for the specific digital pathology task using the small set of annotated training data [106]. The idea of using transfer learning comes from the fact that natural images and histology images, despite their significant differences, share basic image features such as lines and arcs.

- **Data augmentation:** Another technique used to deal with stain color variation and the difficulty of getting sufficient well annotated data is the use of data augmentation

during the training phase. This technique allows to artificially increase the number of training samples and to make the network more robust to noise and color variations [107, 108, 109, 110]. Data augmentation consists in bringing a single or a combination of modifications to the original image such as hue, brightness, and scale jittering, adding uniform noise to the intensity channels, image flips, and random affine transforms.

- **Multiple instance learning:** In order to address the issue of weak labels which is very common in high resolution WSI analysis, many groups investigated multiple instance learning solutions based on considering each WSI as a bag of patches called instances [111, 112, 113, 114]. In this case, a positive bag contains at least one positive instance and a negative bag does not contain any positive instance

- **Multiscale learning:** combining results from different magnifications has been successfully used in several digital pathology applications in order to aggregate different levels of information [115, 116, 117]. For instance, multiscale learning can be useful in the case of tumor segmentation since the cancerous tissue is heterogeneous and can contain small cells and much larger structures at the same time. This technique has been considered as an approximation of pathologists' behavior who progressively integrate features from multiple scales to reach their pathological decision (Fig. 4.1).

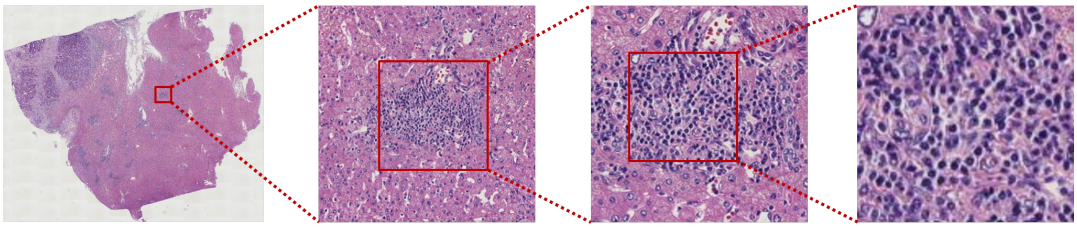


Fig. 4.1 Example of an H&E stained image at different zoom levels.

4.3 Validation and Evaluation

In order to be able to adopt machine learning solutions in real clinical scenarios and greatly benefit from the recent advances of deep learning networks, it is mandatory to thoroughly and intensively validate and standardize these algorithms. Increasing generalisation capabilities and ensuring reproducibility of algorithms is extremely important especially in the case of healthcare systems where patients' safety is the highest priority. The transition from academic research to clinical practice requires algorithms to be validated on very large patient cohorts ideally coming from multiple laboratories to ensure a broad applicability of the methods. Significant efforts have been made in order to allow the adoption and effective use of machine learning solutions by quantitatively and pathologically validating the new approaches, but the goal is still far from being completely reached.

One way of validating computer based analysis is to quantify the similarity between the results of the tested algorithm and the ground truth results of expert pathologists on a specific task using new samples unseen by the algorithm at the training phase. The choice of the task should be coherent with what the algorithm will be used for after validation. This validation

approach is very commonly used in segmentation and classification problems. Several metrics have been used to evaluate the performance of this type of algorithms. Most metrics are based on confusion matrices.

Tab. 4.1 Confusion matrix.

	Ground truth	
Prediction	Positive	Negative
Positive	TP	FP
Negative	FN	TN

In Table 4.1, TP, FP, FN and TN refer to true positives, false positives, false negatives and true negatives respectively. In the following are some common metrics used for this type of validation.

- **Accuracy:** It measures the percentage of predictions that the model correctly labeled. Its formula can be written as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

  Accuracy gives a very general idea about the performance of the model but it is not considered as a reliable indicator in case of class imbalanced datasets.

- **Precision:** It measures the percentage of positive identifications that were actually correct. Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

- **Recall:** It measures the percentage of actual positives that were correctly identified as positives. Recall (also called True Positive Rate TPR) is defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

- **Specificity:** It measures the percentage of actual negatives that were correctly identified as negatives. Specificity is defined as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{4.4}$$

- $F1$ **score (or Dice coefficient):** It corresponds to the harmonic mean of precision and recall. $F1$ score better evaluates the performance of the model than accuracy, particularly in the case of class imbalanced datasets. It is defined as:

$$F1 = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \tag{4.5}$$

- **Area Under the Curve (AUC):** A Receiver Operating Characteristics (ROC) curve is a graph showing the performance of a model at different classification thresholds. The graph plots the model's TPR (recall) against its False Positive Rate (FPR) defined as $1 - Specificity$. AUC measures the entire area under the ROC curve and provides an aggregate measure of performance across the possible thresholds. The higher the AUC, the better is the overall performance.

In other applications such as stain virtualization and color normalization, the validation can be based on assessing the quality of the generated images for example by assessing its similarity to ground truth images. In this case, different metrics have been used. In the following y and $\hat{y}$ correspond to an original and a synthesised image respectively, n corresponds to the number of pixels, y_i and $\hat{y}_i$ correspond to the i^{th} pixels in y and $\hat{y}$ respectively.

- **Mean Squared Error (MSE):** It refers to the average of the squared intensity differences of both compared images. MSE is defined as follows:

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.6}$$

- **Peak Signal to Noise Ratio (PSNR):** It represents the ratio between the maximum possible power of a signal and the power of the distorting noise. PSNR is expressed in the logarithmic Decibel scale and can be defined as:

$$PSNR(y, \hat{y}) = 10\log_{10}(\frac{y_{max}^2}{MSE}), \tag{4.7}$$

  where y_{max} is the maximum value in the original image.

- **Pearson Correlation Coefficient:** This metric is used to measure the degree of linear correlation between two images. It is defined as:

$$r(y, \hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{4.8}$$

  where $\bar{y}$ and $\bar{\hat{y}}$ are the mean values of y and $\hat{y}$ respectively.

- **Structural SIMilarity index (SSIM):** This index is a perceptual measure used to assess the quality of a processed image compared to an original image by measuring the structural information change based on human visual system characteristics [118]. It is defined as:

$$SSIM(y, \hat{y}) = \frac{(2\mu_y\mu_{\hat{y}} + c_1)(2\sigma_{y\hat{y}} + c_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + c_1)(\sigma_y^2 + \sigma_{\hat{y}}^2 + c_2)}, \tag{4.9}$$

  where μ_y and $\mu_{\hat{y}}$ are the averages of y and $\hat{y}$ respectively, σ_y and $\sigma_{\hat{y}}$ are the variances of y and $\hat{y}$ respectively, $\sigma_{y\hat{y}}$ is the covariance of y and $\hat{y}$ and c_1 and c_2 are small positive constants allowing to stabilize the division.

- **Complex Wavelet Structural SIMilarity index (CWSSIM):** This metric is an extension of the Structural SIMilarity index (SSIM) to the complex wavelet domain characterized by its robustness to small translations and rotations. As explained in [119], small image distortions result in consistent phase changes in the local wavelet coefficients which does not affect the structural content of the image. It is defined as:

$$CWSSIM(c_y, c_{\hat{y}}) = \frac{2\sum_{i=1}^{n} |c_{y,i}||c_{\hat{y},i}| + K}{\sum_{i=1}^{n} |c_{y,i}|^2 + \sum_{i=1}^{n} |c_{\hat{y},i}|^2 + K} \times \frac{2|\sum_{i=1}^{n} c_{y,i}c^*_{\hat{y},i}| + K}{2\sum_{i=1}^{n} |c_{y,i}c^*_{\hat{y},i}| + K}, \quad (4.10)$$

where $c_y = \{c_{y,i}|i = \{i = 1, ..., n\}\}$ and $c_{\hat{y}} = \{c_{\hat{y},i}|i = \{i = 1, ..., n\}\}$ are two sets of coefficients in the complex wavelet transform domain extracted at the same spatial location in the same wavelet subbands of y and $\hat{y}$ respectively. c^* denotes the complex conjugate of c and K is a small positive constant.

Part III

Generalization of Multistain Immunohistochemistry Tissue Segmentation

5 Introduction

> *The question of whether a computer can think is no more interesting than the question of whether a submarine can swim.*
>
> **— Edsger Dijkstra**
> (Mechatronics Volume 2: Concepts in Artificial Intelligence)

Image segmentation is one of the key, common and critical steps in medical applications. Medical diagnosis, predictive analysis and quantitative assessment rely heavily on structure and lesion segmentation.

5.1 Motivation

A key challenge in cancer immunotherapy biomarker research is quantification of pattern changes in microscopic WSIs of tumor biopsies and surgical specimens. Different cell types tend to migrate into various tissue compartments and form variable distribution patterns. Drug development requires correlative analysis of various biomarkers in and between the tissue compartments. Specifically in the context of immunotherapy drugs development, the samples are very variable and the slides are typically stained in at least 3 different IHC staining methods with different stain types and colors in addition to the traditional H&E. To enable that, tissue slides are manually annotated by expert pathologists. Manual annotation of tissue slides is a labor intensive, tedious and error-prone task. Additionally, with the tools existing today it is also limited in precision and can be poorly consistent between different experts. Previous studies have shown that pathologist concordance in the manual analysis of slides suffers from a relatively high inter and intra user variability [120, 121, 122] mainly due to the large size of the images. Automation of this annotation process can improve accuracy and consistency while reducing workload and cost in a way that will positively influence drug development research and will assist pathologists in the time consuming operation of manual segmentation.

In this part we develop a supervised deep learning based method to automatically segment and annotate digitized slide images with multiple stainings into compartments of tumor, healthy tissue, and necrosis. We address the task in the context of drug development where multiple stains, tissue types and tumor types exist and look into solutions for generalizations over these image populations. We additionally propose solutions to address the dataset high variability and class imbalance and to gain more trust from expert pathologists.

The content of this part is based on the following publication:

Amal Lahiani, Jacob Gildenblat, Irina Klaman, Nassir Navab, and Eldad Klaiman. "Generalising multistain immunohistochemistry tissue segmentation using end-to-end colour deconvolution deep neural networks". *IET Image Processing, 2019)*.

5.2 Related Work

With the new advances in computer vision and image analysis methods, automatic segmentation has gained more popularity and increased interest in the medical field communities.

5.2.1 Segmentation in Medical Imaging

In recent years researchers have intensively investigated the potential of using computational medical image segmentation based on machine learning techniques for several medical modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), X-ray, ultrasound and mammography. For example, multi-dimensional Recurrent Neural Networks (RNNs) have been adapted to brain segmentation in order to account for the spatio-temporal information [123] and deep CNNs with multi-modality MR images have been used in order to segment isointense infant brain images [124]. Additionally, the challenge of multiple sclerosis lesion segmentation in MRI images has been addressed with different approaches such as applying longitudinal multi-view CNNs to multiple MR modalities [125] and using 3D convolutional encoder networks with shortcut connections [126]. In the context of brain tumor segmentation, Havaei et al. proposed a multi-modality model handling missing imaging modalities [127] based on training a CNN to learn a separate embedding for each modality input. Specific moments of the available modalities are then computed and further processed to predict the final segmentation. Kamnitsas et al. proposed an efficient multi-scale 3D CNN for automatic brain lesion segmentation aiming to alleviate the computational burden of 3D image processing and to handle class imbalance [128]. They additionally used a 3D fully connected Conditional Random Field in order to refine the network's segmentation output. Brain tumor segmentation has been also addressed using an enhanced CNN with loss function optimization by BAT algorithm [129] and a multi-scale CNN with various MR modalities [130]. Moreover, deep learning solutions have been developed to segment neuronal membranes in electron microscopy images [131, 132], pancreas in CT images [133] and knee cartilage segmentation in MRI scans [134].

5.2.2 Segmentation in Digital Pathology

In the context of digital pathology, recent studies have shown significant advancement and increased interest in applying machine learning methods for segmentation, classification and detection. For example, great efforts have been done in breast cancer diagnosis in order to detect and localise breast cancer metastases in WSIs using careful image patch sampling, data augmentations and the inception architecture [76]. Janowczyk et al. used a resolution adaptive deep hierarchical learning scheme for nucleus segmentation in breast cancer in

order to reduce the computation time without affecting the performance [135]. Additionally, Xing et al. proposed a method for automatic nucleus segmentation in brain, pancreatic neuroendocrine and breast cancer using a deep CNN model followed by an iterative region merging [136]. Then a segmentation algorithm, including a robust selection-based sparse model and a local repulsive deformable model, has been used in order to separate between the individual nuclei. Song et al. used multiscale CNNs for unsupervised segmentation of overlapping cervical cell cytoplasm in pap smear images [137]. In the context of colon glands segmentation [138], different deep learning based methods have been developed and used such as introducing task-specific losses [139, 140], designing contour aware multi-task networks [141, 142], combining hand-crafted and deep learning features [142] and using pre-processed versions of the images as inputs to the CNNs [143, 144]. Segmentation of epithelial and stromal regions in breast and colorectal cancer has been addressed using a deep CNN feature learning method [145]. A 2D spatial clockwork RNN has been used to encode global context information into the features of each local image patch and improve the performance of perimysium segmentation in digitized muscle microscopy images [146]. Wang et al. used transfer learning and conditional random field jointly trained with a multiscale fully convolutional network in order to segment messy and muscle regions in histopathological images [147]. Apou et al. combined the outputs of a CNN and a texture classification system to detect lobular structures in breast [148]. Deep learning methods have also been applied for cancer and metastasis detection in histopathology images [149, 150, 151, 152]. Most of these methods were applied to the traditional H&E stained images. Some methods were additionally applied to one IHC staining [136, 145, 148] but the training was done separately for the different modalities.

6 Methodology

We present a method to generalize multistain tissue segmentation based on adapting CNNs to address the challenges we faced during the design of the solution including dataset high variability, limited annotated data and class imbalance. The main objective consists in automatically segmenting H&E and multiple IHC WSIs into compartments of background, healthy tissue, tumor and necrosis.

6.1 Problem Description

When multiple stainings are available, it is preferred to train a single model that will perform well on all stainings rather than multiple models (i.e. for each specific stain). Actually, a combined dataset with multiple stainings has the advantage of making the training dataset larger and more comprehensive thus handling the limitations in terms of available annotated samples and potentially also reducing overfitting to the training dataset. However, it is possible that training a CNN with multiple stainings in the training dataset makes it harder to learn the correct features of the different classes due to increased dataset and intraclass variability. For example, tumor can have different predictive features in H&E and IHC images making the task more complex and the generalization power more challenging to reach. Reducing input color variability has already been explored in digital pathology applications such as segmentation and classification. For instance stain normalization has been used as a pre-processing step [153, 154, 155] in order to standardize the colors of images of the same stain type but obtained with different conditions or different scanners and hence improve the performance on the performed task. Another way of limiting the detrimental effect of stain variability on training which has been described and tested in previous studies is using the concentration of stains as network inputs instead of the RGB pixel values. Color deconvolution has been used on histology images as a pre-processing step showing improved segmentation and classification results [98, 144, 156, 157]. However, stain deconvolution is generally subjective and depends on the stain bases. In the context of drug development, staining procedures, colors, and quality may vary for different reasons such as lab processes, imaging scanners, and staining protocols. Duggal et al. proposed to add a stain separation layer to the network as part of a binary cell classification framework [158], however a pre-processing step to estimate the optical density from the raw image was used. In addition, the method is heavily dependent on the parameter initialization of the stain separation layer. This filter initialization was based on stain parameters making it subjective and limited to the use of a single staining. In our case the training dataset comprises multiple stainings making it difficult to define stain specific parameters.

Another challenge that we had to address in order to ensure a good performance is the class imbalance. Actually, different biopsies contain variable amounts of background, tumor, tissue,

and necrosis. Thus, the generated dataset contained an inherent imbalance of the classes. Such class imbalances can significantly hinder the learning process in deep learning.

6.2 Network Architecture and Solutions

In order to address the dataset variability challenge, we propose to add an inherent color deconvolution segment to a state of the art segmentation network (i.e. UNET fully convolutional network [159]).

UNET contains two paths and has an encoder-decoder architecture. The first path (called encoder) allows to capture the context of the input image by gradually reducing the image size and increasing the depth (number of feature maps). The second path (called decoder) enables the precise localisation by gradually increasing the image size and decreasing the depth. The particularity of UNET architecture is that the coarse contextual information resulting from the encoder part is transferred to the decoder part by means of skip connections (Fig. 6.1). This combination of contextual and localisation information allows to get better precise locations and to significantly improve the performance of the network.

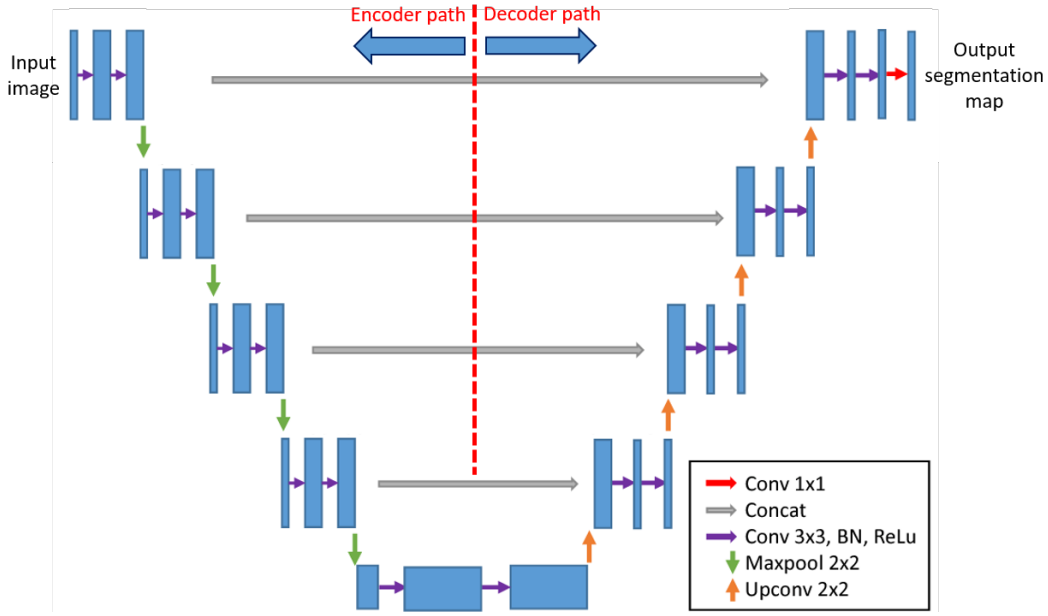


Fig. 6.1 UNET architecture. The encoder (the contraction path) allows to capture the context while the decoder (expansion path) enables the precise localisation. The skip connections allow to transfer simple features from early to later layers in order to recover the spatial information lost during the downsampling layers.

In our workflow, we slightly modified the original UNET architecture as follows. First, we used an appropriate size of zero padding in all convolution layers to preserve the spatial size of the input to the layers. The result is that both the input and the output spatial dimensions of the network are the same. We also considered a smaller network width: each layer has half the number of filters compared to the original UNET network. This helped speeding up the learning process and reduce overfitting. In addition, batch normalization [67] was

applied after every convolutional layer. Dropout layers [160] were also added at the end of the encoder and the decoder parts of the UNET in order to reduce overfitting.

In our approach, the color deconvolution segment parameters are learned as part of the segmentation network without the need to preprocess the images, making it ideal in the case of training on a multistain dataset. We then eliminate the need to run a separate preprocessing color deconvolution step for each different stain type and make both training and inference simpler. The proposed architecture has two additional (1×1) convolution layers preceding the modified UNET segmentation network. As the training dataset is composed of different stainings, the number of principle color shades in the training images is 6: pink, blue, purple, brown, yellow, and red. We chose then the first layer to have 6 (1×1) filters, each filter corresponding to a color. The second layer contains 3 (1×1) filters so as not to change the UNET architecture input size. Many of state of the art stain separation methods are based on Beer-Lambert's law using the optical density space [161, 162, 163]. According to this law, the optical density is defined as $OD = \log_{10}(\frac{I_0}{I_t}) = \varepsilon LC$ where OD is the optical density, I_0 is the intensity of the incident light, I_t is the intensity of the light after passing through the specimen, ε is the absorption coefficient, C is the concentration of the absorbing substance, and L is the thickness of the specimen. This law states then that the observed pixel intensity I_t varies nonlinearly with the concentration of the stain. In order to allow the color deconvolution segment to learn this nonlinearity of the physical model, each of the (1×1) convolution layers is followed by a nonlinear function.

The proposed color deconvolution network architecture (CD-UNET) is composed of 2 main parts (Fig. 6.2). The first part is a color deconvolution segment composed of 2 layers of (1×1) convolution with ReLU and batch normalization. The second part is the modified UNET fully convolutional network [159] resulting in a pixel wise segmentation of the input image.

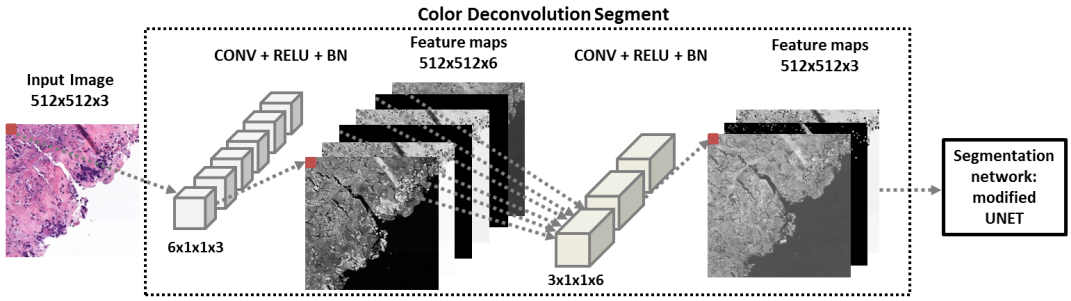


Fig. 6.2 The proposed CD-UNET architecture. The proposed color deconvolution segment is composed of 2 layers of convolutions. The first layer has 6 $(1 \times 1 \times 3)$ filters whereas the second layer has 3 $(1 \times 1 \times 6)$ filters both followed by RELU and batch normalization.

Additionally, in order to increase the amount of labeled data for training the network, we used data augmentation techniques including hue, brightness, scale jittering, adding uniform noise to the intensity channel, image flips, and random but small affine transformations.

In order to address the high class imbalance challenge, we use the Cross-Entropy loss weighted with median frequency balancing [164]. Median frequency balancing is a loss weighting strategy weighting each class loss by the ratio between the median of all class frequencies computed on the entire training set and the actual class frequency. This means that the weight

of the minority classes is higher than the dominant classes making the learning much less biased in favor of the dominant classes. The weighted Cross-Entropy is defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \log(\hat{p}_c^{(n)}) \omega_c, \tag{6.1}$$

where N is the number of samples, $\hat{p}_c^{(n)}$ is the probability of sample n being in class c, $y_c^{(n)}$ is the label of sample n for class c when the label is given in one-hot encoding and C is the number of classes. ω_c is the class weight for class c defined as:

$$\omega_c = \frac{median(\{f_c | c \in C\})}{f_c}, \tag{6.2}$$

where f_c corresponds to the frequency of class c.

7 Application

7.1 Dataset Description

Our dataset contains 77 WSIs of colorectal carcinoma metastases in liver tissue from biopsy slides stained with H&E (blue, pink) and 8 additional IHC assays stains as follows: CEA (brown), CD163/CD68 (brown, red), CD8/CD3 (brown,red), FoxP3 (brown), Ki67/CD3 (brown, red), Ki67/CD8 (purple, yellow), PRF/CD3 (brown, red), PD1 (brown). All these IHC stainings use a blue (Hematoxylin) counterstain. The selection was done according to the most frequent stainings arriving in the system with the idea to have as much data as possible as well the highest possible impact on the routine operational pipeline by automating the annotations for the majority of slides.

We split this dataset into training (51 slides) and testing (26 slides) sets. During training, 10% of the training data was reserved for validation. All the slides were chosen following a review of image quality, tissue quality, staining quality and region annotations by an expert pathologist. The various regions on the slides were annotated with one of the following categories: "Tissue" - i.e. normal tissue, "Tumor", "Necrosis" - i.e. dead tissue, and "Exclude" - i.e. areas not to be used due to irrelevance, artifacts, etc. Areas not included in any of the above categories are labeled "Background."

Each high resolution WSI was split into overlapping $512 \times 512 \times 3$ RGB tiles for processing at a 10x magnification factor (half of the original scanning resolution). The selection of a lower resolution was needed to increase the contextual information when classifying a given tissue pixel. The memory bound on the computing hardware limits the size of the input images to the network and therefore, in order to have enough tissue context for a limited tile size in the image dataset, we opted for the reduced magnification. Other studies have also found 10x magnification to be sufficient for tissue segmentation tasks [76, 152]. For each image tile a corresponding ground truth tile was created using the region annotations of the pathologist. This process yielded 16834 $512 \times 512 \times 3$ RGB tiles and their corresponding $512 \times 512 \times 1$ ground truth tiles. A sample of these tiles can be seen in Fig. 7.1.

The generated training dataset contains variable amounts of background, tumor, tissue, and necrosis. Table 7.1 shows the imbalance in the number of pixels belonging to each category.

Tab. 7.1 Percentage of pixels from each category.

Background (%)	Tumor (%)	Tissue (%)	Necrosis (%)
58	19	16	7

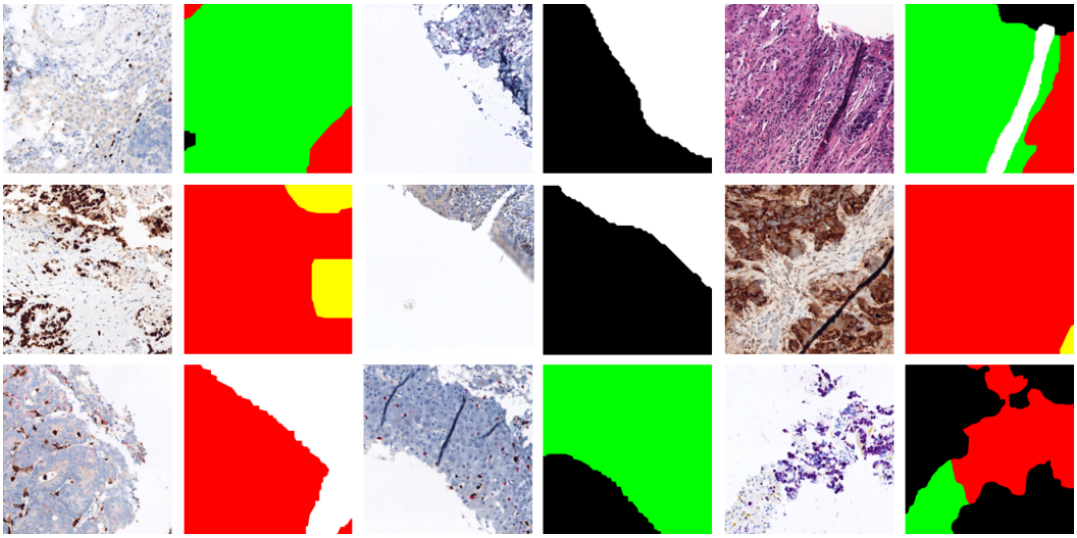


Fig. 7.1 Example tiles from the dataset and their respective label images colored green, red, yellow, black, and white corresponding to the different tissue regions of tissue, tumor, necrosis, background and exclude respectively.

7.2 Experiments and Observations

All network training was done using multiple GPUs of the Roche Pharma High Performance Clusters (HPC) in Penzberg. The evaluation of models was done based on $F1$ scores in all experiments in order to account for precision and recall.

7.2.1 Training with a Single Staining versus Multiple Stainings

In the development of the network architecture we initially examined the results of the modified UNET. We trained the modified UNET with 2 different datasets: The first dataset contains exclusively one specific staining (H&E) and the second dataset contains different staining types (H&E and 8 IHCs). We noticed that validation $F1$ scores of the different classes converge smoothly during training after 200 epochs in the case of the single stain dataset whereas they did not converge smoothly or to the same levels during the same training period in the case of the multiple staining dataset. This difference is most strongly expressed for the classes of necrosis and background. Fig. 7.2 shows validation $F1$ scores for the different classes during the 200 epochs of training with both datasets. The validation and the calculation of the scores were performed every 10 epochs.

7.2.2 Multiple Staining Training with End-to-End Color Deconvolution Network

We trained the modified UNET and CD-UNET for 200 epochs on the same multistain dataset and computed $F1$ scores for the different classes on the validation set every 10 epochs. We

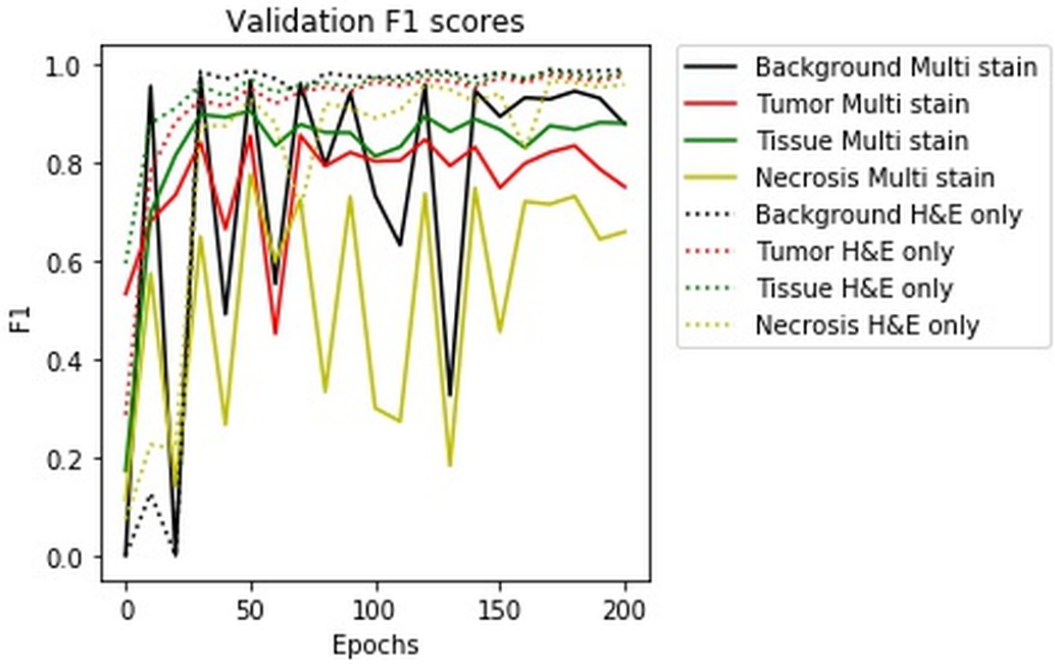


Fig. 7.2 Evolution of validation F1 scores during training. The black, red, green, and yellow curves correspond to background, tumor, tissue, and necrosis respectively. The dotted and solid lines correspond to validation scores of the modified UNET during training for a single stain (H&E) and a multistain dataset respectively.

noticed that the validation $F1$ scores converge faster and more smoothly on this dataset for the CD-UNET architecture compared to the modified UNET architecture (Fig. 7.3).

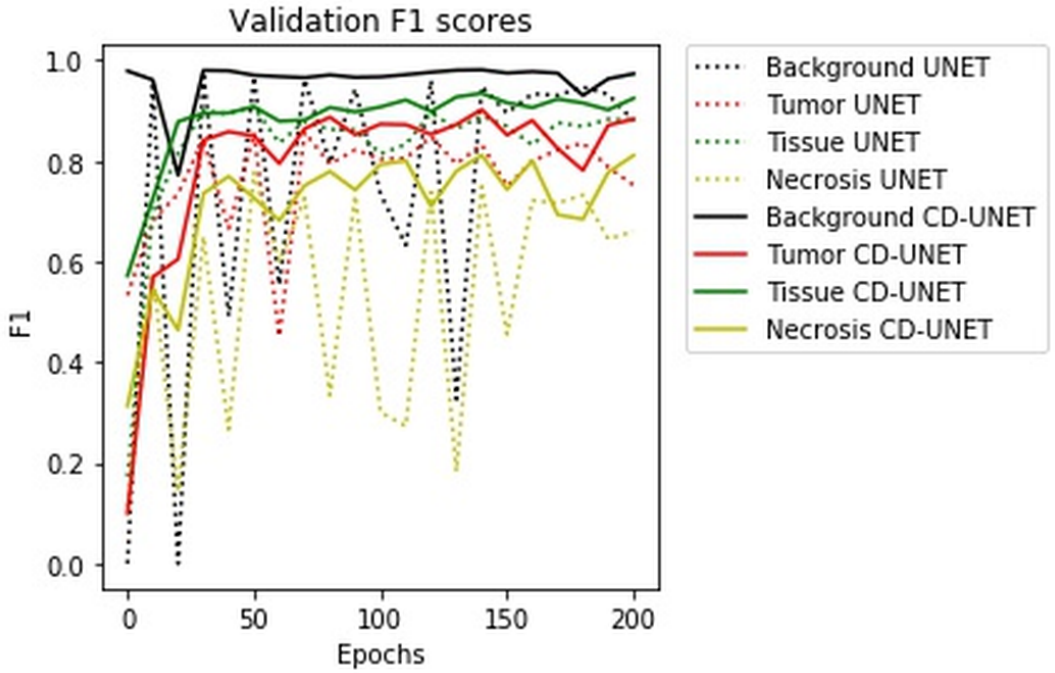


Fig. 7.3 Validation $F1$ scores for the different categories during 200 epochs of training. The dotted and solid lines correspond to the modified UNET and CD-UNET respectively.

7.3 Results and Comparison

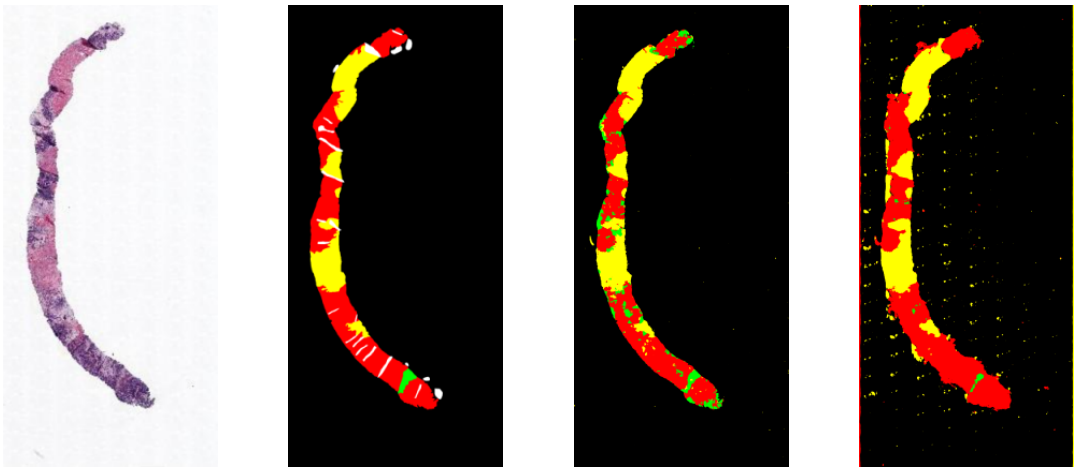
7.3.1 Single Staining versus Multiple Stainings

After 200 epochs of training the modified UNET, we evaluate the networks' performance on the testing dataset. The evaluation shows increased generalization capabilities for the single stain dataset compared to the multistain dataset (Table 7.2). For a fair comparison, we evaluated both networks on exactly the same testing dataset using only H&E images from the testing set.

Tab. 7.2 Testing $F1$ scores for each of the categories: Multiple stains vs H&E only.

	Background	Tumor	Tissue	Necrosis
Multistain	0.63	0.20	**0.78**	0.44
H&E only	**0.99**	**0.85**	0.71	**0.88**

Fig. 7.4 shows an example of an H&E image from the testing set segmented with both networks.



(a) (b) (c) (d)

Fig. 7.4 Example of a slide from the testing set segmented after training the modified UNET with a single stain dataset (H&E) and a multistain dataset. (a) and (b) correspond to the original H&E image and the ground truth respectively. (c) and (d) correspond to the segmentation output of UNET trained with a single stain dataset and a multistain dataset respectively.

7.3.2 Multiple Staining Color Deconvolution Network

After 200 epochs of training, we evaluated the performance of the end-to-end color deconvolution segmentation network on the testing WSIs. In this case, the testing dataset consists of slides from the stain types that were present in the training dataset. We also compare our results to previous state of the art segmentation networks (UNET, SegNet [165], ResNet-18 and ResNet-34 [166]). To ensure a fair comparison, we used the same training and testing

datasets for all networks. The $F1$ scores of the test set are listed in Table 7.3. $F1$ scores on the testing set were generally higher with CD-UNET reflecting better generalization capabilities over multistain images.

Tab. 7.3 Testing F1 scores.

	Background	Tumor	Tissue	Necrosis	Average
UNET	0.92	0.52	0.89	0.60	0.73
SegNet	0.98	0.87	**0.91**	0.57	0.83
ResNet-18	**0.99**	0.86	0.81	0.74	0.85
ResNet-34	**0.99**	0.85	0.86	0.75	0.86
CD-UNET	**0.99**	**0.88**	0.90	**0.80**	**0.89**

Fig. 7.5 shows some example results of the trained segmentation network on 3 different IHC stained slides (CD163/CD68, CD8/Ki67 and Ki67/CD3).

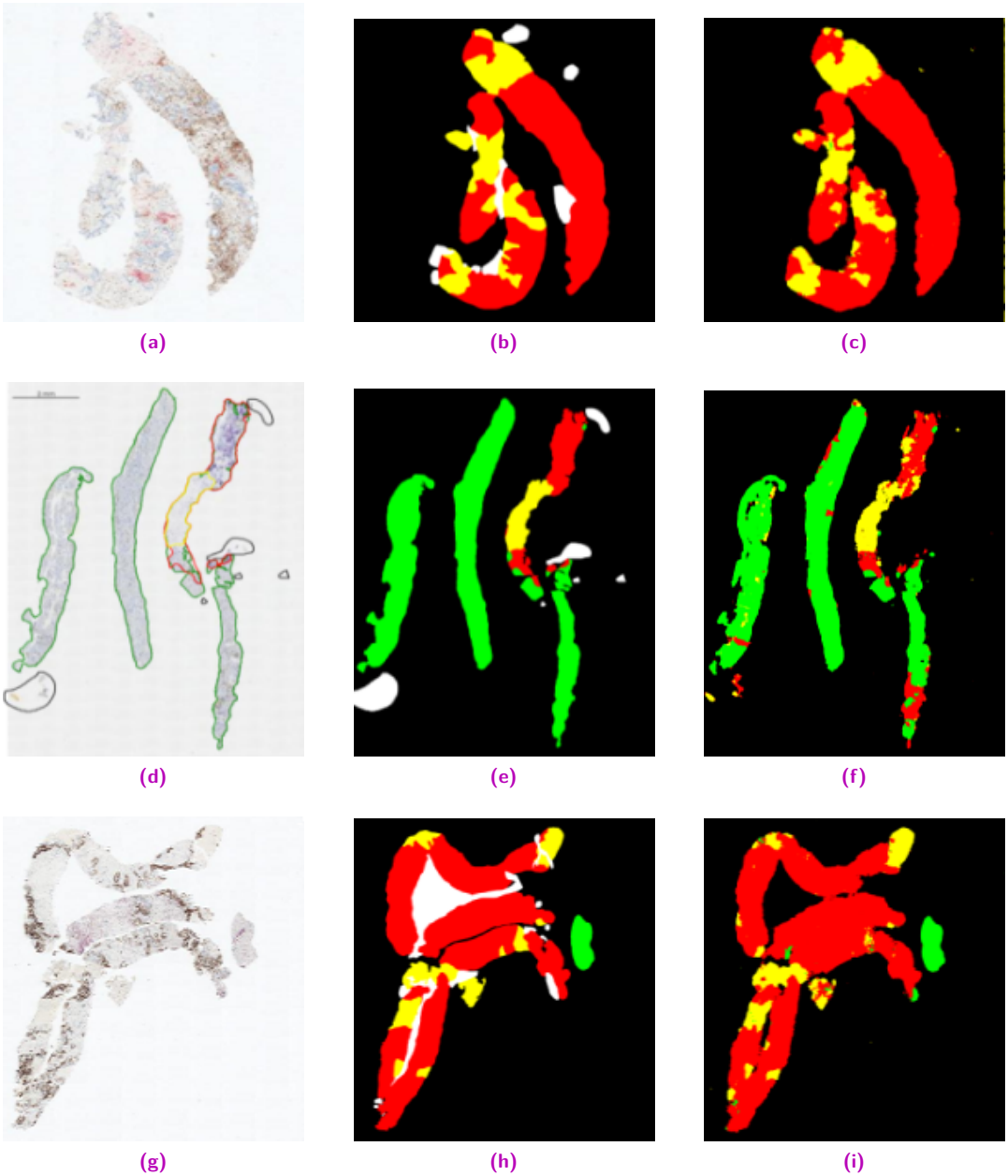


Fig. 7.5 Example of CD-UNET segmentation outputs of IHC images from the testing set. The first, second and third rows correspond to CD163/CD68, CD8/Ki67 and Ki67/CD3 respectively. (a), (d) and (g) correspond to the original images. (b), (e) and (h) correspond to the ground truth labels. (c), (f) and (i) correspond to the CD-UNET segmentation outputs.

8 Tools for Understanding the Network

Visualizing and understanding network decisions and intermediate computations is very important especially in the medical field where medical experts need to understand algorithm decisions in order to trust the results of automated analysis. For this purpose we propose to visualize and highlight pixels in the image that were significant for the network's output (often termed attribution) as well as the features learned by the network [52, 53, 54,55] and the outputs of specific layers.

Color Deconvolution Segment Visualizations The visualization of filters and feature maps has been one of the most prominent tools used in deep learning in order to facilitate the understanding of the network decisions [56, 57]. In order to visualize the effect of the color deconvolution segment on input images, we apply activation maximization to the filters of the first layer. Activation maximisation is a technique aiming to find the kind of information that a particular layer is trying to capture by generating patterns that maximize the mean activation of a chosen feature map in the considered layer of the trained network. In this case, instead of updating the weights to minimize a certain error function, the trained weights are fixed and the input image is updated in order to maximize the activation of a specific layer. Additionally, in order to visualize the effect of the color deconvolution segment we show some examples of its feature map outputs on different stainings.

- **Activation maximization of the first layer filters:** This approach allowed us to generate synthetic images that maximally activate the response of the first layer filters [58]. A noise image is inserted to the network, and several iterations of gradient ascent are run in order to modify the input image pixels to maximize the response of each of the filters. Fig. 8.1 shows the images we obtained following this approach. We notice that the resulted images correspond to stain colors from the training dataset.

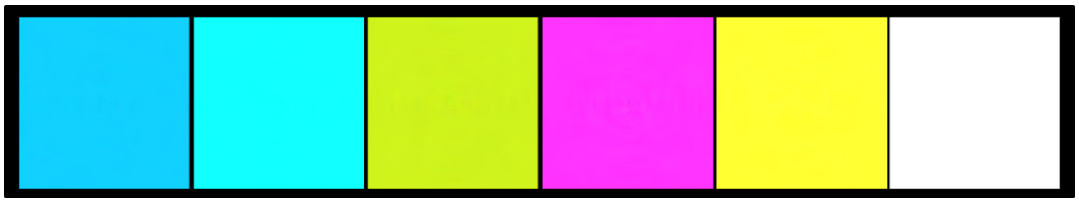
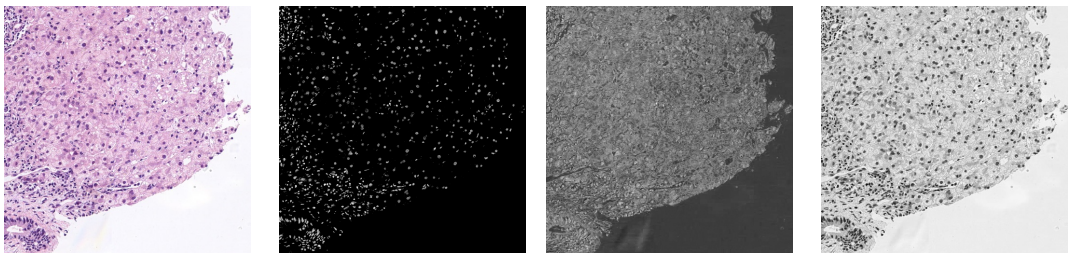


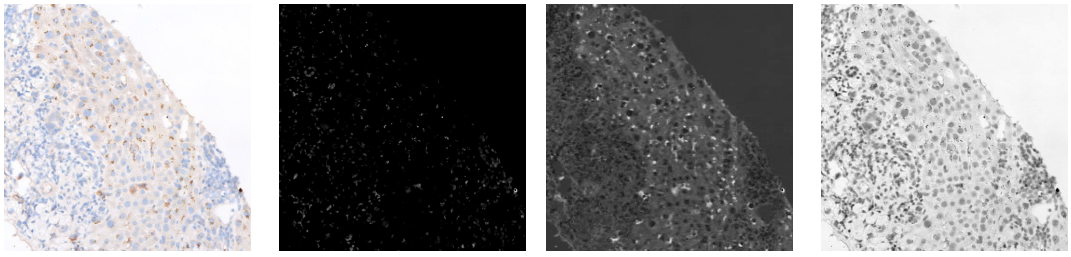
Fig. 8.1 Activation maximization of the filters of the first layer. The obtained images correspond to different stain colors. The white image corresponds to the background color.

- **Color deconvolution segment outputs:** In order to demonstrate the effect of the CD segment, we visualise its outputs using different stains (H&E and IHC). Figs. 8.2, 8.3, 8.4 and 8.5 show examples of input images and their corresponding outputs from the CD segment of the network.



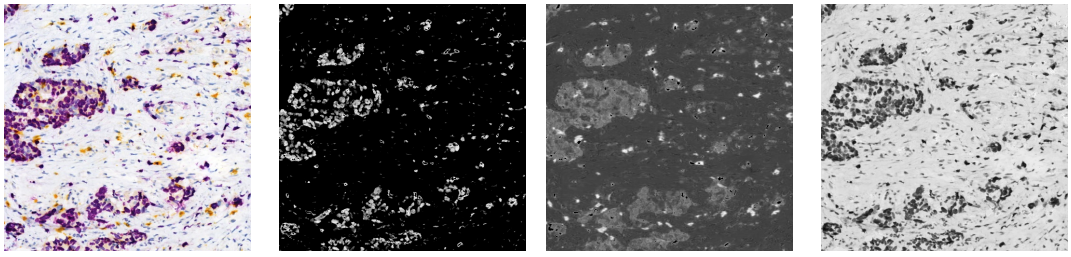
(a) (b) (c) (d)

Fig. 8.2 Example of the output of the color deconvolution segment on an H&E image. (a) corresponds to the Original image, (b) is the first output of the segment corresponding to the hematoxylin channel (blue cell nuclei), (c) is the second output of the segment corresponding to the eosin channel (pink connective tissue). (d) is the third output corresponding to the background in this case.



(a) (b) (c) (d)

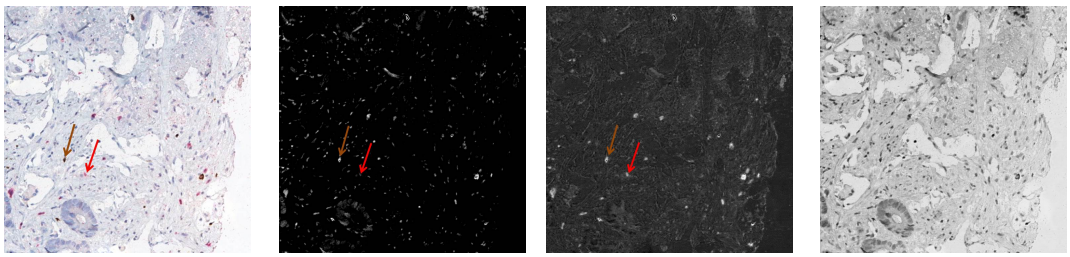
Fig. 8.3 Example of the output of the color deconvolution segment on a CEA image. (a) corresponds to the Original image, (b) is the first output of the segment corresponding to the hematoxylin channel (counterstain: blue cell nuclei), (c) is the second output of the segment corresponding to the CEA channel (cells with brown cytoplasm). (d) is the third output corresponding to the background.



(a) (b) (c) (d)

Fig. 8.4 Example of the output of the color deconvolution segment on a CEA image. (a) corresponds to the Original image, (b) is the first output of the segment corresponding to the Ki67 channel (purple cells), (c) is the second output of the segment corresponding to the CD8 channel (yellow cells). (d) is the third output corresponding to the background.

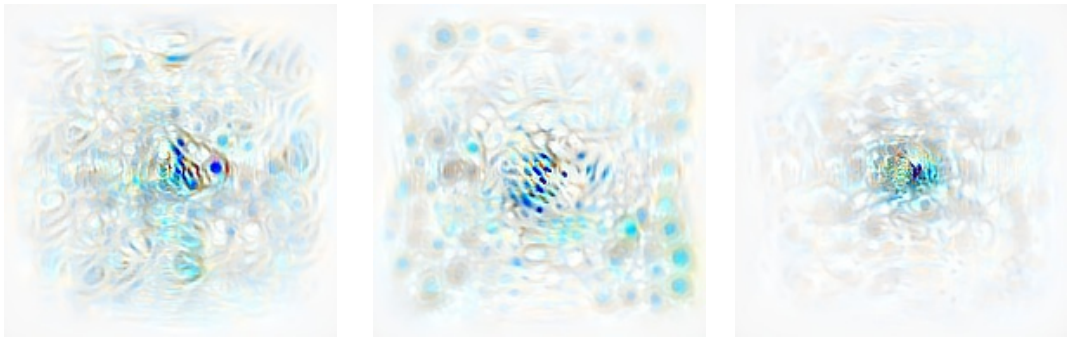
Network Decision Understanding In order for deep learning networks to be applicable in medical applications, it is necessary to make their decisions understandable and interpretable by humans. In this context, we apply two visualization techniques to histology image segmentation based on computing the gradients of the target score with respect to the input image.



(a) (b) (c) (d)

Fig. 8.5 Example of the output of the color deconvolution segment on a CEA image. (a) corresponds to the Original image, (b) is the first output of the segment corresponding to the hematoxylin channel (counterstain: blue cell nuclei), (c) is the second output of the segment corresponding to the CD3 channel (red cells). (d) is the third output corresponding to the background. We notice that the CD8 cells (brown cells) are detected in the first and the second outputs (brown arrows) while CD3 cells (red cells) are detected only in the second output (red arrows). This is probably due to the overlapping spectra of brown and red. Interestingly, the separation between red and brown cells is also a challenging task for pathologists.

- **Target class as seen by the network:** We use the same activation maximisation technique described in the first part of this paragraph and apply it to target class scores. In this case, a noise image is inserted into the network, a category in the network output is set as the target, and several iterations of gradient ascent are run in order to modify the input image pixels to receive a high value in the target pixel. Using this we can create examples of input images that cause a high activation at the target pixel for each of the categories (Fig. 8.6).



(a) (b) (c)

Fig. 8.6 Synthetic images that maximize the scores for tumor, tissue, and necrosis. (a), (b) and (c) correspond to tumor, tissue and necrosis respectively.

An interesting observation is that the area of the pixels in the input image that affect an output pixel (the effective receptive field) is different for the three categories. The tissue category score is maximised when there are tissue cell nuclei far from the target pixel, implying that patterns of multiple tissue cell nuclei around the target pixel are used by the network as clues of tissue presence. The tumor and necrosis categories, on the other hand, seem to look for patterns of condensed large distorted cell nuclei around the target pixel. The synthetic image for tissue shows regular cell structures in the centre but also far from the centre, meaning that patterns of cells around the target pixel were used by the network in order to make the decision.

- **Pixel attribution:** When the trained network makes a class prediction, it is helpful and informative to visualize the input pixels which were important for making this decision. This is called pixel attribution and can be done by computing a feature importance mask (also called a sensitivity map) for an image for a particular prediction. One possible way of generating the feature importance mask is by computing the gradient of a class prediction neuron with respect to the input pixels. The resulting gradient image gives an idea about the effect of a small change of each input pixel on the target class prediction. However, this type of gradient images is very noisy. In order to significantly reduce the aforementioned noise and make the sensitivity maps sharper, a technique called SmoothGrad [59] has been proposed. This technique is based on adding Gaussian noise to several copies of the input image then averaging the resulting gradient images. We used the SmoothGrad technique in order to generate class specific gradient images then we thresholded the resulting images in order to keep only the pixels that were very important for the network decision. Fig. 8.7 corresponds to the gradient images of the different categories.

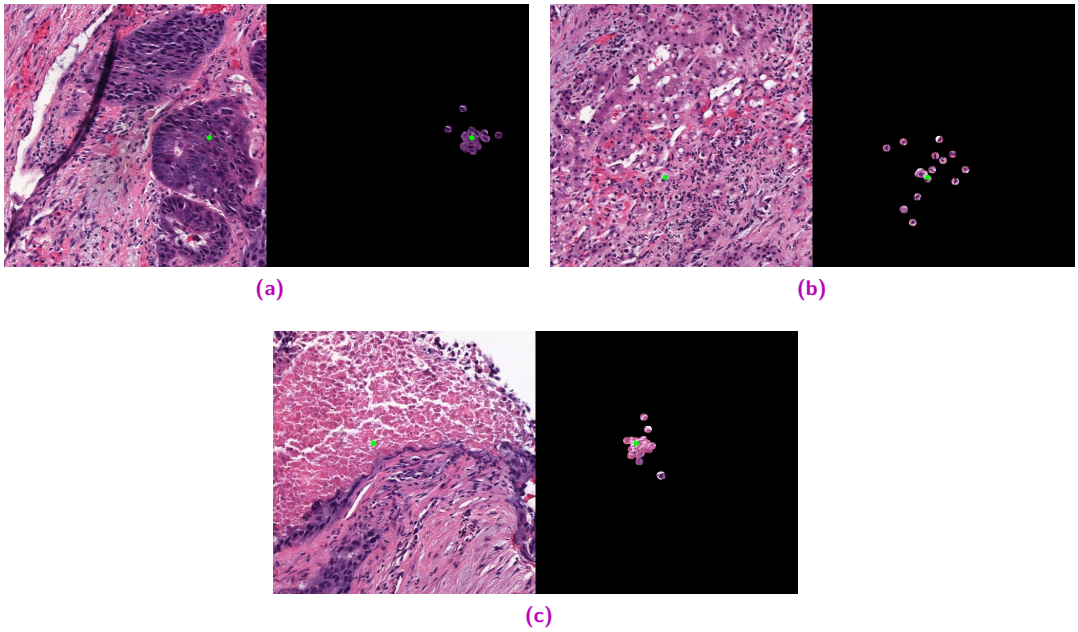


Fig. 8.7 SmoothGrad results. The target pixel is marked by the green circle. (a), (b) and (c) correspond to the SmoothGrad results for a tumor, a tissue and a necrosis pixel respectively.

An interesting observation in the visualisations was that for normal tissue and tumor, the gradients highlight, respectively, healthy looking and tumor cell nuclei in the target pixel surroundings while ignoring other texture. This observation was confirmed by an expert pathologist. Another important observation is that the effective receptive field was different between the different categories. The tissue category has a large effective receptive field compared to tumor and necrosis, which is in harmony with the interpretation of the synthetic images maximising the scores of the classes.

9 Conclusions and Discussions

Computerised segmentation of different tissue compartments has the advantage of being faster, less expensive, less laborious, and more accurate and objective than manual segmentation traditionally performed by expert pathologists. Using expert annotated slides in order to teach an automatic segmentation model has the additional value of reusing expensive annotations and biopsy slide images to generate additional value.

In the context of drug development, multiple IHC stains are used. In this work multiple stains were simultaneously used in order to train a unified segmentation model that deals with multistain histopathology images. Our experiments proved a higher difficulty in training the network for the more complex task of multistain image segmentation compared to the one stain scenario. Our interpretation was that the increased variability of the input image colors presented an additional complexity for the network and made the training process more difficult and erratic.

Several state-of-the-art methods used stain normalization as a pre-processing step in order to reduce color variability in the input space and improve segmentation and classification performance on different datasets. Other methods used color deconvolution as a pre-processing step. However, most state-of-the-art color deconvolution methods are highly dependent on the definition of stain matrices, which is a very subjective and stain dependent task. Actually, every staining has a different color deconvolution matrix and every matrix is optimised for a specific set of histology images obtained with specific experimental conditions. As a result, using these methods to separate between stains may result in inaccurate results if images from different sources or obtained with different experimental setups are used. Other color deconvolution methods are based on supervised learning approaches. However, this category of methods requires good quality annotated data for the specific task of color deconvolution for each specific stain, which is costly and challenging to obtain [167]. Some methods suggested including a stain separation layer as part of the network architecture [158]. However, they show that their method is highly dependent on filter initialisation, which should follow stain basis vectors. In our case, different stains (H&E and 8 IHCs) were used, and more stains are periodically added to our slide processing. In addition, defining stain reference vectors is a laborious and subjective task. We therefore present a method for generalising tissue segmentation over multiple stainings by adding a color deconvolution segment to the segmentation network architecture. The parameters of this segment are optimised during the regular learning process in an end-to-end training scheme. Adding the color deconvolution segment to the modified UNET substantially improved the convergence smoothness and speed of the network when training on a multistain dataset. The generalisation is also substantially improved, as can be seen from the network performance on the testing set. We hypothesize that this segment allows the network to deal with the variation of the input in the first color deconvolution layers and leaves the "rest" of the network with a much easier task to learn. The visualisation of the outputs of the color deconvolution segment using different stainings

as inputs shows the added layers actually learned to separate between different stain channels for the different stain types.

In order to enable understanding of the network architecture, we visualise synthetic images that maximise the scores of the different classes as well as different gradient images. This allowed us to see the effective receptive field that the network needs in order to make a decision for a specific pixel as well as the eigenimages that make the network predict a specific label.

Working with diverse datasets presents unique challenges for deep learning architectures. Our work shows that capturing some of the variability in a small network segment can reduce the complexity for the rest of the network architecture and improve the learning stability, the accuracy, and the generalisation capabilities. An end-to-end training scheme is imperative for applications with big operational pipelines in which manually calculating and inserting data to a system can cause bias and delay.

One possible way of improving the work is to investigate the cases where the network fails to predict the correct category. These cases could be discussed with expert pathologists to understand the reasons for failure and try to adapt the network accordingly. It could also be of interest to check if the missed predictions have an effect on pathological interpretations or if the obtained accuracy is good enough for a clinical diagnosis. A second limitation of the method is that it does not penalise the possible inaccuracy in the ground truth annotations used for training.

Part IV

Unsupervised Stain Virtualization

10 Introduction

> *It happens too often that new technology is frustrating. Therefore a lot of potentially good data is not being used.*
>
> — **Jeroen Windhorst**
> (Advisor within healthcare & IT)

Histopathological evaluation and analysis of tissue samples is a key practice in patient diagnosis and drug development, especially in oncology. As a result, stain virtualization has recently gained interest from the pathology and computer vision communities allowing simulation of stained tissue images thus saving lab and tissue resources.

10.1 Motivation

Medical diagnosis and clinical decisions rely heavily on the histopathological evaluation of tissue samples, especially in oncology. Historically, Hematoxylin and Eosin (H&E) has been used by pathologists as a gold standard staining in classical histopathology. However, in many cases, various target specific stains, including in-situ hybridization (ISH) and IHC, are needed in order to highlight and identify specific structures and/or targets in the tissue. As tissue is scarce and staining procedures are tedious and laborious, it would be beneficial to generate images of stained tissue virtually thus saving lab efforts, time and valuable tissue. Virtual staining could also generate in-silico multiplexing of different stains on the same tissue segment, enabling analysis of different markers in the same whole slide coordinate space and without the need to perform intermediate sensitive washing steps or deal with slide registration. This could enable pathologists to gain a better understanding of the tumor microenvironment. For these reasons, stain virtualization has gained more popularity and general interest over the last decade, especially with the new advances of machine learning based style transfer approaches. Thanks to the success of GANs and the progress of unsupervised learning, unsupervised style transfer GANs have been investigated as possible solutions to virtually generate realistic, clinically meaningful and interpretable images.

In this part of the dissertation we present novel solutions for unsupervised virtual stain synthesis in brightfield microscopy and validate our assumptions and results quantitatively and qualitatevely.

The content of this part is based on the following publications:

Amal Lahiani, Irina Klaman, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. "Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency". *Submitted to Journal of Biomedical and Health Informatics (JBHI), 2020*.

Amal Lahiani, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. "Perceptual embedding consistency for seamless reconstruction of tilewise style transfer". *Medical Image Computing and Computer Assisted Intervention (MICCAI), 2019*.

Amal Lahiani, Jacob Gildenblat, Irina Klaman, Shadi Albarqouni, Nassir Navab, and Eldad Klaiman. "Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach.". *European Congress of Digital Pathology (ECDP), 2019*.

10.2 Related Work

In the field of digital pathology slices of tissue are stained and scanned under a microscope. The staining types determine which parts or targets in the tissue are highlighted with specific colors and there are a multitude of staining techniques using chemical and biological processes e.g. H&E, IHC and ISH. Tissue staining aims to reveal relevant biological structures such as cell boundaries, cell types, tissue types, and biomarker amounts and distribution. Tissue staining materials and procedures can be time consuming and expensive, and typically require special expertise. Additionally, clinical human biopsy tissue can be a scarce and expensive material. These limitations in terms of available tissue and budget usually reduce the number of examinations and stainings performed on a sample. In many cases clinicians must select a subset of the stainings (and respective information) from their original wish-list due to these limitations. This can limit clinicians' ability to obtain all relevant information from a patient biopsy potentially reducing their ability to analyse and understand the tumor microenvironment and to improve biomarker and drug development research. In some tasks a correlative analysis of different biomarker specific stains is needed for a better and more accurate diagnosis [168].

In many cases information exists in the stained slide images about targets and objects not specifically targeted by the stain. For example, pathologists have the ability to identify lymphocytes in a H&E image [169] even without directly staining them for lymphocyte specific markers such as CD3. This fact motivated the research in the direction of generating virtually stained slides from other modalities and different style transfer approaches have been developed and applied for stain virtualization applications.

Style transfer (also called image-to-image translation) is a field with growing interest and use cases in deep learning having various computer vision and medical applications. It allows to learn a mapping between an input and an output image and to render an image in a new style while preserving its original semantic content. Style transfer applications include photos and objects from labels synthesis [46], portrait editing and augmented face generation [170], artistic style transfer [171, 172]], image domain adaptation [173] and super-resolution [174]. In the medical field, style transfer applications have also been used for reconstruction of PET, MRI and CT [175], cross-modality image synthesis [176] and virtualization of histopathological WSI tissue staining.

In the field of digital pathology, a multimodal pseudo-coloring method has been used in order to map confocal reflectance and fluorescence images into H&E images based on the experimentally determined colors of Hematoxylin and Eosin stains in a conventional histologic H&E staining [177, 178, 179]. Other groups used fluorescence images in order to virtually generate H&E images. In [180], Nadarajan et al. converted multiplexed immuno-fluorescence microscopy images to pseudo-color H&E images in order to automate multi-class segmentation ground truth labeling of H&E images. The methods in [181] and [182] aim to generate brightfield images from fluorescent images by using linear approximations between the fluorescence intensity and the concentration of the dyes in the tissue in order to combine the advantages of quantitative analysis and histopathological diagnosis. In [45], Giacomelli et al. demonstrated a physically realistic rendering approach modeling the transmission of a wavelength through a specific thickness of a specimen containing N absorbing dyes using a discrete model based on optical density from Beer–Lambert law in order to generate virtual H&E images from epi-fluorescence multiphoton microscopy data. In [5], we presented our first stain virtualization contribution based on using an optimization method in order to optimize model parameters and convert fluorescence into H&E images enabling tumor annotation by pathologists in multiplex fluorescence images.

Recently, deep learning approaches have been used successfully in image-to-image translation [46, 171]. In the field of histopathology, GAN based methods have been used in [183] and [184] in order to generate brightfield virtually stained images from unlabelled tissue-auto-fluorescence and label-free quantitative phase microscopy images respectively. GANs have also been used in [185], [186] and [187] in order to transform unstained hyperspectral tissue images into H&E images, non stained into H&E stained WSIs and H&E into immunofluorescent images respectively. CNNs have been used in [188] and [189] in order to make fluorescence label predictions from transmitted light images of unstained samples. Borhani et al. used Deep Neural Networks (DNNs) in order to map label-free multi-modal multi-photon microscopy images into H&E images [190]. Fujitani et al. used Fully Convolutional Neural Networks (FC-NNs) in order to convert H&E to Massons Trichrome images [191]. All of the aforementioned approaches are based on supervised training and require the use of paired registered datasets in order to spatially align the first modality input images and the corresponding output images of the second modality. The generation and preparation of these aligned datasets for training is very often a complex task in medical applications. In the case of histology, it includes either staining consecutive sections of each patient block or staining the same section with both protocols consecutively. The first option involves over-using tissue materials and the second, when biologically possible, requires advanced knowledge and expertise about antibody reactions and permanent effect on the tissue. Additionally, both options may result in the disappearance or appearance of structures, tissue deformations, folds or other artifacts, which makes the registration task difficult and inaccurate and may affect the supervised learning based model training.

One way to avoid the need for these aligned datasets and to facilitate dealing with variability present in sets of slide images is to use unsupervised style transfer methods for stain virtualization. In this part of the dissertation we present our contributions about unsupervised virtual stain synthesis. Around the same period of time, several research groups have also been interested in this topic. For example, unpaired image-to-image translation has been explored in [192] and [193] in order to transfer stain styles between brightfield modalities and perform

stain-independent segmentation of WSIs. Unsupervised H&E to IHC translation has also been used in [194] in order to improve the efficiency and accuracy of tissue segmentation.

10.3 Challenges

The size of high resolution WSIs is one of the main challenges in digital pathology. This makes tilewise processing necessary especially in the case of deep learning where memory and computational resources are limited by GPU hardware. In [68], Ulyanov et al. demonstrated that using instance normalization layers is beneficial and helps improve the quality of the synthesized images in the case of style transfer deep learning based applications. However, they also make the mapping functions dependent on the statistics of the input image and those of its corresponding feature maps. Instance normalization layers are applied at test time as well making a pixel in the output image depend not only on the network and the receptive field area but also on these statistics, which in turn results in applying different functions to adjacent pixels belonging to different adjacent tiles. For this reason, inference of trained generators on WSIs in a tilewise manner when instance normalization modules are used results in tiling artifacts between adjacent tiles in the reconstructed WSI.

Demonstration
Let's assume that:

$$y = g(x),$$

where x, y and g correspond to an input tensor, an output tensor and an instance normalization layer respectively. Let $x \in \mathbb{R}^{T \times C \times W \times H}$ be a tensor containing a batch of T images and x_{tijk} the $tijk^{th}$ element, where j and k correspond to spatial dimensions, i corresponds to the feature channel and t corresponds to the batch index. In this case, g can be expressed as:

$$y_{tijk} = g(x_{tijk}) = \frac{x_{tijk} - \mu_{ti}}{\sqrt{{\sigma_{ti}}^2 + \varepsilon}},$$

where μ_{ti} and ${\sigma_{ti}}^2$ are the mean and variance of the input tile. If we consider 2 adjacent pixels x_{tijk} and ${x^{'}}_{tijk}$ with similar values on the edges of two adjacent input tiles x and $x^{'}$ having very different statistics, the instance normalization functions g and $g^{'}$ applied to these two pixels will be completely different. This results in a tiling artifact in the generated whole slide image as adjacent output tiles might have significantly different pixel values on their borders (Fig. 10.1).

With instance normalization it is possible to use the same running mean and variance values for all the tiles at inference time in a way similar to inference with batch normalization. This approach allows indeed to remove the tiling artifact, but the resulting output at inference can have a lower quality and very faint colors. This effect can be explained by the fact that the training datasets are quite variable and containing both tissue and background, this makes the running mean and variance locally irrelevant.

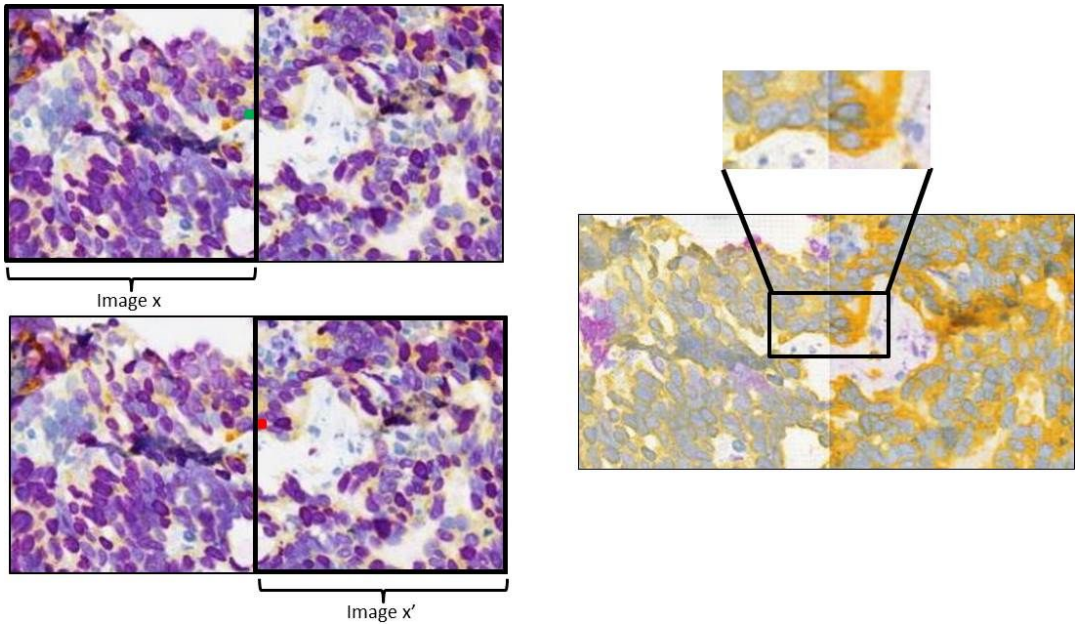


Fig. 10.1 (Right) Tiling effect in adjacent output tiles. (Left) Image x and $x^{'}$ correspond to 2 adjacent input tiles from a WSI. The green and red circles correspond to 2 adjacent pixels belonging to the same cell nucleus but to different input tiles.

Additionally, in neural network based style transfer, separating semantic content from style features is still a difficult and challenging task [172]. One way to improve the quality of the style-changed synthesised images is learning relevant high-level content and style features. In order to achieve this, some groups have investigated the effect of perceptual feature-level losses [174, 195] as a method for learning content related embeddings.

11 Methodologies

11.1 Unpaired Stain-to-Stain Translation

Our contributions to stain virtualization in digital pathology are based on the concept of cycle consistent adversarial networks (called CycleGANs) initially introduced in [171]. We adapted this type of networks to our specific use-case and proposed novel techniques in order to address stain virtualization challenges.

The original CycleGAN architecture allows transferring styles between two different domain spaces in an unsupervised fashion without the need for paired and registered images for training. It is based on the assumption that translating one image from the first domain ($\mathcal{X}$) to the second domain ($\mathcal{Y}$) and then back to the first domain should result in the same input image. The architecture is divided into four networks trained simultaneously: two mapping generators (G_1 and G_2) that transfer the styles between domains in both directions and two adversarial discriminators (D_1 and D_2) that attempt to match the distribution of the real and simulated images. Each generator has an encoder-decoder architecture: an input image x is fed to the encoder e_1 resulting in an embedding em_1, which is fed to the decoder d_1 to output an image $\hat{y}$. Similarly, the image $\hat{y}$ is fed to the encoder e_2 resulting in an embedding em_2, which is fed to the decoder d_2 in order to reconstruct back the input image $\hat{x}$. The objective function is a combination of an adversarial loss forcing the generators to generate images indistinguishable from the target distribution and a cycle consistency loss aiming at regularizing the training and preventing mode collapse by reducing the space of possible mappings. As described in [171], the full CycleGAN objective can be written as:

$$\begin{aligned}\mathcal{L}_{CycleGAN} &= \mathcal{L}_{GAN}(G_1, D_2, X, Y) \\ &+ \mathcal{L}_{GAN}(G_2, D_1, Y, X) \\ &+ \omega_{cyc}\mathcal{L}_{cyc}(G_1, G_2),\end{aligned} \tag{11.1}$$

where $\mathcal{L}_{GAN}(G_1, D_2, X, Y)$ and $\mathcal{L}_{GAN}(G_2, D_1, Y, X)$ correspond respectively to the adversarial losses of G_1 and G_2, ω_{cyc} to the weight of the cycle consistency loss and $\mathcal{L}_{cyc}(G_1, G_2)$ to the cycle consistency loss defined as:

$$\begin{aligned}\mathcal{L}_{cyc}(G_1, G_2) &= \mathbb{E}_{X\sim\mathbb{P}_X}[\|x - G_2(G_1(x))\|] \\ &+ \mathbb{E}_{Y\sim\mathbb{P}_Y}[\|y - G_1(G_2(y))\|]\end{aligned} \tag{11.2}$$

In the following, we will detail our contributions to unsupervised stain virtualization. As has been described in the previous chapter, due to the size of high resolution WSIs and memory bound hardware limitations, all network training and inference of the trained networks on the testing slides were done tilewise. Virtual WSIs of the testing dataset are then obtained by merging back the tile outputs.

11.2 Inference with Overlap

As a first solution we propose to use overlapping tiles during inference in order to reduce the tiling artifact. The solution is based on using a smaller input size and a sliding window for the instance normalization function statistics, allowing to have a smooth transition in the statistic values when deploying on 2 adjacent slides (Fig. 11.1).

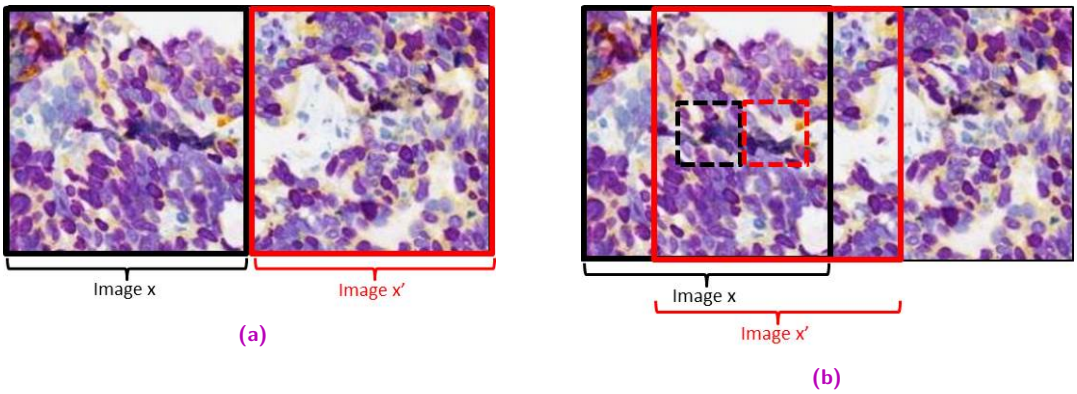


Fig. 11.1 Image (a) corresponds to the inference performed on 2 adjacent slides using the classical method. Image (b) corresponds to the new inference approach. The solid and dotted line squares correspond to the sliding window considered and the effective tile used for inference respectively.

Using a large tile overlap allows to mitigate the tiling artifact problem but the solution is computationally expensive during network inference and only partially solves the problem.

11.3 Perceptual Embedding Consistency

In order to make virtual staining of WSIs more efficient and robust for real world use cases, we aim to better address this instance normalization induced tiling artifact in a way that does not require superfluous processing. We introduce a novel Perceptual Embedding Consistency (PEC) loss function into the CycleGAN network architecture during training in order to regularize the effect of input image contrast, color and brightness perturbations in the generator latent space and hence substantially reduce the tiling artifact in the reconstructed virtual WSIs. We propose to add the PEC loss in the generator bottlenecks in order to minimize the distance between the bottleneck latent features of both generators (Fig. 11.2).

Theoretically this additional loss term would force the network to learn similar features of the same image presented with both stain styles. We propose to add this PEC loss to the overall

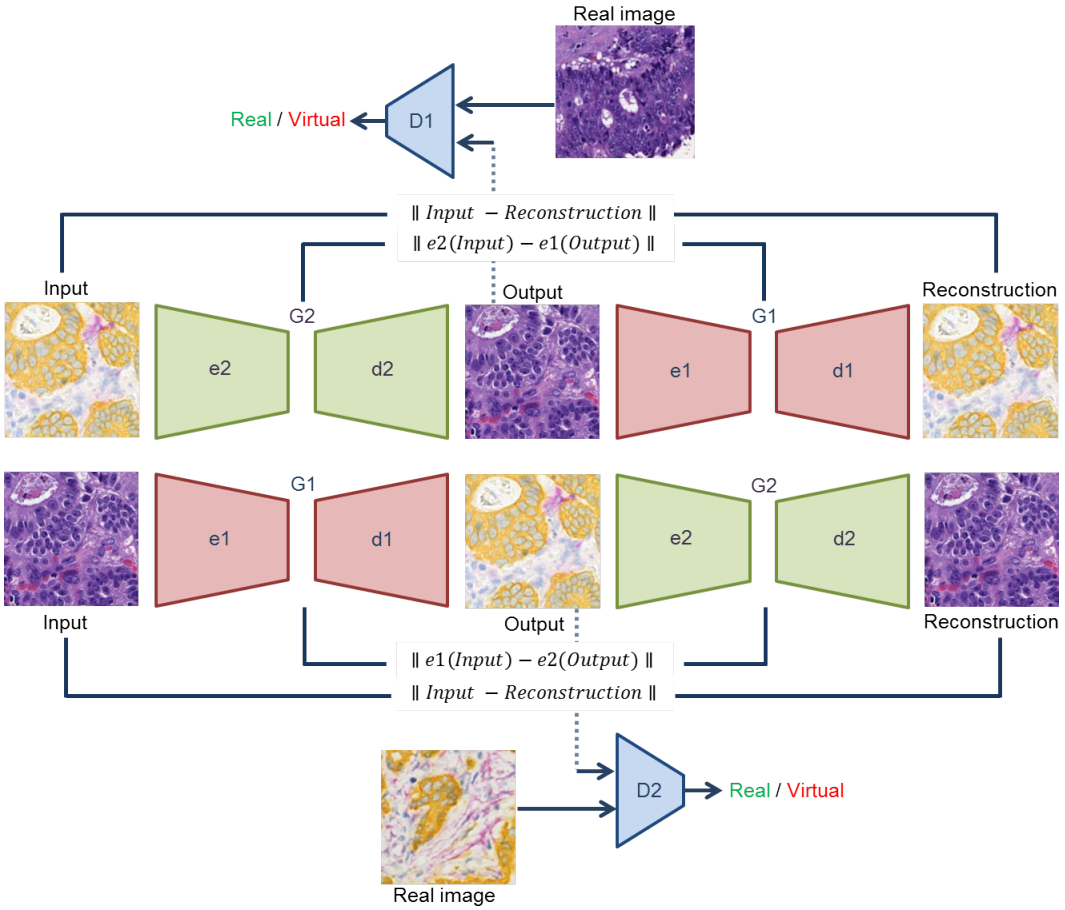


Fig. 11.2 Architecture of the proposed model. The overall objective includes the adversarial losses, the cycle consistency loss and the PEC loss in the generator bottlenecks. G, D, e and d correspond to generator, discriminator, encoder and decoder respectively.

objective function of the network along with the adversarial and cycle consistency losses of the CycleGAN cost function using the ℓ_2 norm between embeddings.

The roles of the adversarial and cyclic losses are generating realistic images and ensuring a structural correlation between input and simulated output images respectively. We argue that the PEC loss helps the network capture the semantics of the images without adding task specific prior knowledge such as shared segmentation or classification of an image in different domains. Our hypothesis is that adding this loss to the objective function forces the network to learn color, contrast and brightness invariant features in the generator bottlenecks. This reduces variability between adjacent output tiles in the reconstructed WSIs. Learning color, contrast and brightness invariant features in the latent space implies that the corresponding learned bottleneck feature maps consist mostly of semantic content information e.g. cell size and morphology, connective tissue texture and nuclear density. With this configuration, the role of the generator encoders is to extract condensed anatomical information and the role of the decoder is to add the style of the target domain to the extracted features. Under this assumption, input pixels belonging to adjacent tiles are translated to the output domain more

homogeneously and the resulting reconstructed WSIs are then substantially more continuous. We define the additional PEC loss term as:

$$\begin{aligned}\mathcal{L}_{embd}(G_1, G_2) = \mathbb{E}_{X\sim\mathbb{P}_X}[\|e_1(x) - e_2(G_1(x))\|_2] \\ + \mathbb{E}_{Y\sim\mathbb{P}_Y}[\|e_2(y) - e_1(G_2(y))\|_2],\end{aligned} \tag{11.3}$$

where e_1 and e_2 correspond to the encoders of the first and the second generator respectively and $\|.\|_2$ to the ℓ_2 norm. The full objective function can then be written as:

$$\mathcal{L} = \mathcal{L}_{CycleGAN} + \omega_{embd}\mathcal{L}_{embd}, \tag{11.4}$$

where ω_{embd} corresponds to the weight of the PEC loss.

12 Applications

12.1 From Ki67-CD8 to FAP-CK

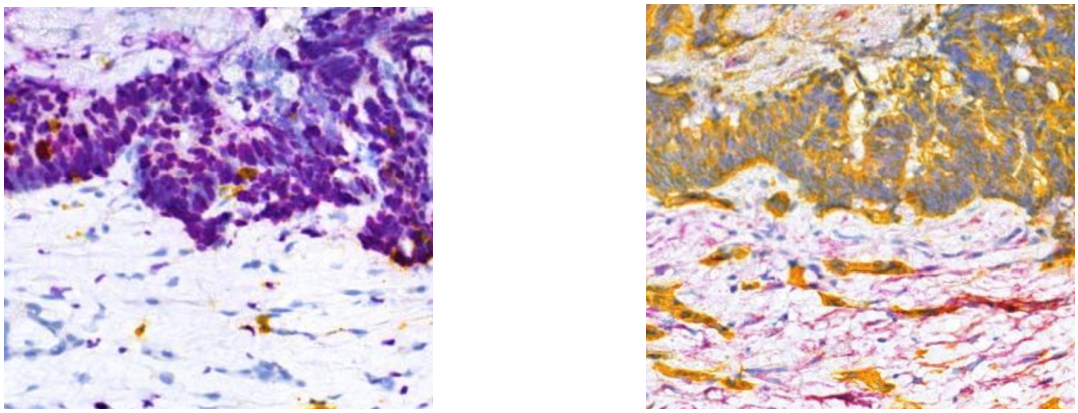
Our first stain virtualization application is based on IHC stainings and aims to virtually generate FAP-CK stained slide images from Ki67-CD8 stained slide images. Ki67 is a marker associated with proliferating cancer cells [196], CD8 corresponds to cytotoxic T cells, CK (CytoKeratin) is a marker for tumor cells [197], fibroblast activation protein (FAP) is expressed by cancer-associated fibroblasts in the stroma of solid tumor [198]. These input and output stainings were chosen for several reasons. First, information about tumor characteristics in FAP-CK could be encoded in the form of proliferation and tumor infiltrating lymphocytes in Ki67-CD8. Furthermore, Ki67-CD8 is one of the classical IHC stainings used in histopathology while FAP-CK is a new duplex IHC protocol allowing to characterize tumor and to advance research in the direction of drug development. Additionally, generating virtual FAP-CK stained slide images from Ki67-CD8 allows the creation of a virtual multiplexed brightfield image, i.e. having 4 target stains on the same whole slide coordinate system, which is technically challenging using classical staining and brightfield microscopy imaging methods due to limitations in stain separation caused by overlapping spectra.

12.1.1 Dataset Description

Our dataset consists of WSIs of Colorectal Carcinoma metastases in liver tissue from biopsy and surgical specimen selected from Roche Pathology image database. All the slides were chosen following a review of tissue, staining and image quality.

The training dataset includes 20 WSIs: 10 from Ki67-CD8 stained slides and 10 from FAP-CK stained slides, each from different patients. These training slides were tiled into overlapping 512×512 images at 10x magnification (half of the original scanning resolution). The reduced magnification allows us to have enough contextual information in the input which is needed in order to learn a meaningful feature set in the model while at the same time facilitates dealing with the computational memory limits. 10x magnification is usually sufficient for many tasks in pathology, for instance in the case of tissue segmentation. The tiling yielded 17025 tiles from Ki67-CD8 slides and 17812 tiles from FAP-CK slides. Fig. 12.1 shows examples of Ki67-CD8 and FAP-CK tiles.

In order to compare virtual and real FAP-CK images of similar tissue sections, the testing dataset consists of 10 pairs of Ki67-CD8 / FAP-CK slides from the same tissue block.



(a) (b)

Fig. 12.1 (a) Ki67-CD8 tile. Purple and yellow cells correspond to Ki67 and CD8 respectively. (b) FAP-CK tile. Purple and yellow correspond to FAP and CK respectively. In both stainings the counterstain (color of all cell nuclei) is Hematoxylin visible in light blue in the images.

12.1.2 Experiment Description

We trained a vanilla CycleGAN model as described in the previous section of this chapter in order to learn a mapping between Ki67-CD8 and FAP-CK staining spaces. As generator networks we used ResNet architectures [166] with 11 residual blocks. This ResNet architecture has a receptive field of 207×207 pixels, roughly corresponding to 190×190 microns on the 10x magnification image of the tissue. This receptive field size allows enough contextual information from the surroundings of the target pixel in the input to find meaningful histology features for the prediction of the virtual stain on the output image. As discriminator networks we used 70×70 PatchGans [199]. In order to distribute the training on multiple GPUs and accelerate the learning process, we implemented the stochastic synchronous ADAM algorithm [200] and used the pytorch distributed computing library which allowed us to use 12 GPUs concurrently on the Roche Pharma HPC cluster in Penzberg.

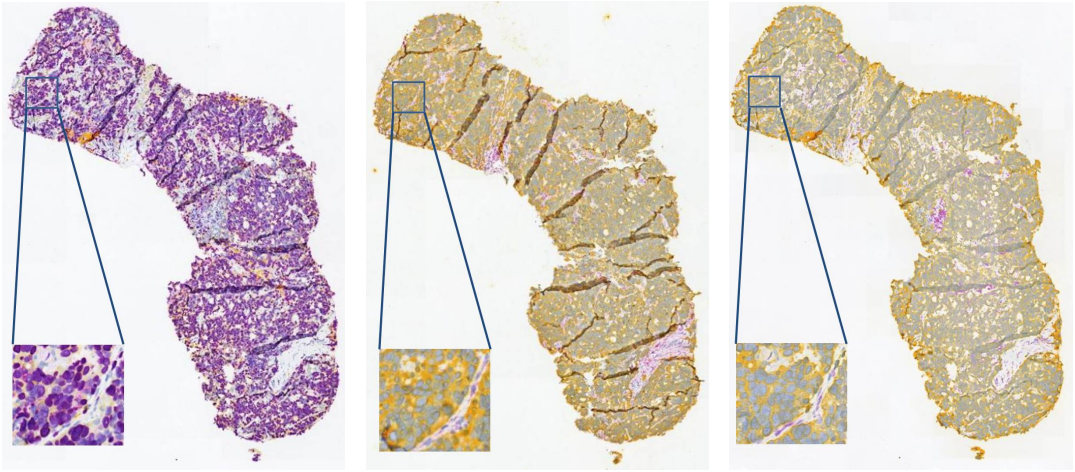
Inference of the trained network on the testing slides was also done tilewise then the tile output is merged back in order to obtain a virtual WSI. As explained in the previous sections we used overlapping tiles and an effective input size during inference in order to mitigate the tiling artifact problem. For this we used a smaller input size of 128×128 instead of 512×512 and a 512×512 sliding window for the instance normalization function statistics.

12.1.3 Results and Validation

In this paragraph, we show results obtained with the described experiment setup on images from the testing dataset. Then we describe the validation of this virtual staining method by comparing the results of an automatic cytokeratin positive (CK+) cell detection and FAP quantification algorithm trained on real stained images and applied to a set of virtual stained images and corresponding stained slides from the same tissue block.

Results

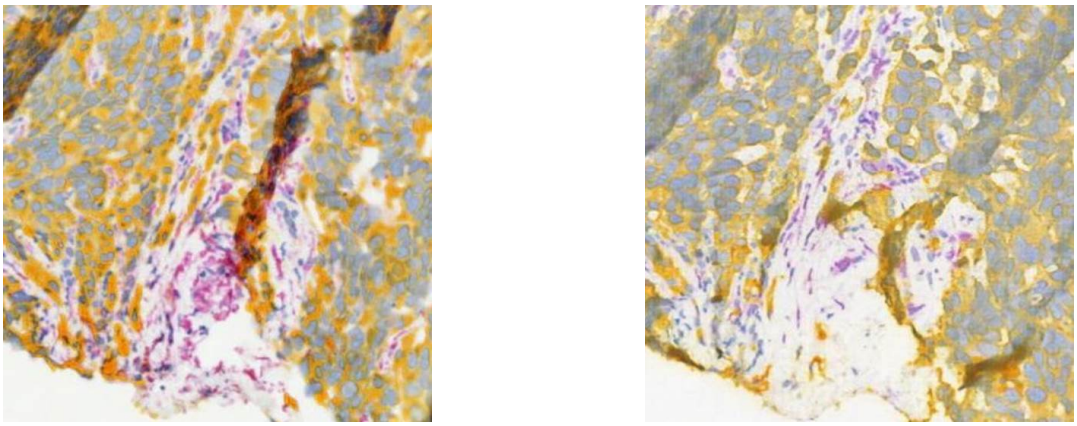
Visual assessment of the virtual generated images shows that the results are visually similar to the real staining of a slide from the same tissue block. Fig. 12.2 shows an example of a virtual FAP-CK image obtained using the trained CycleGAN and overlapping tiles in inference.



(a) (b) (c)

Fig. 12.2 a), (b) and (c) correspond to an input Ki67-CD8 image, the image of a real stained FAP-CK slide from the same tissue block and the virtual FAP-CK slide image.

However, we notice that in several cases FAP expression is different between the real and virtual images (Fig. 12.3). The localisation of FAP is generally successful however the patterns and the amounts are not always matching.



(a) (b)

Fig. 12.3 (a) and (b) correspond to FAP expression in a real and a virtual image respectively. We observe differences in FAP patterns and amounts.

Validation

FAP-CK images are mainly used by pathologists in order to quantify CK+ cells and FAP expressing areas in order to quantify the severity of the tumor and to estimate the success of FAP targeting drugs. We validate our results on a dataset of 10 testing paired images of slides from the same tissue block using an automatic algorithm for CK+ cell detection and

FAP quantification. The algorithm was developed and validated using real FAP-CK images. The results include CK+ cell densities and FAP densities in real and virtual WSIs. Fig. 12.4 shows a representation of the difference between the obtained results from virtual stain slides compared to real stained slides from the same tissue block.

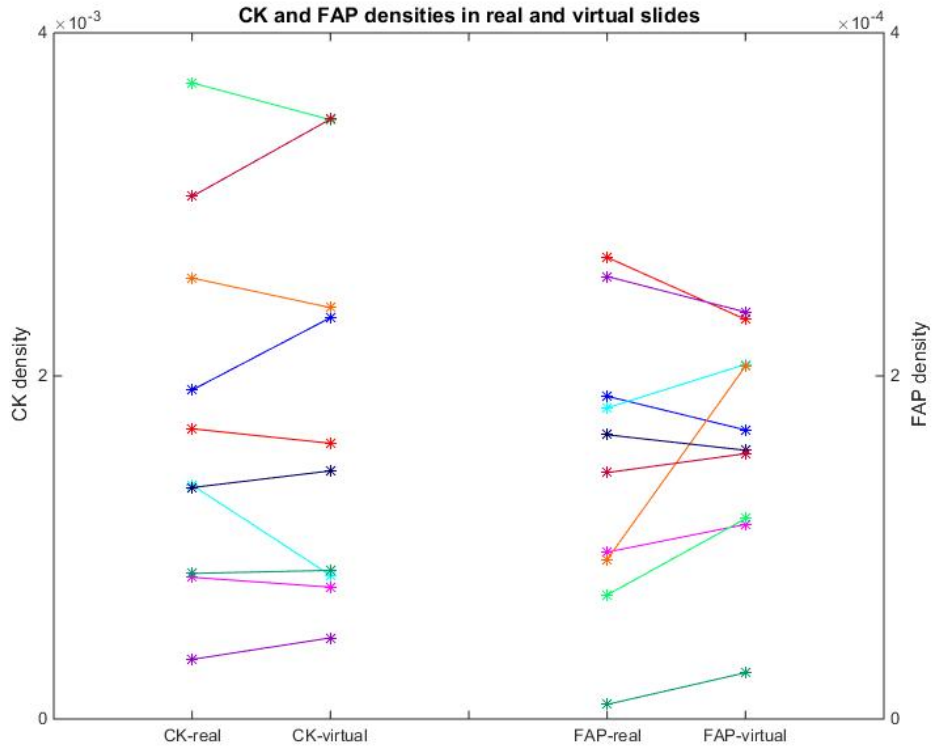


Fig. 12.4 Visualization of CK^+ cell densities (left side) and FAP densities (right side) in real and virtual whole slide images.

We can see from Fig. 12.4 that the difference between results on real and virtual stained slides is not one sided. This implies that our mapping algorithm does not consistently over-generate or under-generate CK or FAP. In order to visualize the difference between these densities we compute the absolute relative difference between the results obtained in the real and virtual slides. We show the results as a boxplot representation in Fig 12.5.

Analysis of the results shows a median absolute relative difference of 8% with 0.016 variance between CK densities in real and virtual slides. This was also confirmed by our expert pathologist who evaluated real and virtual paired slides and reported high correlation in CK expression. For FAP, we report a median of 14% with a variance of 0.466, reflecting a substantially higher variability than for CK. Our expert pathologist also confirmed this observation and mentioned that FAP features are completely not visible for pathologists in Ki67-CD8 staining.

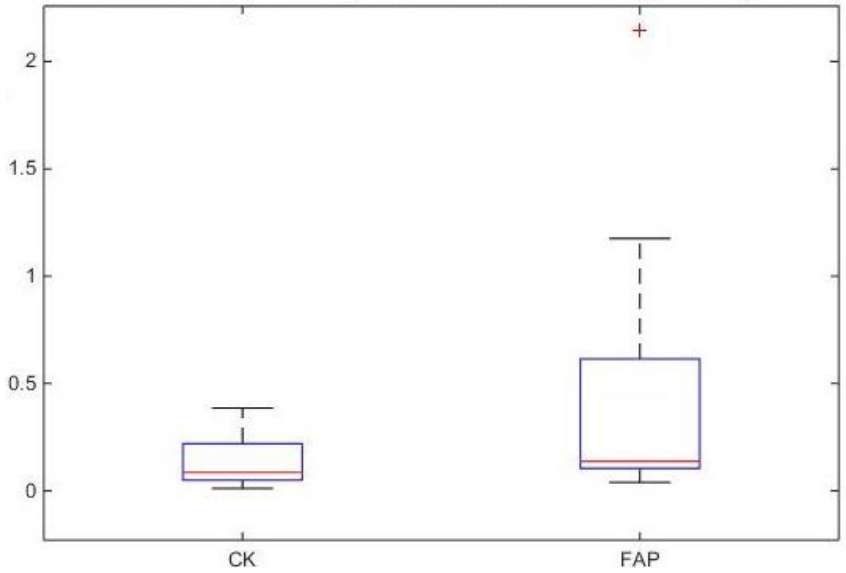


Fig. 12.5 Boxplot representation of the absolute relative difference between real and corresponding virtual slides for CK^+ cell densities(left side) and FAP cell densities (right side).

12.1.4 Discussion

Stain virtualization in IHC brightfield microscopy is an application with significant impacts on tissue diagnostics and drug development research. It has the advantage of making target specific stains more accessible and allows to reuse existing archival tissue images efficiently in order to virtually generate new stains.

Our experiment setup allowed us to synthesize virtual images looking to some extent visually similar to the real images taken from the same tissue block. However, we noticed that some tissue areas had different patterns in real and virtual images. As the real and synthesized slides do not originate from exactly the same tissue section, some of the observed differences can be attributed to differences between the input and reference tissue samples. Additionally, even when staining the same tissue with the same protocol, different images may result due to variations in tissue preparation, staining or scanning conditions. More investigations about the observed differences showed us that FAP reflected a substantially higher variability between real and virtual images than CK. Visual assessment of the results shows that the localization of FAP is partially successful, but the expression and amounts are different from the real slides. One explanation for this effect can be that FAP is associated with tumor growth and increased angiogenesis rather than with anatomical or phenotypic features [201]. If these functional features do not elicit a visible change in the input Ki67-CD8 staining, the model cannot correctly learn the mapping. Human identification and annotation of FAP expression in images of Ki67-CD8 stained tissue are extremely difficult if not virtually impossible. This observation is very interesting for future stain virtualization research directions as it allows to discover the limitations of simulation methods when biological constraints are present. However, in this specific case, the model success in localizing FAP, even if not flawlessly, suggests there are visual features in the input images that indicate the expression of FAP in the tissue. Identifying these visual features might lead to better FAP stain virtualisation as well as to new insights into tumor microenvironment anatomies. It is also worthwhile to note that

obtaining a virtual image which is slightly different from the real one should be acceptable as long as the pathologic interpretations of both of them are similar.

The validation of this first unsupervised stain virtualization application was based on measures obtained from the whole slides. The paired images we had for testing did not originate from the same tissue section and not always from consecutive sections, they were just taken from the same tissue block, making the registration of these paired images not always accurate and the pixelwise (or localized) validation not very meaningful. This is why whole slide level validation was more appropriate in our case. However, this has the limitations of discarding the space information and limiting the number of testing data points. In the next section of this chapter we try to bring more insights from stain virtualization validation by introducing new quantitative and pathological validation methods.

12.2 From H&E to FAP-CK

In this section we extend the stain virtualization framework to include another input staining and generate virtual FAP-CK from real stained H&E images. H&E is the gold standard and the most common staining technique used in histology. Hematoxylin stains cell nuclei in blue and Eosin stains extracellular connective tissue and cytoplasm in different shades of pink. As mentioned in the previous section, FAP-CK is a duplex IHC staining used for tumor diagnosis and drug development research. FAP is expressed in the stroma of solid tumor by cancer-associated fibroblasts and CK is a marker for epithelial tumor cells [197]. In the used FAP-CK assay, FAP staining is purple and CK is yellow. As H&E images are abundantly available, we propose to leverage this existing data in order to virtually generate new and unavailable information about these samples providing additional data for the research teams.

12.2.1 Dataset Description

We used a Roche internal histology dataset composed of 50 WSIs corresponding to 25 different patients. All WSIs are obtained from surgical specimen of Colorectal Carcinoma metastases in liver tissue and each patient is associated with 2 WSIs coming from 2 consecutive sections: the first section is stained with H&E and the second section with FAP-CK. We split this dataset into 10 WSIs (i.e. 5 H&E and 5 FAP-CK images corresponding to 5 patients) for training and 40 WSIs for validation.

Due to the size of high resolution histology images and hardware memory limitations, training WSIs were split into overlapping 512×512 tiles at 10x magnification. The tiling yielded 7592 H&E tiles and 7550 FAP-CK tiles.

The 40 testing WSIs are taken from 20 patients different from those used in training. They include 20 pairs of H&E and their corresponding real FAP-CK WSIs obtained from consecutive sections of the same tissue block. Additionally, in order to compare the difference between real and virtual staining to the upper-bound difference between consecutive real stained FAP-CK images, an additional third consecutive section is stained with the FAP-CK protocol for each patient of the test dataset. This results in three consecutive sections of the same tissue block

for each patient in the testing set: the first section stained with H&E and the second and third sections stained with the FAP-CK protocol. For validation and comparison, we register the consecutive WSIs of tissue sections of the testing set using a geometric point set matching method [202].

12.2.2 Experiment Description

In this H&E to FAP-CK translation experiment, we train a vanilla CycleGAN network (baseline) and the proposed PEC loss based model as described in the "Methodologies" section using the same hyper-parameters (as defined in the original CycleGAN paper). The generator networks and the discriminators are based on ResNet-6 and 70×70 PatchGAN architectures respectively. The generator networks are based on an encoder-decoder architecture where instance normalization is used in all the layers and the size of the bottleneck feature maps is $128 \times 128 \times 256$. All networks are trained from scratch for 50 epochs with a learning rate of 0.0002 and a batch size of 16. Then for both methods, the network configuration corresponding to the epoch with the smallest training loss is chosen for inference. Similar to [171], we set the cycle consistency loss weight to $\omega_{cyc} = 10$, and for our proposed approach we fix the PEC loss weight to be equal to the cycle loss weight ($\omega_{embd} = \omega_{cyc} = 10$). We train the models using the Roche Pharma High Performance Computing (HPC) cluster in Penzberg. We distribute the training on 8 v100 GPUs using the Pytorch distributed computing library and the stochastic synchronous ADAM algorithm [200].

12.2.3 Validation and Evaluation Metrics

We quantitatively evaluate our method and results by comparing virtually generated FAP-CK images and their corresponding consecutive real stained FAP-CK images using similarity measures. For this, we measure the Complex Wavelet Structural Similarity (CWSSIM) index, which is an image similarity metric extending the Structural Similarity (SSIM) index to the complex wavelet domain. CWSSIM is an index bounded between 0 for highly dissimilar images and 1 for perfectly matching images. As described in [119], small image distortions result in consistent phase changes in the local wavelet coefficients which does not change the structure of the local image features. As a result, CWSSIM is not very sensitive to small translations and rotations. This property was useful in our case since due to biological constraints our testing dataset is composed of consecutive sections stained each with one of the protocols. As small differences in tissue structures can be observed between consecutive sections especially in high resolution, perfect registration of consecutive slides is impossible. We report the metric for both virtual-real and real-real registered image pairs. Measurements on real-real registered image pairs serve as a baseline for the calculated evaluation metrics. Difference between images of real stained consecutive sections are typically caused by staining machine and reagent variability, imaging pipeline variability (e.g. microscope and scanner lighting and focus) and differences in the tissue anatomy of consecutive slides.

12.2.4 Ablation Test and Comparison

In order to visualize and quantitatively evaluate the effect of the proposed PEC loss, we compare virtually generated images and their corresponding consecutive real stained images by conducting an ablation test. The test consists in training the same model using the same generator and discriminator architectures and the same hyper-parameters with and without the embedding loss. Visual assessment of the resulting virtual stained tissue images shows that the learned mapping between staining domains is reasonable on the tile level for both networks. However, when considering the reconstructed WSIs, the network with PEC loss yields images with substantially reduced tiling artifact (Figs. 12.6c and 12.6f).

We additionally compare our results to the methods introduced in [3] and [203] as solutions to reduce the tiling artifact. The first solution is based on using a large overlapping window at inference time in order to smooth the transitions in instance normalization layer statistics. As described in [3] and [203], we use a small input size of 128×128 and an overlap of 384 pixels. While this solution partially reduces the tiling artifact, it is computationally very expensive: the inference has to be computed 16 times instead of once in order to get the output on a 512×512 tile. The second solution is based on freezing instance normalization layers at test time: i.e. using the running mean and variance obtained during training instead of the actual statistics when computing the network inference. This approach allowed to generate virtual images without the tiling artifact but the quality of the generated images was lower because the running statistics are not necessarily locally relevant. The resulted images are shown in Fig. 12.6.

For the quantitative validation, virtual stained and real stained WSIs of consecutive tissue sections are registered and tiled into 1024×1024 tiles at 10x magnification. The 20 testing WSIs yielded 2233 virtual tiles and their corresponding 2233 real tiles. We compute the CWSSIM between the real and virtual tiles for the models trained with and without the PEC loss. For the baseline model, we compute the similarity measure in the case of the normal inference, the inference with overlap and the inference with frozen instance normalization layers. We additionally compare these measures to those obtained from consecutive real stained registered images. Table 12.1 shows the obtained results.

Tab. 12.1 CWSSIM between virtual and real samples of the testing set and between consecutive real samples.

Mean (Median) ± Std	CWSSIM
Real consecutives (upper-bound)	0.94 (0.97) ± 0.085
CycleGAN (normal inference)	0.81 (0.83) ± 0.178
CycleGAN (overlapping inference)	0.81 (0.85) ± 0.179
CycleGAN (running statistics)	0.73 (0.75) ± 0.236
Our approach	**0.83 (0.88) ± 0.170**

The median CWSSIM measure for our approach reflects 6%, 3.5% and 17.3% of relative improvement compared to CycleGAN with normal inference, CycleGAN with overlapping inference and CycleGAN with frozen instance normalization layers at inference respectively. Additionally, we calculate the similarity measures per patient and we observe higher CWSSIM

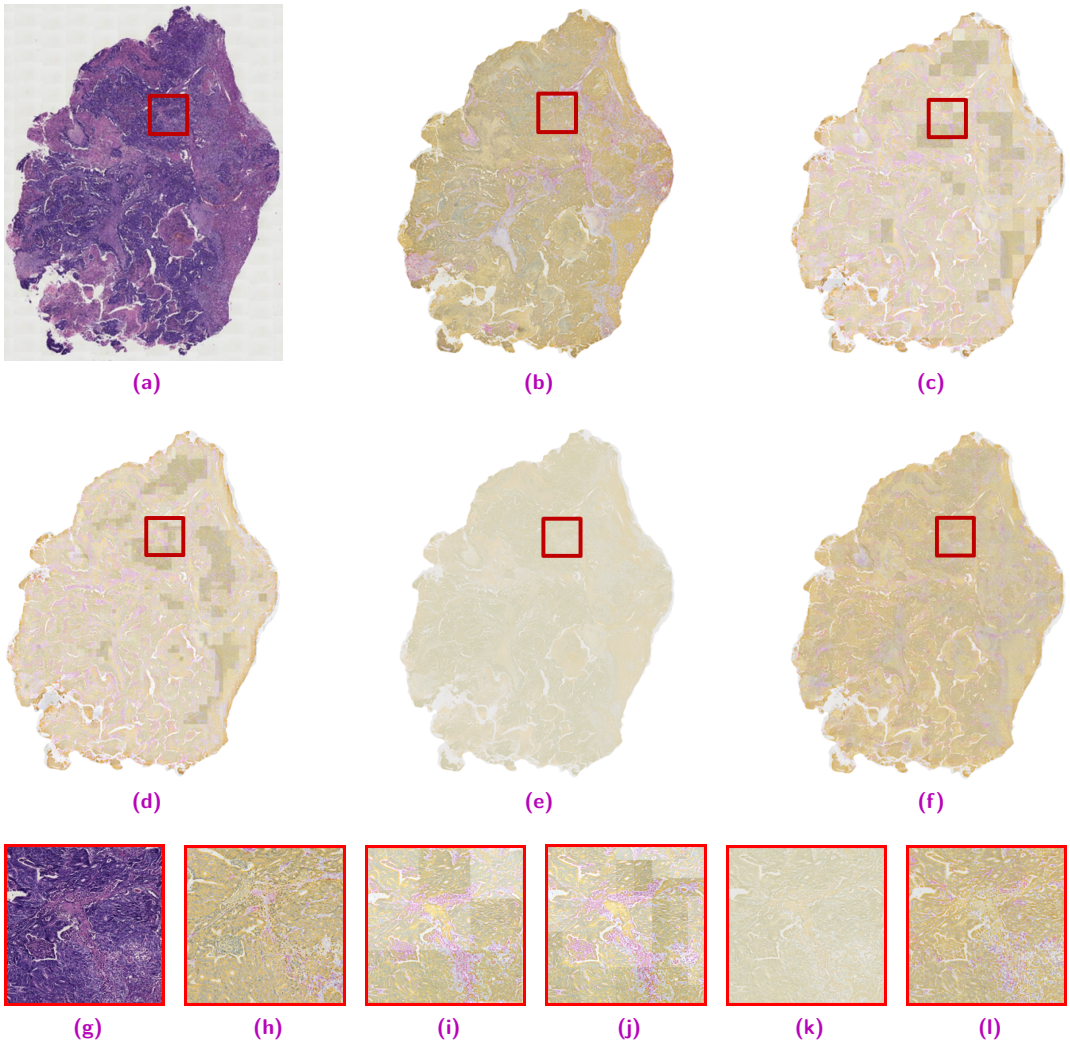


Fig. 12.6 (a) and (b) correspond to an input H&E image from the testing set and its corresponding consecutive registered real stained FAP-CK respectively. (c), (d) and (e) correspond to the virtual FAP-CK images obtained when training with the baseline model. (c) is obtained when using the normal inference, (d) is obtained when using large overlap inference and (e) is obtained when the learned training statistics are used in instance normalization layers during inference. (f) corresponds to the virtual image obtained when the proposed PEC loss is added and the normal inference is used. (g), (h), (i), (j), (k) and (l) correspond to zooms of the red boxes in (a), (b), (c), (d), (e) and (f) respectively. The effect of the proposed approach in generating more homogeneous and seamless reconstructed WSIs is clearly visible.

for 90% of the patients, 85% of the patients and all the patients compared to the baseline model with normal inference, with overlapping inference and with frozen instance normalization layers at inference respectively. We also observe that the upper-bound measurements from real consecutive sections stained under the same conditions present higher similarity as expected.

12.2.5 Embedding Visualization

In order to understand the effect of the added loss on the learned features, we propose to visualize the feature maps in the bottleneck layer where the PEC loss is introduced. Figs. 12.8a

and 12.8c show the first 25 bottleneck feature maps of one tile from the testing set in the case of a network trained without and with the PEC loss respectively. Interestingly the feature maps are much smoother when the PEC loss is introduced. The different structures and tissue types are more clearly highlighted and the noise features are reduced. This observation supports our assumption about the effect of the PEC loss on learning more semantic content based features. In order to verify the observation, we compute and compare the average Signal to Noise Ratio (SNR) of the bottleneck feature maps for different tissue compartments including background (Fig. 12.7a), necrosis (Fig. 12.7b), tumor cell area (Fig. 12.7c) and a selected structure inside tumor (Fig. 12.7d). We use SNR defined as the ratio between the mean and the standard deviation of the considered area. We report the results for the networks trained with and without the PEC loss in Table 12.2.

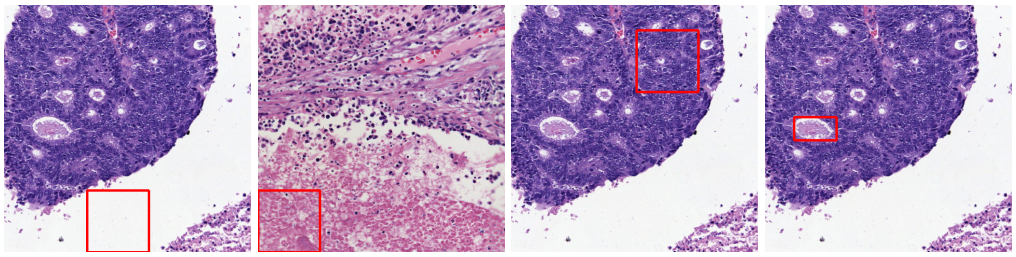


Fig. 12.7 The red crops in all the tiles correspond to the different image compartments used for SNR comparison. The crops in (a), (b), (c) and (d) correspond to background, necrosis, tumor cell area and a selected elliptic structure inside tumor respectively.

Tab. 12.2 Average SNR of bottleneck feature maps for different tissue compartments.

	CycleGAN	Ours
Background crop (Fig. 12.7a)	1.053	**4.175**
Necrosis crop (Fig. 12.7b)	0.917	**1.029**
Tumor cell area crop (Fig. 12.7c)	0.975	**1.132**
Elliptic structure crop (Fig. 12.7d)	0.934	**1.022**

The obtained results show higher SNR values with our approach compared to the baseline method for the different tissue compartments. This characteristic is visually verified in Figs. 12.8a and 12.8c where it can be observed that the different semantic regions in the input tile are more homogeneous areas in the embedding space of the proposed approach compared to the baseline network.

We further validate the assumption by visualizing the first 25 bottleneck feature maps of a contrast perturbed version of the original image and comparing these feature maps to those of the original image for both the baseline network (Figs. 12.8a and 12.8b) and the proposed network (Figs. 12.8c and 12.8d). The feature maps of the original and contrast perturbed images are much more similar and smooth for the proposed approach compared to the baseline network. This observation shows that the bottleneck feature maps are less sensitive to contrast variations when the PEC loss is used.

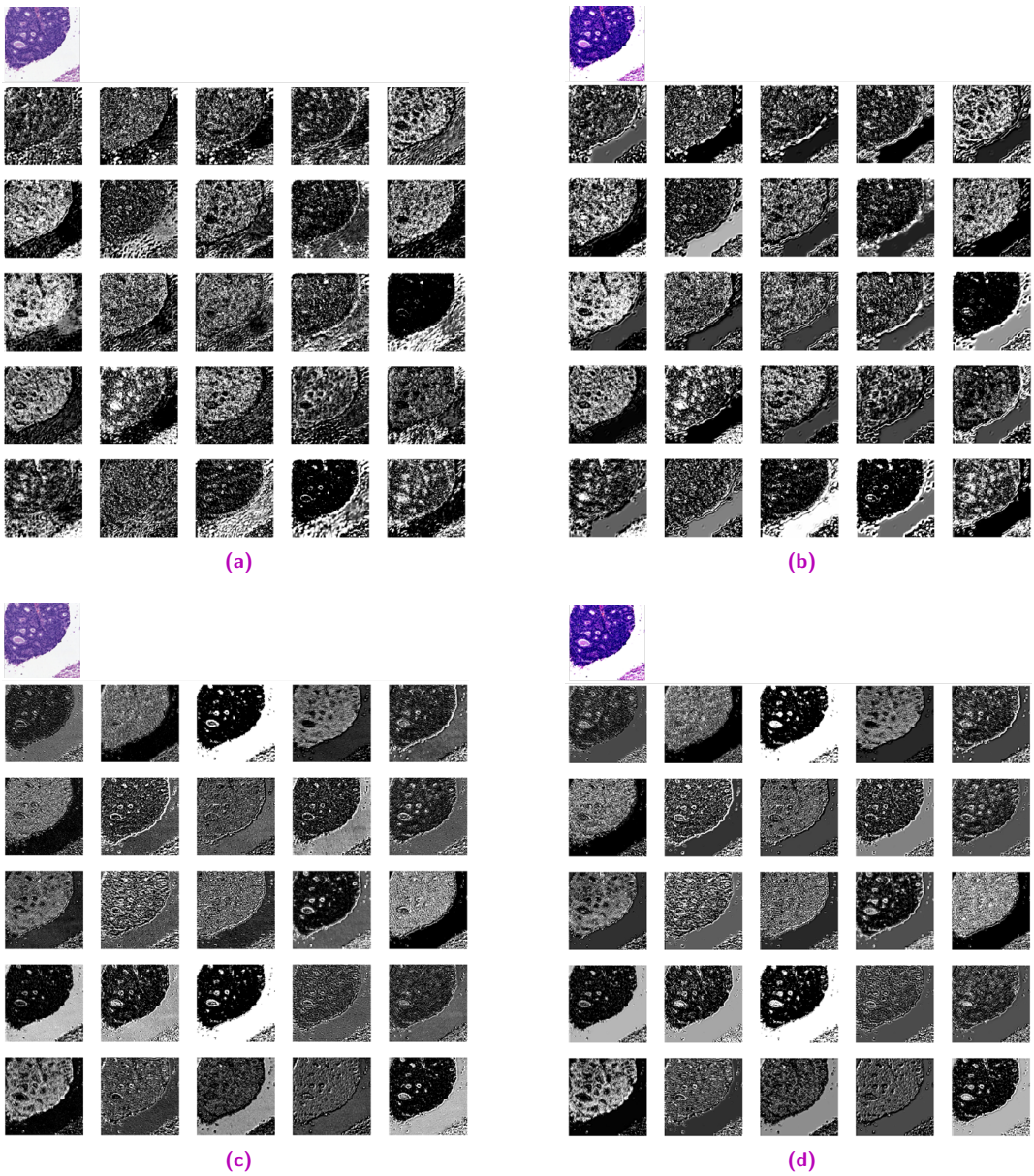


Fig. 12.8 (a) and (b) correspond respectively to the first feature maps of a clean image and its contrast perturbed version in the case of a network trained without the PEC loss. (c) and (d) correspond to the same images in the case of a network trained with the PEC loss.

12.2.6 Tumor segmentation from embeddings.

In addition to embedding visualization and SNR measurements, we further validate the robustness of bottleneck feature maps to color, contrast and brightness perturbations by training a tumor segmentation network using bottleneck embeddings of H&E images as inputs to the network. For this, we use H&E images of the same training and testing datasets that we used to train and test the stain virtualization networks. First, we compute bottleneck embeddings of the training dataset tiles using the trained stain virtualization networks in the case of the baseline model and our proposed approach respectively. For each of the cases, we use these generated embeddings in order to train a ResNet based segmentation

network with 2 classes: tumor and non-tumor. Then we evaluate the performance of both trained segmentation networks on embeddings of the testing dataset H&E images with random perturbations in color, contrast and brightness using F1 scores of both segmentation classes. The results are summarized in Table 12.3.

Tab. 12.3 Testing F1 segmentation scores.

	Tumor	Non-tumor
CycleGAN (baseline)	0.78	0.93
Our approach	**0.81**	**0.94**

The results show 3.8% and 1% relative improvement in testing $F1$ scores for tumor and non-tumor respectively when the embeddings generated with PEC loss based network are used. This demonstrates that adding the PEC loss allows to generate bottleneck embeddings that are more robust to color, contrast and brightness perturbations and that those embeddings are also better in an auxiliary downstream pathological task.

12.2.7 Sensitivity Analysis

Our theoretical assumption is that the PEC loss allows the network to learn content semantics rather than color, contrast and brightness related features. In order to further verify the validity of this assumption, we propose to study the sensitivity of the networks to different color, contrast and brightness perturbations. We compare these sensitivity values between the baseline and proposed networks. The experiment consists in introducing different perturbations to input H&E tiles from the testing set and evaluating the sensitivity of the bottleneck embeddings to these perturbations. For this we randomly select 100 512×512 input H&E tiles from the testing set, we apply different perturbations and we extract the bottleneck feature maps for the clean and perturbed versions. Then for each tile we compute the MSE in the embedding space between the original and perturbed tiles. We evaluate the results for the networks trained with and without the PEC loss. Fig. 12.9 shows the curves of the average MSE measures for the selected input images in the case of contrast, brightness and color perturbations. The difference in the embedding space between the clean and perturbed tiles is consistently and significantly smaller for all the perturbations when the PEC loss is used (Fig. 12.9). This demonstrates that the PEC loss helps the network learn less color, contrast and brightness dependent features. The semantic features make the trained network more robust and less sensitive to the variations of adjacent input tiles which in turn enables creation of more homogeneous and seamless reconstructed WSIs.

12.2.8 Pathological Validation

In order to pathologically validate the stain virtualization application, we propose to compare virtual and corresponding real images by comparing the pathological interpretation of two pathologists on randomly selected pairs of real/virtual tiles.

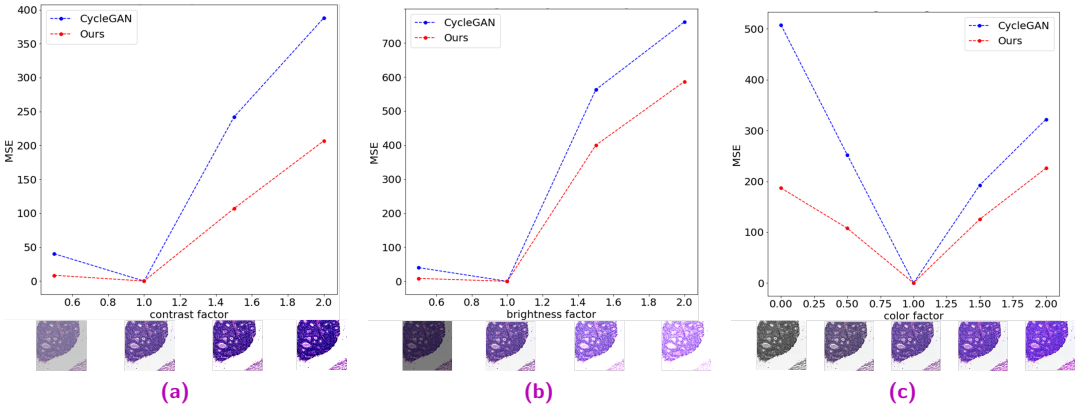


Fig. 12.9 Average MSE between the bottleneck embeddings of the original tiles and their corresponding different perturbed versions in the case of the network trained with (red curve) and without (blue curve) the PEC loss. (a), (b) and (c) correspond to contrast, brightness and color perturbations respectively.

Experiment Description

For the pathological validation, we randomly selected 100 FAP-CK tiles from the testing dataset WSIs consisting of 50 pairs of registered virtual/real images and we recorded the blind interpretation of pathologists on these images. For this we presented 2 pathologists with 100 images in a random order and asked them to pathologically interpret the image by deciding if it is CK+, FAP+ and/or necrotic. In order to measure the agreement between the pathologists' responses on real and virtual images and the inter-pathologist agreement, we calculated the accuracy and the Kappa coefficient corrected for prevalence and bias [204, 205]. This coefficient is bounded between -1 and 1 describing perfect disagreement and perfect agreement respectively.

Pathological Interpretation

We analyse the results for each clinical interpretation category (CK, FAP and necrosis). The images were classified by pathologists into CK+/CK-, FAP+/FAP- and necrosis/non-necrosis categories. For each of the categories, we calculate the accuracy and the Kappa coefficient between the evaluations made by each pathologist on real vs. virtual images. Then we compare to the measures of inter-pathologist agreement for evaluations made by both pathologists on real images. The results are summarized in Table 12.4. P1 and P2 refer to the first and second pathologist respectively.

The first two rows of each category correspond to the consistency of pathologist assessment on the virtual images vs. corresponding real images. The higher these agreements are, the closer the pathological interpretations of virtual and corresponding real images for the classification task. The third row corresponds to the average agreement of both pathologists for that task. The fourth row corresponds to the inter-pathologist agreement on real images only.

The average agreement of both pathologists for the interpretation of virtual and real images is higher than the inter-pathologist agreement on real images for CK and necrosis classification. This reflects that the "human error" (inter-pathologist disagreement) in the interpretation of real stained tissue images for these classes is higher than the ”errors” (intra-pathologist disagreement) between interpretations performed on real vs. virtual stained tissue images by

Tab. 12.4 Kappa coefficients for CK, FAP and necrosis classification.

	Accuracy	Kappa	Agreement
CK			
P1: Virtual vs. Real	0.96	0.92	Almost perfect agreement
P2: Virtual vs. Real	0.94	0.88	Almost perfect agreement
Average Virtual vs. Real	0.95	0.9	Almost perfect agreement
P1 vs. P2 on real images	0.88	0.76	Substantial agreement
FAP			
P1: Virtual vs. Real	0.78	0.56	Moderate agreement
P2: Virtual vs. Real	0.7	0.40	Fair agreement
Average Virtual vs. Real	0.74	0.48	Moderate agreement
P1 vs. P2 on real images	0.84	0.68	Substantial agreement
Necrosis			
P1: Virtual vs. Real	0.96	0.92	Almost perfect agreement
P2: Virtual vs. Real	0.94	0.88	Almost perfect agreement
Average Virtual vs. Real	0.95	0.9	Almost perfect agreement
P1 vs. P2 on real images	0.88	0.76	Substantial agreement

the same pathologist. As it is common to observe a certain degree of disagreement between experts in human expert interpretation of pathology images, using virtual staining for this clinical task would not increase the clinical interpretation error/disagreement above the normal amount observed for this task.

In the case of FAP classification, pathologist average interpretation agreement on real vs. virtual images is lower than the inter-pathologist agreement on real images. We also record lower agreement measures for FAP classification compared to CK and necrosis classification. This is mainly due to 2 major factors. First of all, FAP patterns are not perfectly generated in the virtual images and visual differences in FAP staining between real and virtual images exist. Additionally, FAP pathological interpretation is still a controversial task and the limit between FAP+ and FAP- is still subjective.

12.2.9 Discussion

In this subsection, we introduced a new PEC loss to an unsupervised style transfer network and applied the proposed approach to a stain virtualization application in order to simulate high resolution brightfield FAP-CK WSIs from real stained H&E WSIs. The proposed solution allows the generation of more homogeneous and seamless reconstructed virtual WSIs by adding a feature level constraint to the network. We show and demonstrate that this additional loss term forces the network to learn more content related information in the latent space which in turn reduces the effect of color and contrast variability of input tiles and helps substantially

reduce the tiling artifact resulting from the combination of tilewise processing and instance normalization layers in the network.

Visual results, feature visualization and quantitative measures demonstrate the effectiveness of the proposed method in solving the reconstruction tiling artifact, which is one of the major challenges in WSI stain virtualization. However, the synthesized images are not uniformly identical to the consecutive real stained images considered as ground truths. As has been concluded from the pathologist clinical validation, the main differences are related to different FAP patterns in virtual and real images. Similar to the discussions of the previous section of this chapter, these differences might be caused by the nature of FAP as a functional target without known anatomical histology features [201]. Under this assumption, it becomes likely that FAP predictive features in H&E images might be insufficient to reconstruct the correct FAP patterns in the virtual FAP-CK images. We believe that studying the effects of biological constraints on the success of stain virtualization and defining the histological requirements for a successful simulation would allow to define the scope of in-silico stain translation and to make the application more efficient and trustable.

Image Visual Realism

In order to test the visual realism of the simulated images, we used a set of 100 FAP-CK tiles from the testing dataset (i.e. 50 pairs of registered virtual/real images) and asked 2 pathologists to decide if the evaluated images come from the real stained tissue image set or the virtually generated image set. Then we calculated the Kappa coefficient corrected for prevalence and bias of the real vs. virtual classification. Both pathologists were shown the tiles one by one in a random order and were asked to decide if the image is real or simulated. Table 12.5 shows the agreement results for each pathologist (first couple of rows), the average (third row) and the inter-pathologist agreement (fourth row).

Tab. 12.5 Kappa coefficients for real vs. virtual classification.

	Kappa	Agreement
P1: Prediction / Ground truth	0.6	Moderate agreement
P2: Prediction / Ground truth	0.96	Almost perfect agreement
Average	0.78	Substantial agreement
P1 prediction / P2 prediction	0.56	Moderate agreement

On average the pathologists were in substantial agreement regarding the prediction whether the images are real or virtual. This means that while pathologist interpretations on the pairs of real and virtual tiles showed high agreement particularly for CK and necrosis classification, both pathologists were better than random in discriminating between real and virtual images. This reflects that some visual features in the synthesized images are not completely natural and identical to the real image features. While this does not present a serious clinical issue if the real and virtual images share the same pathological interpretations, it is interesting to investigate these predictive visual features and when possible synthesize more realistically looking virtual WSIs.

Additional Similarity Measures

For more objective validation measurements, in addition to the CWSSIM, we compute the PSNR and the Pearson correlation coefficient between the real and virtual tiles for the models trained with and without the PEC loss. We additionally compare these measures to those obtained from consecutive real stained registered images. Table 12.6 shows the obtained results.

Tab. 12.6 Similarity measures between virtual and real samples of the testing set and between consecutive real samples.

Mean (Median) ± Std	CWSSIM	PSNR	Pearson correlation
Real consecutives	0.94 (0.97) ± 0.085	18.48 (16.17) ± 7.64	0.41 (0.37) ± 0.249
CycleGAN (baseline)	0.81 (0.83) ± 0.178	17.01 (14.46) ± 7.11	0.30 (0.21) ± 0.245
Our approach	**0.83 (0.88) ± 0.170**	**18.05 (15.55) ± 7.31**	**0.35 (0.25) ± 0.270**

The median measures for our approach compared to the baseline reflect 7.5% and 19% of relative improvement for PSNR and Pearson correlation coefficient respectively. Additionally, these similarity measures are higher for all the patients. However, we noticed that the standard deviation (SD) of the PSNR and the Pearson Correlation coefficient are large in all cases, even in the case of real/real comparison. We believe that this is mainly related to the fact that the compared real and virtual slides come from consecutive sections and not from the same section. Thus, even after careful registration, small differences and distortions could still exist in some areas. As the PSNR and the Pearson correlation coefficient are very sensitive to small distortions, we observe high SD values. This is proved by the fact that the SD values are also large in the case of real registered pairs (upper-bound: first row of Table 12.6). For this reason, we think that these high values are correlated to biological constraints and that CWSSIM is the most convenient measure in our case because it is not very sensitive to small distortions.

Part V

Conclusions and Outlook

13 Summary and Findings

In this dissertation we explored digital pathology applications and introduced novel deep learning solutions aiming to address several challenges in the context of patient diagnosis and drug development research. In this part, we wish to sum up our work and contributions and to discuss some interesting directions for potential future research.

In the first part of this dissertation, we introduced the fundamental concepts of histology and digital pathology research. We presented the history and evolution of histology from the inception of the discipline in the 18th century until the recent advancements of digital pathology findings and solutions enhanced by the progress of computational power and image processing approaches. We additionally detailed the general workflows in histology, the main digital pathology applications and the different technical and computational challenges that need to be addressed in order to ensure a successful and efficient digital transition and to increase confidence and adoption in the very conservative field of human pathology.

In the second part, we provided an overview of deep learning based data driven approaches which have revolutionized computer vision tasks and recently also medical applications. We presented the fundamental principles of deep learning approaches including supervised and unsupervised learning with a focus on the most relevant deep learning algorithm subsets (i.e. CNNs and GANs). Then we introduced the recent advances in deep learning solutions for digital pathology image analysis and we discussed the common practices and the main related challenges and difficulties.

The third and fourth parts of the dissertation are dedicated to the details and discussions of our contributions to two different digital pathology applications. The first application consists in generalizing segmentation of different functional compartments over multiple brightfield stainings. In cancer immunotherapy development, quantification of pattern changes in microscopic WSIs of tumor biopsies and surgical specimen is one of the key steps allowing to control the evolution of the tumor, to improve the quality of the diagnosis and to enhance drug development research. In many cases, a correlative analysis of various biomarkers is needed in order to better understand the tumor microenvironment. As expert manual slide annotation is tedious, expensive and sometimes poorly reproducible, automating the segmentation task has the advantage of being faster, cheaper, more accurate and less prone to intra and inter-pathologist inconsistencies. Data driven segmentation approaches have the additional value of reusing existing expensive annotations and biopsy slide images in order to generate new insights and value from these data and hence improve the operational pipelines. For all these reasons, we introduced a supervised deep learning method based on CNNs in order to train a unified segmentation model that deals with multi stain histopathology images including H&E and multiple IHCs. The proposed solution is based on adding a color deconvolution module trained end to end and allowing the network to better deal with the variability of input image colors, to improve the convergence smoothness and speed during

training and to increase the generalization capabilities. The second application consists in virtually synthesizing new staining images from different real staining images in order to generate in-silico multiplexing of different stains on the same tissue segment, to enable analysis of different biomarkers on the same whole slide coordinate space and to save efforts and resources. As generating and preparing paired registered training datasets is a very sensitive and complex task especially in the case of medical applications, we used unsupervised deep learning style transfer approaches in order to learn a mapping function between the input and output staining spaces. Throughout this work, we addressed one of the main challenges of processing large images using deep learning based style transfer methods. The issue results from combining the use of instance normalization layers and tilewise inference and causes an undesirable tiling artifact. Initially, we proposed to mitigate this artifact by using large tile overlap during inference. Then we introduced a more efficient solution based on adding a perceptual embedding consistency loss during training in order to make the network more robust and less sensitive to the variations of adjacent input tiles and hence enable the creation of more homogeneous reconstructed WSIs. We additionally addressed the validation of the stain virtualization application using different methods including quantitative and qualitative approaches.

Throughout these projects, we had a lot of interest in understanding and explaining deep neural networks as a first step towards deploying this category of methods in real clinical scenarios and gaining more trust from experts to use the system. In the context of the segmentation project, we applied several visualization techniques in order to help understanding the network's decisions and behavior. For this we visualized synthetic images that maximise the scores of the different classes and the response of specific filters as well as different gradient images. We additionally visualized specific feature maps using several staining types as inputs in order to understand the effect of specific layers or modules. In the context of the stain virtualization project, we visualized specific embeddings and evaluated their noise levels. We additionally validated the effect of one of the loss terms by studying the sensitivity of the network to different input perturbations and by using bottleneck embeddings in an auxiliary downstream pathological task.

14 Outlook

Although this work allowed us to achieve promising results and very good performances, there are several possible future directions and extensions that can be studied and addressed.

Multistain Segmentation

Segmentation of different functional compartments has been an active research topic in digital pathology since the introduction of digital WSIs. In this dissertation, we presented a segmentation method based on CNNs that builds on state-of-the-art supervised deep learning approaches and provides solutions to some domain specific challenges.

One of the main challenges is that the number of annotated slides per stain is limited making training a different model for each stain very challenging. In our work we proposed to train a single model for all available stains as a solution to increase the size of the training dataset. However, one possible solution would be to use semi-supervised learning [206] where both labeled and unlabeled data are used in order to train the model. This is particularly useful in medical applications where the amounts of labeled data are limited. A second option would be to integrate active learning in the segmentation framework where experts are interactively and iteratively queried to label new selected data. This "smart" data selection aims to find the most informative samples which, when added to the training dataset, can result in a better accuracy. There are several strategies for this selective data sampling such us uncertainty sampling [101], query-by-committee [102], expected impact [207] and density-weighted methods [208].

Additionally, in patient diagnosis and disease detection, it is very important to quantify and control the uncertainty in a model's decision. In this case it is possible to mimic typical medical decision making and mark the uncertain samples as requiring additional attention by experienced medical experts or incorporate the uncertainty measures in the training phase in order to improve the performance of the model. In our case, discussing the non confident data points with pathologists can help understand the reasons for failure and when possible incorporate the uncertainty estimate or domain-specific knowledge in the learning process in order to improve the generalization performance. This would also allow to estimate the quality of the predictions and to increase the reliability of deep learning models.

Detecting and understanding models' failure situations is of extreme significance particularly in medical diagnosis where some errors are not tolerable. This would allow to analyze the limits of machine learning algorithms and to find appropriate solutions to make them more robust and trustable. Recent work has shown that deep neural networks can be fooled by adding small imperceptible perturbations to normal inputs. These carefully engineered malicious examples are called adversarial attacks and rise a lot of safety concerns in the context of clinical settings. A possible extension of our work would be to understand which adversarial

attacks can fool the deep learning segmentation models. While the adversarial attacks are unlikely to occur in real data, finding these attacks and training the networks to be more robust against them would help understand the limits of the model and to appropriately decrease the confidence of the model's decision on uncertain inputs.

In the field of histopathology data labeling is done on high resolution WSIs by expert pathologists. Segmenting the different functional compartments on very large images by humans can result in some segmentation inaccuracies particularly on the borders between different classes. A further future direction of our work would be to investigate the effect of these ground truth annotation inaccuracies on the model's performance. One possible way of doing this would be penalizing a margin area on the borders between the different compartments during training.

Stain Virtualization

In this dissertation, we presented different stain virtualization applications aiming to synthesize a new staining from another input staining using unsupervised deep learning approaches. While the achieved results were very promising, there is still a lot of room for improvements and potential future directions.

As has been deducted from FAP-CK image synthesis, studying the effect of biological constraints on the success of stain virtualization would be a natural extension of the work presented in this dissertation. One possible way of investigating this topic would be by visualizing the predictive features in the input staining that were important for the generation of each output marker then discussing these features with expert pathologists in order to assess their relevance.

A further possible future direction would be to investigate the possibility of synthesizing a virtual staining using other modalities as input. One option that we have started investigating is to generate H&E images from ultramicroscopy images. Ultramicroscopy is based on light sheet illumination and allows optical sectioning and 3D imaging of cleared samples with micrometer resolution without the need for mechanical slicing. Synthesizing H&E and IHC stains from ultramicroscopy images would revolutionize the field of histology. However, several challenges are still to be addressed in order to make the application useful and efficient. Ultramicroscopy imaging resolution is one of the current limitations. Although significant advancements have been recently reached in improving the imaging resolution of light sheet fluorescence microscopy systems, the current systems are still unable to generate images with the same resolution as confocal microscopes. An additional challenge of this application is the difficulty to validate the models in order to make them reliable and ready to be used in clinical settings. A possible way of validation would be to image the testing samples with ultramicroscopy then rehydrate them and prepare and cut them for real H&E staining. Real and virtual images can then be compared quantitatively or qualitatively by an expert in order to evaluate their similarity. However, this setup is not straight forward and requires a lot of sensitive work. For instance, it is very difficult to find the exact section in the ultramicroscopy block that matches the real stained H&E section.

I hope that the work and challenges presented in this dissertation will inspire the conception of new ideas and the development of new solutions and will promote further research. As it is the case in all medical applications, the deployment of these methods in real clinical settings

is sensitive and challenging. However, I believe that this would be possible if all contributors, including expert pathologists and researchers and clinical partners, collaborate and combine their knowledge and efforts in order to make digital pathology more efficient and patient's life more pleasant. I believe that the results and conclusions presented in this dissertation are promising and encouraging and that our work has contributed to open exciting and promising directions for further research and improvements.

Bibliography

[1] A. Lahiani, I. Klaman, N. Navab, S. Albarqouni, and E. Klaiman. “Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency”. In: *IEEE Journal of Biomedical and Health Informatics* (Feb. 2020) (cit. on pp. 1, 24).

[2] A. Lahiani, N. Navab, S. Albarqouni, and E. Klaiman. “Perceptual embedding consistency for seamless reconstruction of tilewise style transfer”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham: Springer International Publishing, 2019, pp. 568–576 (cit. on pp. 1, 24).

[3] A. Lahiani, J. Gildenblat, I. Klaman, S. Albarqouni, N. Navab, and E. Klaiman. “Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach”. In: *European Congress on Digital Pathology*. Springer. 2019, pp. 47–55 (cit. on pp. 1, 24, 72).

[4] A. Lahiani, J. Gildenblat, I. Klaman, N. Navab, and E. Klaiman. “Generalising multistain immunohistochemistry tissue segmentation using end-to-end colour deconvolution deep neural networks”. In: *IET Image Processing* 13.7 (2019), pp. 1066–1073 (cit. on p. 1).

[5] A. Lahiani, E. Klaiman, and O. Grimm. “Enabling histopathological annotations on immunofluorescent images through virtualization of hematoxylin and eosin”. In: *JPI* 9 (2018) (cit. on pp. 1, 57).

[6] G. Musumeci. *Past, present and future: overview on histology and histopathology*. 2014 (cit. on p. 9).

[7] L. Pantanowitz, P. N. Valenstein, A. J. Evans, et al. “Review of the current state of whole slide imaging in pathology”. In: *Journal of pathology informatics* 2 (2011) (cit. on p. 9).

[8] C. R. Taylor. “From microscopy to whole slide digital images: a century and a half of image analysis”. In: *Applied Immunohistochemistry & Molecular Morphology* 19.6 (2011), pp. 491–493 (cit. on pp. 9, 10).

[9] I. Hussein, M. Raad, R. Safa, R. A. Jurjus, and A. Jurjus. “Once Upon a Microscopic Slide: The Story of Histology”. In: *Journal of Cytology & Histology* 6 (2015) (cit. on p. 9).

[10] X. Chen, B. Zheng, and H. Liu. “Optical and digital microscopic imaging techniques and applications in pathology”. In: *Analytical Cellular Pathology* 34.1-2 (2011), pp. 5–18 (cit. on p. 9).

[11] D. Evanko, A. Heinrichs, and C. Karlsson. “Milestones in Light Microscopy”. In: *Nature* (2011) (cit. on p. 9).

[12] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz. “Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives”. In: *Journal of pathology informatics* 9 (2018) (cit. on pp. 9, 10).

[13] H. A. Alturkistani, F. M. Tashkandi, and Z. M. Mohammedsaleh. “Histological stains: a literature review and case study”. In: *Global journal of health science* 8.3 (2016), p. 72 (cit. on p. 9).

[14] S. Dunst and P. Tomancak. “Imaging Flies by Fluorescence Microscopy: Principles, Technologies, and Applications”. In: *Genetics* 211.1 (2019), pp. 15–34 (cit. on p. 9).

[15] R. S. Weinstein, A. R. Graham, L. C. Richter, et al. “Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future”. In: *Human pathology* 40.8 (2009), pp. 1057–1069 (cit. on p. 9).

[16] S. M. Williams, W. H. Henricks, M. J. Becich, M. P. Toscano, and A. B. Carter. “Telepathology for patient care: what am I getting myself into?” In: *Advances in anatomic pathology* 17 2 (2010), pp. 130–49 (cit. on p. 9).

[17] C. V. Hedvat. “Digital microscopy: past, present, and future”. In: *Archives of Pathology and Laboratory Medicine* 134.11 (2010), pp. 1666–1670 (cit. on p. 10).

[18] R. S. Weinstein, M. J. Holcomb, and E. A. Krupinski. “Invention and early history of telepathology (1985-2000)”. In: *Journal of pathology informatics* 10 (2019) (cit. on p. 10).

[19] L. Pantanowitz. “Digital images and the future of digital pathology”. In: *Journal of pathology informatics* 1 (2010) (cit. on p. 10).

[20] G. Bueno, M. M. Fernández-Carrobles, O. Deniz, and M. García-Rojo. “New trends of emerging technologies in digital pathology”. In: *Pathobiology* 83.2-3 (2016), pp. 61–69 (cit. on p. 10).

[21] P. W. Hamilton, P. Bankhead, Y. Wang, et al. “Digital pathology and image analysis in tissue biomarker research”. In: *Methods* 70.1 (2014), pp. 59–73 (cit. on pp. 10, 13, 14).

[22] M. H. Ross, W. Pawlina, and A Histology. *A text and Atlas*. 2006 (cit. on p. 10).

[23] T. D. Hewitson and I. A. Darby. *Histology protocols*. Springer, 2010 (cit. on p. 10).

[24] S. M. Hewitt, F. A. Lewis, Y. Cao, et al. “Tissue handling and specimen preparation in surgical pathology: issues concerning the recovery of nucleic acids from formalin-fixed, paraffin-embedded tissue”. In: *Archives of pathology & laboratory medicine* 132.12 (2008), pp. 1929–1935 (cit. on pp. 10, 11).

[25] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever. “Breast cancer histopathology image analysis: A review”. In: *IEEE Transactions on Biomedical Engineering* 61.5 (2014), pp. 1400–1411 (cit. on p. 11).

[26] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. “Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology”. In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2008, pp. 284–287 (cit. on p. 12).

[27] H. Fatakdawala, J. Xu, A. Basavanhally, et al. “Expectation–maximization-driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology”. In: *IEEE Transactions on Biomedical Engineering* 57.7 (2010), pp. 1676–1689 (cit. on p. 12).

[28] G. Li, T. Liu, J Nie, et al. “Segmentation of touching cell nuclei using gradient flow tracking”. In: *Journal of Microscopy* 231.1 (2008), pp. 47–58 (cit. on p. 12).

[29] M. N. Gurcan, J. Kong, O. Sertel, B. B. Cambazoglu, J. Saltz, and U. Catalyurek. “Computerized pathological image analysis for neuroblastoma prognosis”. In: *AMIA Annual Symposium Proceedings*. Vol. 2007. American Medical Informatics Association. 2007, p. 304 (cit. on p. 12).

[30] D. Comaniciu and P. Meer. “Cell image segmentation for diagnostic pathology”. In: *Advanced algorithmic approaches to medical image segmentation*. Springer, 2002, pp. 541–558 (cit. on p. 12).

[31] N. Signolle, M. Revenu, B. Plancoulaine, and P. Herlin. “Wavelet-based multiscale texture segmentation: Application to stromal compartment characterization on virtual slides”. In: *Signal Processing* 90.8 (2010), pp. 2412–2422 (cit. on p. 12).

[32] K. Mosaliganti, F. Janoos, O. Irfanoglu, et al. “Tensor classification of N-point correlation function features for histology tissue segmentation”. In: *Medical image analysis* 13.1 (2009), pp. 156–166 (cit. on p. 12).

[33] O. Sertel, U. V. Catalyurek, H. Shimada, and M. N. Gurcan. “Computer-aided prognosis of neuroblastoma: Detection of mitosis and karyorrhexis cells in digitized histological images”. In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2009, pp. 1433–1436 (cit. on p. 12).

[34] A. N. Basavanhally, S. Ganesan, S. Agner, et al. “Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology”. In: *IEEE Transactions on biomedical engineering* 57.3 (2009), pp. 642–653 (cit. on p. 12).

[35] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee. “Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data”. In: *Computerized medical imaging and graphics* 35.7-8 (2011), pp. 506–514 (cit. on p. 12).

[36] O. Sertel, G. Lozanski, A. Shana’ah, and M. N. Gurcan. “Computer-aided detection of centroblasts for follicular lymphoma grading using adaptive likelihood-based cell segmentation”. In: *IEEE Transactions on Biomedical Engineering* 57.10 (2010), pp. 2613–2616 (cit. on p. 12).

[37] S. J. Keenan, J. Diamond, W Glenn McCluggage, et al. “An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN)”. In: *The Journal of pathology* 192.3 (2000), pp. 351–362 (cit. on p. 13).

[38] A. Tabesh, M. Teverovskiy, H.-Y. Pang, et al. “Multifeature prostate cancer diagnosis and Gleason grading of histological images”. In: *IEEE transactions on medical imaging* 26.10 (2007), pp. 1366–1378 (cit. on p. 13).

[39] P. Khurd, C. Bahlmann, P. Maday, et al. “Computer-aided Gleason grading of prostate cancer histopathological images using texton forests”. In: *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2010, pp. 636–639 (cit. on p. 13).

[40] D. Gancberg, T. Järvinen, A. Di Leo, et al. “Evaluation of HER-2/NEU protein expression in breast cancer by immunohistochemistry: an interlaboratory study assessing the reproducibility of HER-2/NEU testing”. In: *Breast cancer research and treatment* 74.2 (2002), pp. 113–120 (cit. on p. 13).

[41] M.-Y. C. Polley, S. C. Leung, L. M. McShane, et al. “An international Ki67 reproducibility study”. In: *Journal of the National Cancer Institute* 105.24 (2013), pp. 1897–1906 (cit. on p. 13).

[42] A. C. Wolff, M. E. H. Hammond, J. N. Schwartz, et al. “American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer”. In: *Archives of pathology & laboratory medicine* 131.1 (2007), pp. 18–43 (cit. on p. 13).

[43] L. Cooper, O. Sertel, J. Kong, G. Lozanski, K. Huang, and M. Gurcan. “Feature-based registration of histopathology images with different stains: An application for computerized follicular lymphoma prognosis”. In: *Computer methods and programs in biomedicine* 96.3 (2009), pp. 182–192 (cit. on p. 13).

[44] M. Bello, A. Can, and X. Tao. “Accurate registration and failure detection in tissue micro array images”. In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2008, pp. 368–371 (cit. on p. 13).

[45] M. G. Giacomelli, L. Husvogt, H. Vardeh, et al. “Virtual hematoxylin and eosin transillumination microscopy using epi-fluorescence imaging”. In: *PLoS One* 11.8 (2016), e0159337 (cit. on pp. 13, 57).

[46] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134 (cit. on pp. 13, 56, 57).

[47] A. Madabhushi and G. Lee. “Image analysis and machine learning in digital pathology: Challenges and opportunities”. In: *Medical image analysis* 33 (2016), pp. 170–175 (cit. on p. 14).

[48] K. W. Foster. “Medical education in the digital age: Digital whole slide imaging as an e-learning tool”. In: *Journal of pathology informatics*. 2010 (cit. on p. 14).

[49] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323 (1986), pp. 533–536 (cit. on p. 17).

[50] S. Ruder. “An overview of gradient descent optimization algorithms”. In: *ArXiv* abs/1609.04747 (2016) (cit. on p. 17).

[51] H. Robbins and S. Monro. “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), 400–407 (cit. on p. 17).

[52] N. Qian. “On the momentum term in gradient descent learning algorithms”. In: *Neural networks : the official journal of the International Neural Network Society* 12 1 (1999), pp. 145–151 (cit. on p. 17).

[53] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2014) (cit. on p. 17).

[54] S. B. Kotsiantis. “Supervised Machine Learning: A Review of Classification Techniques”. In: *Informatica (Slovenia)* 31 (2007), pp. 249–268 (cit. on p. 17).

[55] H. Barlow. “Unsupervised Learning”. In: *Neural Computation* 1.3 (1989), 295–311 (cit. on p. 19).

[56] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. “The "wake-sleep" algorithm for unsupervised neural networks.” In: *Science* 268 5214 (1995), pp. 1158–61 (cit. on p. 19).

[57] Y. Bengio, A. C. Courville, and P. Vincent. “Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives”. In: *ArXiv* abs/1206.5538 (2012) (cit. on p. 19).

[58] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. “Extracting and composing robust features with denoising autoencoders”. In: *ICML ’08*. 2008 (cit. on p. 19).

[59] Y. LeCun, Y. Bengio, and G. Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), 436–444 (cit. on p. 19).

[60] N. Aloysius and M. Geetha. “A review on deep convolutional neural networks”. In: *2017 International Conference on Communication and Signal Processing (ICCSP)* (2017), pp. 0588–0592 (cit. on p. 19).

[61] H.-J. Yoo. “Deep Convolution Neural Networks in Computer Vision: a Review”. In: 2015 (cit. on p. 19).

[62] W. Rawat and Z. Wang. “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review”. In: *Neural Computation* 29 (2017), pp. 2352–2449 (cit. on p. 19).

[63] M. Egmont-Petersen, D. de Ridder, and H. Handels. “Image processing with neural networks - a review”. In: *Pattern Recognition* 35 (2002), pp. 2279–2301 (cit. on p. 19).

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *NIPS*. 2012 (cit. on p. 20).

[65] V. Nair and G. E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *ICML*. 2010 (cit. on p. 20).

[66] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun. “Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), pp. 1–8 (cit. on p. 20).

[67] S. Ioffe and C. Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *ArXiv* abs/1502.03167 (2015) (cit. on pp. 20, 38).

[68] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. “Instance Normalization: The Missing Ingredient for Fast Stylization”. In: *ArXiv* abs/1607.08022 (2016) (cit. on pp. 20, 58).

[69] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. “Improving neural networks by preventing co-adaptation of feature detectors”. In: *ArXiv* abs/1207.0580 (2012) (cit. on p. 20).

[70] A. Radford, L. Metz, and S. Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. In: *CoRR* abs/1511.06434 (2015) (cit. on p. 21).

[71] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. “Generative Adversarial Nets”. In: *NIPS*. 2014 (cit. on p. 21).

[72] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein Generative Adversarial Networks”. In: *ICML*. 2017 (cit. on p. 21).

[73] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. “Unrolled Generative Adversarial Networks”. In: *ArXiv* abs/1611.02163 (2016) (cit. on p. 22).

[74] A. Serag, A. Ion-Margineanu, H. Qureshi, et al. “Translational AI and Deep Learning in Diagnostic Pathology”. In: *Front. Med.* 2019 (cit. on p. 23).

[75] D. Komura and S. Ishikawa. “Machine Learning Methods for Histopathological Image Analysis”. In: *Computational and structural biotechnology journal*. 2018 (cit. on p. 23).

[76] Y. Liu, K. Gadepalli, M. Norouzi, et al. “Detecting Cancer Metastases on Gigapixel Pathology Images”. In: *ArXiv* abs/1703.02442 (2017) (cit. on pp. 23, 24, 34, 41).

[77] T. Wollmann and K. Rohr. “Automatic breast cancer grading in lymph nodes using a deep neural network”. In: *ArXiv* abs/1707.07565 (2017) (cit. on p. 23).

[78] A. Foucart, O. Debeir, and C. Decaestecker. “SNOW: Semi-Supervised, Noisy And/Or Weak Data For Deep Learning In Digital Pathology”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (2019), pp. 1869–1872 (cit. on p. 23).

[79] A. Mahendran and A. Vedaldi. “Understanding deep image representations by inverting them”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 5188–5196 (cit. on p. 24).

[80] L. A. Gatys, A. S. Ecker, and M. Bethge. “Texture and art with deep neural networks”. In: *Current Opinion in Neurobiology* 46 (2017), pp. 178–186 (cit. on p. 24).

[81] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. “SmoothGrad: removing noise by adding noise”. In: *ArXiv* abs/1706.03825 (2017) (cit. on p. 24).

[82] K. Simonyan, A. Vedaldi, and A. Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *CoRR* abs/1312.6034 (2013) (cit. on p. 24).

[83] D. Erhan, Y. Bengio, A. C. Courville, and P. Vincent. “Visualizing Higher-Layer Features of a Deep Network”. In: 2009 (cit. on p. 24).

[84] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. “Experimental perspectives on learning from imbalanced data”. In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 935–942 (cit. on p. 24).

[85] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357 (cit. on p. 24).

[86] I. Mani and I Zhang. “kNN approach to unbalanced data distributions: a case study involving information extraction”. In: *Proceedings of workshop on learning from imbalanced datasets*. Vol. 126. 2003 (cit. on p. 24).

[87] M. Kubat, S. Matwin, et al. “Addressing the curse of imbalanced training sets: one-sided selection”. In: *Icml*. Vol. 97. Nashville, USA. 1997, pp. 179–186 (cit. on p. 24).

[88] Y.-A. Chung, H.-T. Lin, and S.-W. Yang. “Cost-aware pre-training for multiclass cost-sensitive deep learning”. In: *arXiv preprint arXiv:1511.09337* (2015) (cit. on p. 24).

[89] V. Raj, S. Magg, and S. Wermter. “Towards effective classification of imbalanced data with convolutional neural networks”. In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer. 2016, pp. 150–162 (cit. on p. 24).

[90] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy. “Training deep neural networks on imbalanced data sets”. In: *2016 international joint conference on neural networks (IJCNN)*. IEEE. 2016, pp. 4368–4374 (cit. on p. 24).

[91] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. “Cost-sensitive learning of deep feature representations from imbalanced data”. In: *IEEE transactions on neural networks and learning systems* 29.8 (2017), pp. 3573–3587 (cit. on p. 24).

[92] A. Janowczyk and A. Madabhushi. “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases”. In: *Journal of pathology informatics*. 2016 (cit. on p. 24).

[93] A. M. Khan, N. M. Rajpoot, D. Treanor, and D. R. Magee. “A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution”. In: *IEEE Transactions on Biomedical Engineering* 61 (2014), pp. 1729–1738 (cit. on p. 25).

[94] F. Ciompi, O. Geessink, B. E. Bejnordi, 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)* (2017), pp. 160–163 (cit. on p. 25).

[95] H. Cho, S. Lim, G. Choi, and H. Min. “Neural Stain-Style Transfer Learning using GAN for Histopathological Images”. In: *ArXiv* abs/1710.08543 (2017) (cit. on p. 25).

[96] B. E. Bejnordi, G. J. S. Litjens, N. Timofeeva, et al. “Stain Specific Standardization of Whole-Slide Histopathological Images”. In: *IEEE Transactions on Medical Imaging* 35 (2016), pp. 404–415 (cit. on p. 25).

[97] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni. “Staingan: Stain style transfer for digital histological images”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 953–956 (cit. on p. 25).

[98] K. Sirinukunwattana, S. e Ahmed Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot. “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images”. In: *IEEE transactions on medical imaging* 35 5 (2016), pp. 1196–1206 (cit. on pp. 25, 37).

[99] H. Wu, J. H. Phan, A. K. Bhatia, C. A. Cundiff, B. M. Shehata, and M. D. Wang. “Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015), pp. 727–730 (cit. on p. 25).

[100] S. Kothari, J. H. Phan, and M. D. Wang. “Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade”. In: *Journal of pathology informatics*. 2013 (cit. on p. 25).

[101] M. Nalisnik, M. Amgad, S. Lee, et al. “Interactive Phenotyping Of Large-Scale Histology Imaging Data With HistomicsML”. In: 2017 (cit. on pp. 25, 85).

[102] S. Doyle, J. Monaco, M. D. Feldman, J. E. Tomaszeweski, and A. Madabhushi. “An active learning based classification strategy for the minority class problem: application to histopathology annotation”. In: *BMC Bioinformatics*. 2010 (cit. on pp. 25, 85).

[103] R. K. Padmanabhan, V. H. Somasundar, S. D. Griffith, et al. “An Active Learning Approach for Rapid Characterization of Endothelial Cells in Human Tumors”. In: *PloS one*. 2014 (cit. on p. 25).

[104] T. Brunye, P. A. Carney, K. H. Allison, L. G. Shapiro, D. L. Weaver, and J. G. Elmore. “Eye Movements as an Index of Pathologist Visual Expertise: A Pilot Study”. In: *PloS one*. 2014 (cit. on p. 25).

[105] V. Raghunath, M. O. Braxton, S. A. Gagnon, et al. “Mouse cursor movement and eye tracking data as an indicator of pathologists’ attention when viewing digital whole slide images”. In: *Journal of pathology informatics*. 2012 (cit. on p. 25).

[106] M. W. Lafarge, J. P. W. Pluim, K. A. J. Eppenhof, P. Moeskops, and M. Veta. “Domain-Adversarial Neural Networks to Address the Appearance Variability of Histopathology Images”. In: *ArXiv* abs/1707.06183 (2017) (cit. on p. 25).

[107] Y. Xu, Z. Jia, L. Wang, et al. “Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features”. In: *BMC Bioinformatics*. 2017 (cit. on p. 26).

[108] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh. “Convolutional neural networks for histopathology image classification: Training vs. Using pre-trained networks”. In: *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)* (2017), pp. 1–6 (cit. on p. 26).

[109] N. Bayramoglu and J. Heikkilä. “Transfer Learning for Cell Nuclei Classification in Histopathology Images”. In: *ECCV Workshops*. 2016 (cit. on p. 26).

[110] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li. “Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model”. In: *Scientific Reports*. 2017 (cit. on p. 26).

[111] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu. “Weakly supervised histopathology cancer image segmentation and classification”. In: *Medical image analysis* 18 3 (2014), pp. 591–604 (cit. on p. 26).

[112] Y. Xu, Y. Li, Z. Shen, et al. “Parallel multiple instance learning for extremely large histopathology image analysis”. In: *BMC Bioinformatics*. 2017 (cit. on p. 26).

[113] N. Ing, J. M. Tomczak, E. J. Miller, et al. “A deep multiple instance model to predict prostate cancer metastasis from nuclear morphology”. In: 2018 (cit. on p. 26).

[114] Z. Jia, X. Huang, E. I.-C. Chang, and Y. Xu. “Constrained Deep Weak Supervision for Histopathology Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 36 (2017), pp. 2376–2388 (cit. on p. 26).

[115] Y. Song, L. Zhang, S. Chen, D. Ni, B. Y. Lei, and T. Wang. “Accurate Segmentation of Cervical Cytoplasm and Nuclei Based on Multiscale Convolutional Network and Graph Partitioning”. In: *IEEE Transactions on Biomedical Engineering* 62 (2015), pp. 2421–2433 (cit. on p. 26).

[116] D. Romo, J. D. García-Arteaga, P. Arbeláez, and E. Romero. “A discriminant multi-scale histopathology descriptor using dictionary learning”. In: *Medical Imaging*. 2014 (cit. on p. 26).

[117] S. Doyle, A. Madabhushi, M. D. Feldman, and J. E. Tomaszeweski. “A Boosting Cascade for Automated Detection of Prostate Cancer from Digitized Histology”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention: MICCAI* 9 Pt 2 (2006), pp. 504–11 (cit. on p. 26).

[118] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13 (2004), pp. 600–612 (cit. on p. 28).

[119] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey. “Complex Wavelet Structural Similarity: A New Image Similarity Index”. In: *IEEE Transactions on Image Processing* 18 (2009), pp. 2385–2401 (cit. on pp. 29, 71).

[120] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever. “Corrections to "Breast cancer histopathology image analysis: A review"[May 14 1400-1411]”. In: *IEEE Trans. Biomed. Eng.* 61.11 (2014), pp. 2819–2819 (cit. on p. 33).

[121] R. Bhargava and A. Madabhushi. “Emerging themes in image informatics and molecular analysis for digital pathology”. In: *Annu. Rev. Biomed. Eng.* 18 (2016), pp. 387–412 (cit. on p. 33).

[122] J. S. Meyer, C. Alvarez, C. Milikowski, et al. “Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index”. In: *Mod. Pathol.* 18.8 (2005), p. 1067 (cit. on p. 33).

[123] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber. “Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation”. In: *NIPS*. 2015, pp. 2998–3006 (cit. on p. 34).

[124] W. Zhang, R. Li, H. Deng, et al. “Deep convolutional neural networks for multi-modality isointense infant brain image segmentation”. In: *NeuroImage* 108 (2015), pp. 214–224 (cit. on p. 34).

[125] A. Birenbaum and H. Greenspan. “Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks”. In: *Deep Learn Data Label Med Appl*. Springer, 2016, pp. 58–67 (cit. on p. 34).

[126] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam. “Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation”. In: *TMI* 35.5 (2016), pp. 1229–1239 (cit. on p. 34).

[127] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. “HeMIS: Hetero-modal image segmentation”. In: *MICCAI*. Springer. 2016, pp. 469–477 (cit. on p. 34).

[128] K. Kamnitsas, C. Ledig, V. F. Newcombe, et al. “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation”. In: *Med Image Anal* 36 (2017), pp. 61–78 (cit. on p. 34).

[129] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. “Brain tumor segmentation using convolutional neural networks in MRI images”. In: *TMI* 35.5 (2016), pp. 1240–1251 (cit. on p. 34).

[130] L. Zhao and K. Jia. “Multiscale cnns for brain tumor segmentation and diagnosis”. In: *Comput Math Methods Med* 2016 (2016) (cit. on p. 34).

[131] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. “Deep neural networks segment neuronal membranes in electron microscopy images”. In: *NIPS*. 2012, pp. 2843–2851 (cit. on p. 34).

[132] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *MICCAI*. Springer. 2015, pp. 234–241 (cit. on p. 34).

[133] H. R. Roth, A. Farag, L. Lu, E. B. Turkbey, and R. M. Summers. “Deep convolutional networks for pancreas segmentation in CT imaging”. In: *Medical Imaging 2015: Image Processing*. Vol. 9413. SPIE. 2015, 94131G (cit. on p. 34).

[134] A. Prasoon, 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: *MICCAI*. Springer. 2013, pp. 246–253 (cit. on p. 34).

[135] A. Janowczyk, S. Doyle, H. Gilmore, and A. Madabhushi. “A resolution adaptive deep hierarchical (RADHicaL) learning scheme applied to nuclear segmentation of digital pathology images”. In: *Comput Methods Biomech Biomed Eng Imaging Vis* 6.3 (2018), pp. 270–276 (cit. on p. 35).

[136] F. Xing, Y. Xie, and L. Yang. “An automatic learning-based framework for robust nucleus segmentation”. In: *TMI* 35.2 (2016), pp. 550–566 (cit. on p. 35).

[137] Y. Song, E.-L. Tan, X. Jiang, et al. “Accurate cervical cell segmentation from overlapping clumps in pap smear images”. In: *TMI* 36.1 (2017), pp. 288–300 (cit. on p. 35).

[138] K. Sirinukunwattana, J. P. Pluim, H. Chen, et al. “Gland segmentation in colon histology images: The glas challenge contest”. In: *Med Image Anal* 35 (2017), pp. 489–502 (cit. on p. 35).

[139] A. BenTaieb and G. Hamarneh. “Topology aware fully convolutional networks for histology gland segmentation”. In: *MICCAI*. Springer. 2016, pp. 460–468 (cit. on p. 35).

[140] A. BenTaieb, J. Kawahara, and G. Hamarneh. “Multi-loss convolutional networks for gland analysis in microscopy”. In: *IEEE 13th ISBI*. IEEE. 2016, pp. 642–645 (cit. on p. 35).

[141] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng. “DCAN: Deep contour-aware networks for object instance segmentation from histology images”. In: *Med Image Anal* 36 (2017), pp. 135–146 (cit. on p. 35).

[142] W. Li, S. Manivannan, S. Akbar, J. Zhang, E. Trucco, and S. J. McKenna. “Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks”. In: *IEEE 13th ISBI*. IEEE. 2016, pp. 1405–1408 (cit. on p. 35).

[143] Y. Xu, Y. Li, M. Liu, et al. “Gland instance segmentation by deep multichannel side supervision”. In: *MICCAI*. Springer. 2016, pp. 496–504 (cit. on p. 35).

[144] P. Kainz, M. Pfeiffer, and M. Urschler. “Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation”. In: *arXiv preprint arXiv:1511.06919* (2015) (cit. on pp. 35, 37).

[145] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi. “A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images”. In: *Neurocomputing* 191 (2016), pp. 214–223 (cit. on p. 35).

[146] Y. Xie, Z. Zhang, M. Sapkota, and L. Yang. “Spatial clockwork recurrent neural network for muscle perimysium segmentation”. In: *MICCAI*. Springer. 2016, pp. 185–193 (cit. on p. 35).

[147] J. Wang, J. D. MacKenzie, R. Ramachandran, and D. Z. Chen. “A deep learning approach for semantic segmentation in histology tissue images”. In: *MICCAI*. Springer. 2016, pp. 176–184 (cit. on p. 35).

[148] G. Apou, N. S. Schaadt, B. Naegel, et al. “Detection of lobular structures in normal breast tissue”. In: *Comput. Biol. Med* 74 (2016), pp. 91–102 (cit. on p. 35).

[149] A. Cruz-Roa, A. Basavanhally, F. González, et al. “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks”. In: *Medical Imaging 2014: Digital Pathology*. Vol. 9041. SPIE. 2014, p. 904103 (cit. on p. 35).

[150] G. Litjens, C. I. Sánchez, N. Timofeeva, et al. “Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis”. In: *Sci Rep.* 6 (2016), p. 26286 (cit. on p. 35).

[151] H. Rezaeilouyeh, A. Mollahosseini, and M. H. Mahoor. “Microscopic medical image classification framework via deep learning and shearlet transform”. In: *J Med Imaging* 3.4 (2016), p. 044501 (cit. on p. 35).

[152] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. “Deep learning for identifying metastatic breast cancer”. In: *arXiv preprint arXiv:1606.05718* (2016) (cit. on pp. 35, 41).

[153] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. “Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images”. In: *TMI* 35.5 (2016), pp. 1313–1321 (cit. on p. 37).

[154] 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”. In: *IEEE Trans. Biomed. Eng* 61.6 (2014), pp. 1729–1738 (cit. on p. 37).

[155] A. M. Khan, H. El-Daly, E. Simmons, and N. M. Rajpoot. “HyMaP: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images”. In: *JPI* 4.Suppl (2013) (cit. on p. 37).

[156] A. A. Youssif, A. S. Gawish, and M. E. Moussa. “Automated periodontal diseases classification system”. In: *Editorial Preface* 3.1 (2012) (cit. on p. 37).

[157] T. Chen and C. Chefd’Hotel. “Deep learning based automatic immune cell detection for immunohistochemistry images”. In: *MLMI*. Springer. 2014, pp. 17–24 (cit. on p. 37).

[158] R. Duggal, A. Gupta, R. Gupta, and P. Mallick. “SD-Layer: Stain Deconvolutional Layer for CNNs in Medical Microscopic Imaging”. In: *MICCAI*. Springer. 2017, pp. 435–443 (cit. on pp. 37, 51).

[159] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on MICCAI*. Springer. 2015, pp. 234–241 (cit. on pp. 38, 39).

[160] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”. In: *J Mach Learn Res* 15.1 (2014), pp. 1929–1958 (cit. on p. 39).

[161] X. Li and K. N. Plataniotis. “A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics”. In: *IEEE Trans. Biomed. Eng.* 62.7 (2015), pp. 1862–1873 (cit. on p. 39).

[162] A. C. Ruifrok, D. A. Johnston, et al. “Quantification of histochemical staining by color deconvolution”. In: *Anal Quant Cytol Histol* 23.4 (2001), pp. 291–299 (cit. on p. 39).

[163] J. Xu, L. Xiang, G. Wang, et al. “Sparse Non-negative Matrix Factorization (SNMF) based color unmixing for breast histopathological image analysis”. In: *Comput Med Imaging Graph.* 46 (2015), pp. 20–29 (cit. on p. 39).

[164] D. Eigen and R. Fergus. “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture”. In: *ICCV*. 2015, pp. 2650–2658 (cit. on p. 39).

[165] V. Badrinarayanan, A. Kendall, and R. Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *arXiv preprint arXiv:1511.00561* (2015) (cit. on p. 44).

[166] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *CVPR*. 2016, pp. 770–778 (cit. on pp. 44, 66).

[167] 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”. In: *PloS one* 12.1 (2017), e0169875 (cit. on p. 51).

[168] B. Hu, N. El Hajj, S. Sittler, N. Lammert, R. Barnes, and A. Meloni-Ehrig. “Gastric cancer: Classification, histology and application of molecular pathology”. In: *JGO* 3.3 (2012), p. 251 (cit. on p. 56).

[169] M. Kuse, T. Sharma, and S. Gupta. “A classification scheme for lymphocyte segmentation in H&E stained histology images”. In: *Recognizing Patterns in Signals, Speech, Images and Videos*. Springer, 2010, pp. 235–243 (cit. on p. 56).

[170] H. Chang, J. Lu, F. Yu, and A. Finkelstein. “Pairedcyclegan: Asymmetric style transfer for applying and removing makeup”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 40–48 (cit. on p. 56).

[171] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: *ICCV*. 2017 (cit. on pp. 56, 57, 61, 71).

[172] L. A. Gatys, A. S. Ecker, and M. Bethge. “Image style transfer using convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2414–2423 (cit. on pp. 56, 59).

[173] Y. Taigman, A. Polyak, and L. Wolf. “Unsupervised cross-domain image generation”. In: *arXiv preprint arXiv:1611.02200* (2016) (cit. on p. 56).

[174] J. Johnson, A. Alahi, and L. Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *European conference on computer vision*. Springer. 2016, pp. 694–711 (cit. on pp. 56, 59).

[175] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. “Image reconstruction by domain-transform manifold learning”. In: *Nature* 555.7697 (2018), p. 487 (cit. on p. 56).

[176] S. Kaji and S. Kida. “Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging”. In: *Radiological Physics and Technology* (2019), pp. 1–14 (cit. on p. 56).

[177] J. M. Bini, J. Spain, K. S. Nehal, M. Rajadhyaksha, V. Hazelwood, and C. A. DiMarzio. “Confocal mosaicing microscopy of human skin ex vivo: spectral analysis for digital staining to simulate histology-like appearance”. In: *J. Biomed. Opt.* 16.7 (2011), p. 076008 (cit. on p. 57).

[178] D. S. Gareau. “Feasibility of digitally stained multimodal confocal mosaics to simulate histopathology”. In: *J. Biomed. Opt.* 14.3 (2009), p. 034050 (cit. on p. 57).

[179] J. Dobbs, S. Krishnamurthy, M. Kyrish, A. P. Benveniste, W. Yang, and R. Richards-Kortum. “Confocal fluorescence microscopy for rapid evaluation of invasive tumor cellularity of inflammatory breast carcinoma core needle biopsies”. In: *Breast Cancer Res. Treat.* 149.1 (2015), pp. 303–310 (cit. on p. 57).

[180] G. Nadarajan, T. Hope, D. Wang, et al. “Automated multi-class ground-truth labeling of H&E images for deep learning using multiplexed fluorescence microscopy”. In: *Medical Imaging 2019: Digital Pathology*. Vol. 10956. International Society for Optics and Photonics. 2019, 109560J (cit. on p. 57).

[181] A. Can, M. O. Bello, M. J. Gerdes, and Q. Li. *System and methods for mapping fluorescent images into a bright field color space*. US Patent 8,269,827. 2012 (cit. on p. 57).

[182] Y. K. Tao, D. Shen, Y. Sheikine, et al. “Assessment of breast pathologies using nonlinear microscopy”. In: *PNAS* 111.43 (2014), pp. 15304–15309 (cit. on p. 57).

[183] Y. Rivenson, H. Wang, Z. Wei, et al. “Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning”. In: *Nature biomedical engineering* 3.6 (2019), p. 466 (cit. on p. 57).

[184] Y. Rivenson, T. Liu, Z. Wei, Y. Zhang, K. de Haan, and A. Ozcan. “PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning”. In: *Light: Science & Applications* 8.1 (2019), p. 23 (cit. on p. 57).

[185] N. Bayramoglu, M. Kaakinen, L. Eklund, and J. Heikkilä. “Towards Virtual H&E Staining of Hyperspectral Lung Histology Images Using Conditional Generative Adversarial Networks.” In: *ICCV Workshops*. 2017, pp. 64–71 (cit. on p. 57).

[186] A. Rana, G. Yauney, A. Lowe, and P. Shah. “Computational Histological Staining and Destaining of Prostate Core Biopsy RGB Images with Generative Adversarial Neural Networks”. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2018, pp. 828–834 (cit. on p. 57).

[187] E. A. Burlingame, A. Margolin, J. W. Gray, and Y. H. Chang. “SHIFT: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks”. In: *Medical Imaging 2018: Digital Pathology*. Vol. 10581. SPIE. 2018, p. 1058105 (cit. on p. 57).

[188] E. M. Christiansen, S. J. Yang, D. M. Ando, et al. “In silico labeling: Predicting fluorescent labels in unlabeled images”. In: *Cell* 173.3 (2018), pp. 792–803 (cit. on p. 57).

[189] C. Ounkomol, S. Seshamani, M. M. Maleckar, F. Collman, and G. R. Johnson. “Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy”. In: *Nature methods* 15.11 (2018), p. 917 (cit. on p. 57).

[190] N. Borhani, A. J. Bower, S. A. Boppart, and D. Psaltis. “Digital staining through the application of deep neural networks to multi-modal multi-photon microscopy”. In: *Biomedical optics express* 10.3 (2019), pp. 1339–1350 (cit. on p. 57).

[191] M. Fujitani, Y. Mochizuki, S. Iizuka, et al. “Re-staining Pathology Images by FCNN”. In: *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE. 2019, pp. 1–6 (cit. on p. 57).

[192] M. Gadermayr, V. Appel, B. M. Klinkhammer, P. Boor, and D. Merhof. “Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 165–173 (cit. on p. 57).

[193] A. Kapil, T. Wiestler, S. Lanzmich, et al. “DASGAN–Joint Domain Adaptation and Segmentation for the Analysis of Epithelial Regions in Histopathology PD-L1 Images”. In: *arXiv preprint arXiv:1906.11118* (2019) (cit. on p. 57).

[194] Z. Xu, C. F. Moro, B. Bozóky, and Q. Zhang. “Gan-based virtual re-staining: A promising solution for whole slide image analysis”. In: *arXiv preprint arXiv:1901.04059* (2019) (cit. on p. 58).

[195] A. Royer, K. Bousmalis, S. Gouws, et al. “Xgan: Unsupervised image-to-image translation for many-to-many mappings”. In: *arXiv preprint arXiv:1711.05139* (2017) (cit. on p. 59).

[196] L. T. Li, G. Jiang, Q. Chen, and J. N. Zheng. “Ki67 is a promising molecular target in the diagnosis of cancer”. In: *Mol Med Rep* 11.3 (2015), pp. 1566–1572 (cit. on p. 65).

[197] S.-H. Lu, W.-S. Tsai, Y.-H. Chang, et al. “Identifying cancer origin using circulating tumor cells”. In: *Cancer Biol. Ther.* 17.4 (2016), pp. 430–438 (cit. on pp. 65, 70).

[198] A. Research. *AACR 2018 Proceedings: Abstracts 3028-5930*. CTI Meeting Technology, 2018 (cit. on p. 65).

[199] C. Li and M. Wand. “Precomputed real-time texture synthesis with markovian generative adversarial networks”. In: *ECCV*. Springer. 2016, pp. 702–716 (cit. on p. 66).

[200] J. Chen, X. Pan, R. Monga, S. Bengio, and R. Jozefowicz. “Revisiting distributed synchronous SGD”. In: *arXiv preprint arXiv:1604.00981* (2016) (cit. on pp. 66, 71).

[201] T. Kelly, Y. Huang, A. E. Simms, and A. Mazur. “Fibroblast activation protein-α: a key modulator of the microenvironment in multiple pathologies”. In: *Int Rev Cell Mol Biol*. Vol. 297. Elsevier, 2012, pp. 83–116 (cit. on pp. 69, 79).

[202] S. Umeyama. “Least-squares estimation of transformation parameters between two point patterns”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 4 (1991), pp. 376–380 (cit. on p. 71).

[203] T. de Bel, M. Hermsen, J. Kers, J. van der Laak, and G. Litjens. “Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology”. In: *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*. Vol. 102. 2018, pp. 151–163 (cit. on p. 72).

[204] M. L. McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica: Biochemia medica* 22.3 (2012), pp. 276–282 (cit. on p. 77).

[205] K. Gwet et al. “Inter-rater reliability: dependency on trait prevalence and marginal homogeneity”. In: *Statistical Methods for Inter-Rater Reliability Assessment Series* 2.1 (2002), p. 9 (cit. on p. 77).

[206] M. Peikari, J. T. Zubovits, G. M. Clarke, and A. L. Martel. “Clustering Analysis for Semi-supervised Learning Improves Classification Performance of Digital Pathology”. In: *MLMI*. 2015 (cit. on p. 85).

[207] A. Vezhnevets, J. M. Buhmann, and V. Ferrari. “Active learning for semantic segmentation with expected change”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 3162–3169 (cit. on p. 85).

[208] B. Settles. “Active Learning Literature Survey”. In: 2009 (cit. on p. 85).

List of Figures

List of Tables