UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE MATHÉMATIQUES, SCIENCES DE

L'INFORMATION ET DE L'INGÉNIEUR

Laboratoire ICube – UMR 7357

THÈSE présentée par :

[Jelica Zlatko VASILJEVIĆ]

soutenue le : **22 Septembre 2022**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Informatique

Generative Adversarial Networks in Digital Histopathology: Stain Transfer and Deep Learning Model Invariance to Stain Variation

THÈSE dirigée par :

M. WEMMERT Cédric	Professeur, Université de Strasbourg, France
M. STANKOVIĆ Srdjan	Professeur Émérite, Université de Belgrade, Serbie

RAPPORTEURS :

M. DESCOMBES Xavier	Directeur de Recherche, INRIA Sophia Antipolis, France
Mme. TEMERINAC OTT Maja	Professeur, Hochschule Furtwangen, Germany

---

AUTRES MEMBRES DU JURY :

Mme. MERVEILLE Odyssée	Maître de Conférences, INSA Lyon, France
Mme. LECLERC Sarah	Maître de Conférences, Université de Bourgogne, France
M. LAMPERT Thomas (invité)	Chaire Industrielle Sciences des Données et Intelligence Artificielle, Université de Strasbourg, France

Jelica Zlatko VASILJEVIĆ

Generative Adversarial Networks in Digital Histopathology: Stain Transfer and Deep Learning Model Invariance to Stain Variation

Résumé

L'histopathologie numérique est un domaine d'innovation très riche, tant dans les applications cliniques que dans la recherche, où les solutions basées sur l'apprentissage profond connaissent un succès remarquable. Cependant, les méthodes actuelles d'apprentissage profond sont des approches gourmandes en données qui nécessitent d'énormes bases de données annotées pour obtenir des modèles performants. Or, le domaine médical est connu pour sa difficulté à obtenir des données et des annotations - la collecte de données relève d'une réglementation stricte et contraignante, tandis que seuls des experts peuvent effectuer des annotations de haute qualité, ce qui est un processus laborieux et coûteux. De plus, compte tenu des variations qui peuvent se produire en raison du processus et des protocoles de coloration, les données déjà collectées et annotées ne peuvent être réutilisées qu'avec un succès limité. Une telle variation de la coloration représente un changement de domaine et affecte considérablement les solutions basées sur l'apprentissage profond dans la pratique. Cela devient plus évident encore lorsque l'apprentissage se focalise sur des structures biologiques visibles avec plusieurs colorations, car les solutions développées en utilisant les données d'une coloration sont susceptibles d'échouer lorsqu'elles sont appliquées à une autre. Cette thèse étudie le potentiel des réseaux adversaires génératifs (GAN) dans deux directions pour résoudre ces problèmes - le transfert de colorations pour permettre la réutilisation de bases de données déjà disponibles et le développement de modèles invariants aux colorations qui réduiraient le besoin d'acquisition de données ou d'annotations supplémentaires. L'application principale de la thèse est la segmentation des glomérules en pathologie rénale avec de multiples colorations.

Résumé en anglais

Digital histopathology has become a rich area of innovation in both clinical application and research, where deep-learning-based solutions have remarkable success. However, current state-of-the-art deep learning methods are data-hungry approaches which require huge, annotated data collections to perform well. Nevertheless, the medical domain is known for its scarcity of data and annotations — collecting data falls under strict low regulations while experts only can perform high-quality annotations, which is a laborious and expensive process. Moreover, considering the variations that can occur due to the staining process and staining protocols, already collected and annotated datasets can only be reused with limited success. Such stain variation represents a source of domain shift and significantly affects deep learning-based solutions in practice. This becomes more evident when a deep learning task tackles problems related to structures visible under multiple stains as solutions developed using the data from one staining are likely to fail when applied to the other. This thesis investigates the potential of Generative Adversarial Networks (GANs) in two directions for addressing these problems — stain transfer to enable reusing already available data collections; and developing stain invariant solutions which would alleviate the need for additional data acquisition or annotations. The application focus of the thesis is glomeruli segmentation in renal pathology with multiple stainings.

Université de Strasbourg

École doctorale MSII

Generative Adversarial Networks in Digital Histopathology: Stain Transfer and Deep Learning Model Invariance to Stain Variation

THÈSE
pour obtenir le grade de

Doctorat de l'Université de Strasbourg

mention Informatique

Soutenue le 22 septembre 2022

présentée par

Jelica Zlatko VASILJEVIĆ

Composition du jury

Directeur de thèse:	Cédric WEMMERT, Professor, University of Strasbourg, France
Co-encadrant:	Srdjan STANKOVIĆ, Professor Emeritus, University of Belgrade, Serbia
Rapporteurs:	Xavier DESCOMBES, Research Director, INRIA Sophia Antipolis, France Maja TEMERINAC OTT, Professor, Hochschule Furtwangen, Germany
Examinateurs:	Odyssée MERVEILLE, Lecturer, INSA Lyon, France Sarah LECLERC, Lecturer, University of Bourgogne, France
Invité :	Thomas LAMPERT, Chair of Data Science and AI, University of Strasbourg, France

Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie
ICUBE — UMR 7357

Acknowledgements

Finalising this thesis marks the end of an extraordinary period in my life. The years I have spent in France have challenged me in many ways. Nevertheless, I firmly believe that they've made me a better person. This wouldn't have been possible without the many people who have encouraged and supported me and to whom I would like to express my deepest gratitude.

I am extremely grateful to my French supervisors, Professor Cédric Wemmert and Thomas Lampert, for exceptional guidance during my PhD, especially for letting me be a "researcher" at the beginning of my PhD, being open to my ideas and letting me learn from my own mistakes. I would like to sincerely thank my Serbian supervisor, Professor Srdjan Stanković, for his paternal care, his commitment to my progress and his willingness to fight various windmills with me. I also had the great pleasure of working with Professor Friedrich Feuerhake, who helped me understand complex medical concepts and provided excellent constructive criticism on all my work. I would also like to thank my lab colleague Zeeshan Nisar, with whom I had fantastic collaboration and productive discussions. Working in such a supportive and encouraging environment was very important for me.

However, this endeavour wouldn't have been possible without Thomas Lampert, who trusted me from my very beginnings in the research, pushed me forward and strongly shaped my approach to the research. He helped me get through the first (re)submissions, rejections, many negative results and administrative difficulties. He should be awarded a special medal for his patience and tolerance towards my chaotic ideas (in the order of thousands!) and his fight for all the GPU hours and terabytes that were crucial for this work. I owe him a huge debt of gratitude for all the opportunities that helped me grow as a researcher - hackathons, lessons, seminars, summer schools, conferences... and of course for teaching me English grammar (or at least for trying to) and endless text corrections (especially articles, which are always somehow misplaced in my texts!). I would also like to thank Tatjana Aleksic Lampert for her encouragement and support. Finally, I would like to thank both of them for their support during the difficult Covid period.

Nevertheless, I couldn't have undertaken this journey without my family, especially my parents Ankica and Zlatko, who truly believe in me and give me enormous support and unconditional love. Since my earliest childhood, they have put great efforts to make an amazing environment where I was able to explore and develop all my talents. This opened my mind to many possible fields, from poetry and maths

to car repair and cooking, which greatly influenced and built my research spirit. I truly believe that this must have required superhuman strength in the tough post-war years. One such magic trick was spending half of a family budget to buy the computer, which I used to take my first steps in computer science! And here I am now, writing the acknowledgements for my dissertation in the same room where I wrote the first programme in Pascal! Without such steps, I wouldn't be able to learn what really matters in life - family and knowledge. In this sense, I would like to express a big gratitude to my brothers Srdjan, Pavle and Mihajlo, as well as my sisters-in-law, Vanja, Andjela and Milja, who support me in every possible way and give me the warmest love. Words cannot express my gratitude to my grandmother Ljubica, who is my inspiration and motivation and whose vast life experience and advice have helped me see the big picture and crystallise my own goals - Bakickon, hvala!

Apart from my family, I found great support in my close friends Darko Antonijević, Simona Vulović and Jelena Mitrović, who had a remarkable influence on my decisions, my spirit and my strength. Being separated from my country, my language and my family, I would like to thank Bojana Marković, Olja and Nemanja Alabić, Vuk Vuković, Damjan and Ema Marković and Katarina Bačević for being my big Serbian family abroad. From hours of discussions to amazing excursions (Damjan is to thank for the Alsace discovery trips!) and wonderful food (in usually highly improvised dishes) - all this made my life outside the lab seem like a never-ending fairy tale. I would especially like to thank my friends, colleagues and fellow travellers on this PhD trip - Mihailo and Aleksandra Obrenović, with whom I took the first international and research steps. Without them, everything would have been different (and much more stressful)! I owe very special moments to Sarah Zenasni, who became my closest friend and accomplice in many, many crazy things (breaking all kinds of traffic rules to be under the Eiffel Tower at midnight on my birthday is priceless!). Her support is simply unforgettable! I am also grateful for the many wonderful people I had the pleasure to meet: Hajer Akid, Nadja Groysbeck, Julián Martín del Fiore, Melanie Castagno, Sebastian Sampayo and Antonella Zerpa, with whom I have become very good friends. Last but not least, I would like to thank Anthony Fischer for the enormous love, patience and support he has given me over the last two years of my PhD! Merci beaucoup (read with a Serbian accent)! Besides the PhD, I consider meeting these wonderful people as one of the greatest accomplishments I have achieved in France.

Finally, I would like to thank the institutions that supported me during my PhD - the French government for the scholarship for the cotutelle studies, the Fund for Young Talents of the Republic of Serbia (Dositeja) for the scholarships for study abroad and the Faculty of Science at the University of Kragujevac, notably Institute for Mathematics and Informatics, for the support during my international PhD, for being great colleagues and making an environment to which is always a pleasure to return.

”Сви ми дугујемо себе некој деци, будимо потомци да би били преци” (Брана Црнчевић)

Посвећено мојим родитељима и њиховим напорима да ме изведу на прави пут.

Abstract

The deep learning revolution opens the door for remarkable applications of artificial intelligence in the medical domain. Many everyday clinical tasks have great potential to be fully automated, which has triggered a staggering amount of research. In such an environment, digital histopathology is not an exception. However, current state-of-the-art deep learning methods are data-hungry approaches which require huge annotated data collections to perform well. Nevertheless, digital histopathology, like the other fields of the medical domain, is known for its scarcity of data and annotations - collecting data falls under strict law regulations while experts only can perform high-quality annotations, which is a laborious and expensive process. Moreover, considering the variations that can occur due to the staining process and staining protocols, already collected and annotated datasets can only be reused with limited success. Such stain variation represents a source of domain shift and significantly affects deep learning-based solutions in practice. This thesis investigates the potential of Generative Adversarial Networks (GANs) in two directions for addressing these problems — stain transfer to enable reusing already available data collections; and developing stain invariant solutions which would alleviate the need for additional data acquisition or annotations.

The application focus of the thesis is glomeruli segmentation in renal pathology with multiple stainings. The thesis proposes the usage of GAN-based methods for the stain transfer between multiple stainings and gives extensive discussion related to the limitations of these methods and potentially misleading results that can occur. Some of the observed limitations lead to the discovery of hidden noise which encodes stain-related characteristics. This finding is further exploited to propose an unsupervised augmentation strategy which has a positive effect on model performances in a supervised setting, even with a limited number of annotated samples. Moreover, the thesis introduces the first method that encourages empirical stain invariance whose benefits are demonstrated on numerous stainings, including some unseen. Furthermore, the thesis proposes HistoStarGAN, the first end-to-end trainable model that simultaneously performs stain transfer, stain normalisation and stain invariant segmentation. HistoStarGAN model is able to obtain diverse translations between multiple stainings at the same time and to generalise to unseen stainings as well. This property is exploited to generate an artificially created fully annotated dataset, KidneyArtPathology, which will be released to encourage the progress of deep learning-based solutions in the field of renal pathology.

List of Publications During PhD

International Journals

[a] **Jelica Vasiljević**, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks. Neurocomputing 460: 277-291, 2021.

[b] **Jelica Vasiljević**, Zeeshan Nisar, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. CycleGAN for virtual stain transfer: is seeing really believing? Artificial Intelligence in Medicine 133, 2022.

[c] **Jelica Vasiljević**, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. HistoStarGAN: A Unified Approach to Stain Normalisation, Stain Transfer and Stain Invariant Segmentation in Renal Histopathology, Under Review.

International Conferences

[a] **Jelica Vasiljević**, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Self Adversarial Attack as an Augmentation Method for Immunohistochemical Stainings. In Proceedings of the International Symposium on Biomedical Imaging (ISBI), pp. 1939-1943, 2021.

[b] Zeeshan Nisar, **Jelica Vasiljević**, Pierre Gançarski, and Thomas Lampert. Towards Measuring Domain Shift in Histopathological Stain Translation in an Unsupervised Manner, In Proceedings of the International Symposium on Biomedical Imaging (ISBI), 2022.

Contents

Chapter 1

Introduction

It is only with the heart that one can see rightly; what is essential is invisible to the eye.

Antoine de Saint-Exupéry, The Little Prince

"Let me start by saying a few things that seem obvious. I think if you work as a radiologist, you are like the coyote that's already over the edge of the cliff but has not yet looked down, so he does not know there is no ground underneath him. People should stop training radiologists now. It is just completely obvious that within five years, deep learning is going to do better than radiologists because it's going to be able to get a lot more experience. It might be ten years, but we have got plenty of radiologists already. I said this at a hospital, and it did not go down too well." [21] stated Geoffrey Hinton in 2016 at the Creative Destruction Lab (CDL) seminar on "Machine Learning and the Market for Intelligence" in Toronto, Canada. A few years later, in 2018, three prominent computer scientists, including Geoffrey Hinton (Yoshua Bengio, Geoffrey Hinton and Yann LeCun), received the prestigious Turing award "for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing". This Nobel Prize for computing crowns decades of research effort to model intelligence and make machines able to assist humans in various areas. Today, artificial intelligence solutions are part of everyday life, starting from personal assistants in smartphones to self-driving cars. Such outstanding abilities raise hope and fear that machines could replace many expert jobs in the near future. The most sensitive area that attracts a lot of attention is undoubtedly the medical domain.

The deep learning revolution [1] opens the door for remarkable artificial intelligence applications in the medical domain. Plenty of everyday clinical tasks have great potential to be fully automated, which triggered a staggering amount of research [2–4]. In such an environment, digital histopathology is not an exception. Nowadays, deep learning-based methods achieve outstanding results in various tasks such as cancer detection, disease classification, and transplant assessment [22]. In a specific experimental setting, these solutions are able to perform on par with ex-

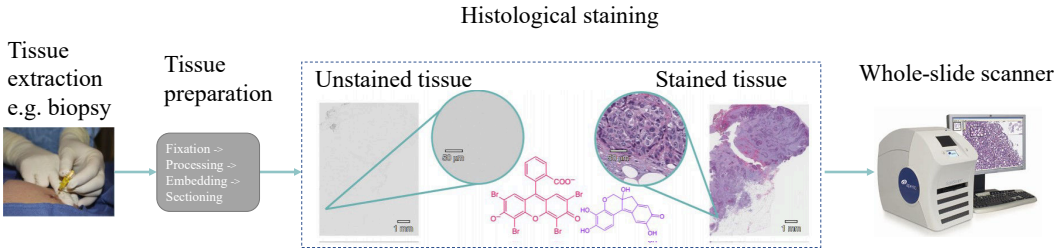


Figure 1.1: Illustration of the routine histological examination process [1]. Image credits to Rivenson et al. [32].

perienced pathologists [4]. Therefore, the same Hinton's statement could refer to pathologists as well.

Nevertheless, the clinical applications of deep learning-based solutions are still very limited [23]. The current understanding of applied methods gives strong evidence that it is very challenging to replace an expert's knowledge and experience with a machine and that tasks are more complex than they might first look [24, 25]. Additionally, theoretical advances in deep learning bring to the table some essential questions such as model bias towards training dataset, explainability, and interpretability [25–28]. In the context of the medical domain, further questions are raised, such as responsibility for a given diagnosis, patient privacy, and ethical issues [29, 30]. Thus, at the current state of deep learning, it seems that the path towards fully automatic medical expertise is long.

Recently, Generative Adversarial Networks (GANs) [5] bring new opportunities for deep learning in the field of digital histopathology. The demonstration of impressive results obtained using GANs in image generation already surpass the complexity of routine pathological examination. However, is seeing really believing?

This thesis represents a contribution to understanding what GANs can do and what they cannot do in digital histopathology, when and in which amount visually impressive results are trustworthy, and more importantly, in which situations such results can be misleading. Moreover, the thesis investigates the ways GANs can be used to build better deep learning models whose path to clinical practice can be shorter.

1.1 Digital Pathology

Histology (originates from Greek, *histos* — tissue + *logos* — science) is a branch of biology which studies the microscopic structures of healthy animal or human tissue. The microscopic study of changes that appear in the tissue as a consequence of disease (pathology) is known as histopathology [31]. Combined with other fields such as biochemistry and physiology, histological analysis is the gold standard in the diagnosis of many diseases.

Histopathological examination starts with the physical removal of a tissue sample from the body by biopsy or surgery. To be microscopically examined, the sample

[1] Image of the scanner taken from https://tmalab.jhmi.edu/scanning.html.

undergoes several preparation steps, as illustrated in Figure 1.1. Ideally, the preparation process should preserve structural features, so that tissue on the slide contains the same structures as in the body [33]. The preparation usually contains five basic stages: fixation, processing, embedding, sectioning and staining [7, 34], followed by scanning, which produces a digitalised version of the slide. As soon as possible, the extracted tissue is fixed using fixatives (e.g. formalin) to prevent decay. The tissue is further processed by dehydration, followed by clearing and infiltration, ending with embedding, usually by paraffin. At this stage, the hardened block with tissue and the surrounding embedding medium is placed in a microtome, an instrument for sectioning [33], which extracts very thin sections ($3-10\mu m$) that are placed on a glass slide. Once sections are obtained, they are colourless, and to be microscopically analysed they need to be stained (dyed).

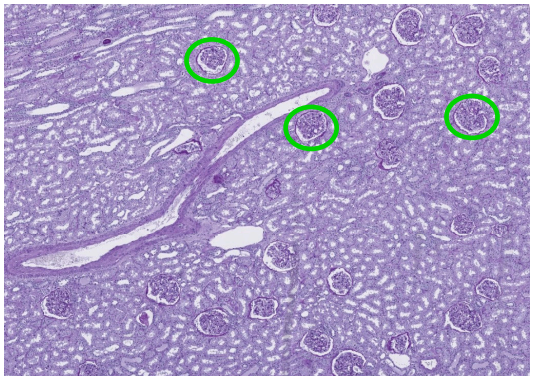
The staining process chemically introduces contrast into tissue sections, making visible particular tissue components or cells and enabling their microscopic analysis. The stains are intended to be selective by making different chemical compounds (e.g. acidic or basic) with tissue in the section. That way, different stains highlight different tissue components, enabling various analyses. For example, hematoxylin is a basic dye that binds acidic components such as cell nuclei resulting in a purplish-blue colour, while eosin is an acidic dye that enhances basic components such as cell stroma or cytoplasm by magenta-red colour. This combination of hematoxylin and eosin (H&E) is a routinely used staining to inspect general structures in the tissue. More information about the staining process is provided in Appendix A.

Stained glass slides, as physical objects, are fragile and prone to fading over time. Nowadays, a histological preparation process usually ends with whole-slide scanners that digitalise the glass slide. Whole-Slide Imaging (WSI) refers to the creation of a digital representation of a histological glass slide at a level of detail provided by a light microscope [35]. For example, a WSI, where each pixel corresponds to a square of $0.5\mu m$ ($0.25\mu m$) in the slide, is regarded as providing an equivalent level of detail as seen with a $\times 20$ ($\times 40$) objective of a high-quality microscope [35]. In this way, the examination can be performed physically far from a microscope, using a personal computer and specialised software for image manipulation. Moreover, multiple pathologists, not necessarily at the same physical location (e.g. hospital, city, country), can analyse the slide simultaneously.

The examination performed by pathologists largely consists of recognising specific patterns, e.g. detection of cancer cells and their distribution. Depending on the analysis, specific staining might be required as each highlights particular components. Thus, a common practice in histology is to stain multiple slices from the same biopsy with different stains, which provides experts with various information regarding tissue structure. Although different stains highlight different tissue components, some general analyses can be performed in multiple stainings. For example, in the case of kidney pathology, glomeruli [2] are visible under multiple stains. As illustrated in Figure 1.2, albeit different parts of glomeruli are highlighted in each stain, it is possible to detect them regardless of the staining.

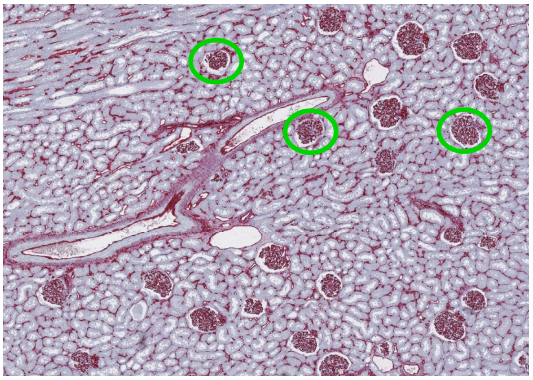
The term 'Digital pathology' is used for a variety of processes related to WSI manipulation, such as storage, annotations and analysis. Particularly interesting

[2] A glomerulus contains a ball-like network of specialised capillaries, representing the filter of the functional renal unit called a nephron [6]. More information is provided in Appendix A.



(a) Periodic Acid Schiff (PAS)

(b) Jones H&E



(c) CD34

Figure 1.2: An example of three consecutive WSIs of a kidney nephrectomy sample stained with different stains. Each staining provides different information on the tissue but some common structures, such as glomeruli, are visible in all stainings (some of them are marked in green circles).

is the exponential growth of publications related to digital pathology and artificial intelligence over the last ten years, where 40.57% of the total number of publications has been published in the last three years (2021, 2020 and 2019). According to such trends, it is expected that interest and usage will continue to increase as the technology progresses. For example, Generative Adversarial Networks (GANs) [5] introduced in 2014 already occupy almost 7.5% of all publications related to artificial intelligence and digital pathology. Thus, it is of crucial importance to explore the limitations and raise awareness about what is reasonable to expect from this technology and its progress.

In the following, common challenges related to digital histopathology and deep learning-based methods will be discussed. One of the greatest obstacles when developing deep learning solutions comes from the large variance introduced by the staining process. This will be discussed in more details in Section 1.1.1. Another challenge is data availability and quality, which is considered in Section 1.1.2.

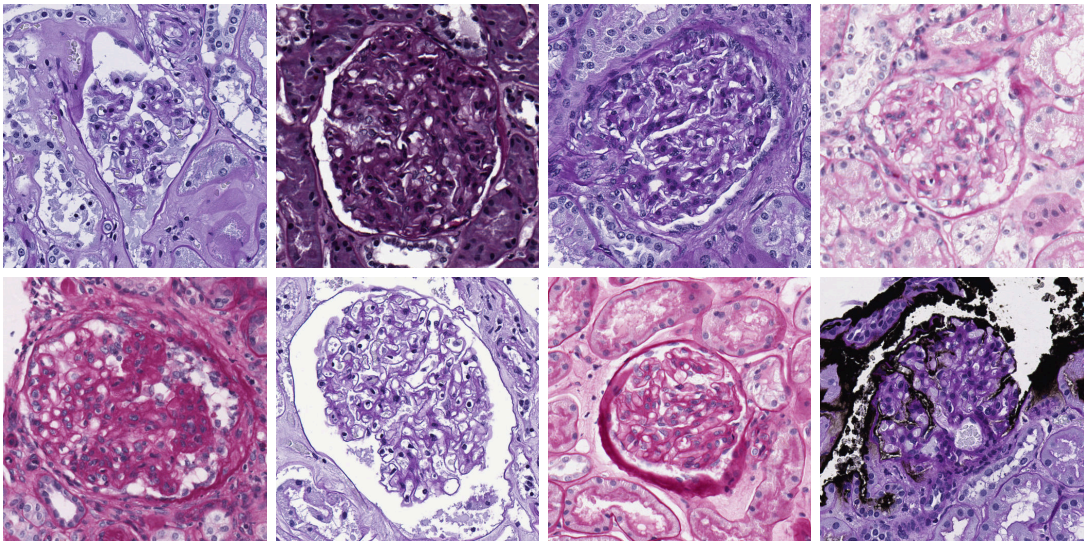


Figure 1.3: PAS-stain variation in kidney pathology. Each image represents the glomerulus in PAS staining.

1.1.1 Staining Process

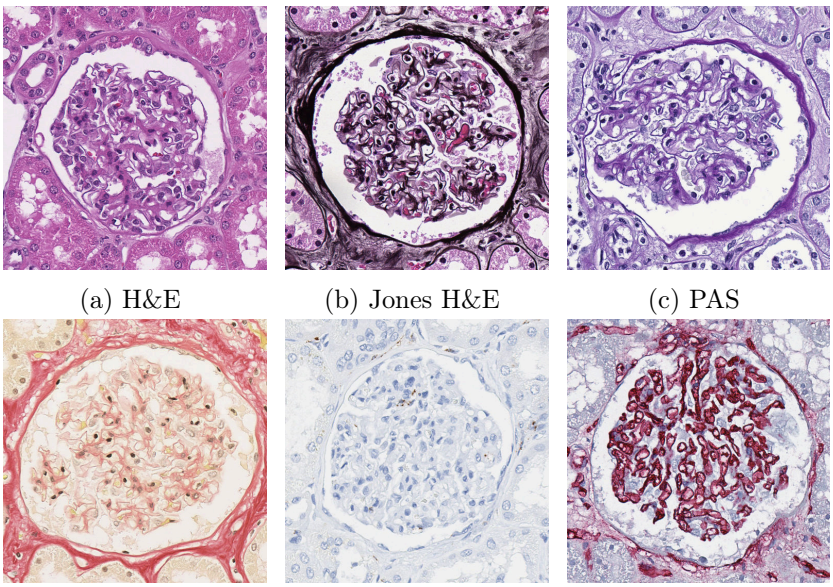
Staining is a crucial step in tissue preparation for histological examination as it visualises the chemical nature of the tissue and cell structures. However, the final WSI can take on a very different appearance, as illustrated in Figure 1.2. These differences can be attributed to stain variation and stain differences, which will be discussed in the following.

1.1.1.1 Stain Variation

Each of the above-mentioned preparation steps (fixation, processing, embedding, sectioning, staining and scanning) depends on multiple parameters, which can strongly affect the diagnostic quality and visual appearance of the resulting slide/image. The preparation of high-quality tissue slides requires careful manipulation and processing of the tissue since each step can introduce artefacts [36, 37]. For example, during the staining phase, impurities present in the dye or leaching of certain substances from tissues into the dye can affect the staining's intensity [38]. Apart from artefacts, most commonly, differences in raw materials, exposure time, quality of the substances used or scanner characteristics are the factors that introduce variation. Figure 1.3 illustrates PAS-stain variation in kidney pathology and its effect on glomeruli appearance. Such variation can be overcome in manual analysis by pathologist due to experience and special training but represent a great challenge for automated solutions.

1.1.1.2 Stain Differences

The difference between tissue sections stained with different stains is not just in their visual appearances (Figure 1.2). The chemical reaction provoked in the tissue



(a) H&E (b) Jones H&E (c) PAS

(d) Sirius Red (e) CD68 (f) CD34

Figure 1.4: Different stains used in kidney pathology. Each image represents a glomerulus (functional unit of a kidney). Each stain provides specific information about the structure and chemical environment of the glomerulus.

by the stain results in highlighting particular tissue components; thus, different structures, or at least different components of the same structures, are visible under each stain. An important aspect of digital histopathology, also considered in this thesis, is the analysis of multiple WSIs from the same tissue stained with different stainings. Usually, consecutive sections (which contain corresponding microscopic structures) are stained differently to enable the analysis of underlying tissue from different aspects. The analysis is usually performed with respect to a specific organ or structure. For example, to diagnose pathologies such as kidney allograft rejection, it is necessary to study the inflammatory micro-environment of the kidney (see Appendix A). In this context, the relevant information could be the distribution of immune cells such as macrophages in relation to glomeruli [39]. To automatically perform such an analysis, the structure of interest, in this case glomeruli, should be detected in each of the consecutive sections regardless of the staining used to stain that section. Figure 1.4 provides examples of glomeruli under different stainings. However, the differences between stains represent an important obstacle for a deep learning-based solution. Therefore, the development of stain invariant deep learning-based solutions, as the main subject of this thesis, is particularly relevant for the advancement of automatic analysis in digital histopathology.

1.1.2 Data Availability

The introduction of Whole Slide Imaging (WSI) scanners enables the production of vast amounts of histological image data. Dataset sizes reported in research papers have increased by several orders of magnitude in the last years [23]. In the light

of a recent discovery that dataset size significantly affects reaching the human-level performance of deep learning modes on visual tasks [40], one can expect considerable advancements toward automating routine histopathological analysis.

However, not all the data produced is of sufficient quality to be directly used. Thus, additional effort related to data preparation is usually required. Nevertheless, the most tedious task is related to the annotation process since a structure of interest could be very sparse or very dense. For example, the glomeruli are sparse structures, covering approximately 2% of the renal tissue area [41]. Therefore, the average kidney WSI, having a size of 100K $\times$80K pixels (at 40$\times$ magnification), contains around 500 glomeruli, where the glomerulus' average size is 500×500 pixels. Contrarily, cancerogenic cells can appear in big portions of WSIs, occupying dense areas of the image. Thus, obtaining annotated dataset is time-consuming, expensive and task-specific. Since it needs to be performed by experts [42], it is usually unfeasible to collect high-quality annotations for all the data acquired daily in hospitals. In addition to other important concerns related to data privacy, these enormous collected datasets are usually left aside. Instead, deep learning-based models are trained on specifically collected and prepared datasets, usually alongside some publicly available collections used for pre-training or initialisation.

Several histological datasets have been made publicly available. The Cancer Genome Atlas Program (TCGA) hosts a huge dataset collection related to different cancer types. However, datasets related to other histological analyses are not so numerous. For the case of renal pathology, the publicly available AIDPATH [18] dataset contains only 47 WSIs of human kidney tissue compared to, e.g. 1399 WSIs available in the CAMELYON dataset [43] (breast cancer metastases of lymph node sections). Dataset availability strongly influences which applications are studied more frequently [3]. Consequently, the application areas with a few available dataset collections may have a lower chance of being explored. One of the greatest obstacles in advancing automatic solutions (deep learning based) in general digital pathology is obtaining representative datasets with high-quality annotations and enough diversity to reflect the real world. Thus, methods that facilitate the annotation process or do not require huge datasets could be more easily adopted in practice.

One of this thesis' contributions is the introduction of methods able to learn from limited annotated datasets with good generalisation properties (Chapter 4 and Chapter 5). Moreover, this thesis proposes an automatic way to generate fully annotated dataset collections for renal pathology, which has the potential to create high-quality datasets for other data-hungry approaches (Chapter 5).

1.2 Generative Adversarial Networks (GANs)

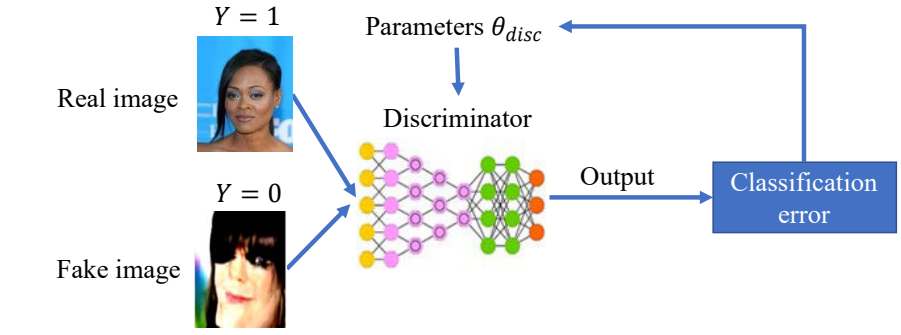
Deep learning has demonstrated remarkable abilities in discriminative tasks, e.g. recognising patterns in input data. However, in generative approaches, success was limited. Some of the attempts [44, 45] didn't attain convincing enough quality. That changed in 2014, with the introduction of Generative Adversarial Networks (GANs). Yann LeCun, the chief AI Scientist at Facebook and ACM Turing Award Laureate, described Generative Adversarial Networks (GANs) as "the most interesting idea in the last ten years in machine learning" (Quora Session, 28th July 2016). GANs have

Figure 1.5: StyleGAN — face generation. All images are generated; these people are not real. The image is taken from [50].

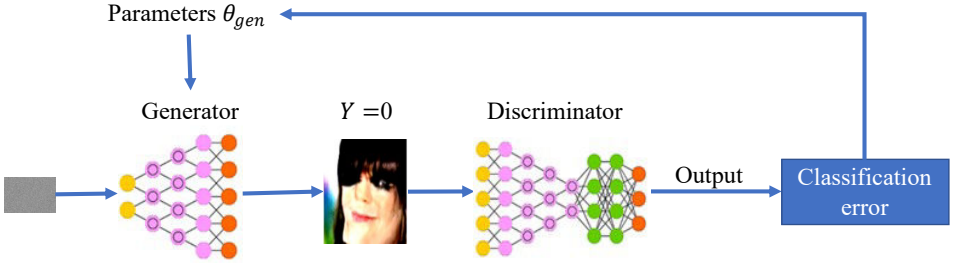
changed the perspective of deep learning and opened a path for impressive applications. Nowadays, the advances in GAN theory and practice enable the generation of high resolution, realistically looking images that are indistinguishable from real pictures (see Figure 1.5). The applications quickly spread to other types of data such as text [46], music [47] and videos [48]. The best indicator of the realism of generated samples is the legal consequences recently raised. For example, in 2020, the state of California put into effect a law regarding the usage of fake images in the public space [49].

In more technical terms, generative methods can be broadly classified into explicit and implicit approaches, and GANs belonging to the latter. Explicit generative models assume that there is a model likelihood function, and typically they are trained by maximising it. Well-known examples of this class of generative models are Variational auto-encoders [44] and PixelCNNs [45]. Contrarily, implicit models are able to generate data without knowing the explicit formulation of its distribution. Thus, GANs do not directly fit or estimate a data distribution. Instead, the model learns to sample from the data distribution by a two-player adversarial game. The players are called the Generator (G) and the Discriminator (D), which are usually represented as neural networks. The Generator learns to generate new data points, and the Discriminator learns to distinguish between real data samples and those produced by the Generator. The learning of these two models is a competitive (adversarial) game since the players' objectives are opposing. The Discriminator aims to discriminate between real and generated data as best as possible, while the Generator aims to create samples which are indistinguishable from the real data.

The Discriminator has a standard classification task to classify what is real and what is fake. The input is a data sample, and since it is known which samples come from the training set and which are generated by the Generator, the training of the Discriminator is fully-supervised. Thus, the objective is the maximisation of classification accuracy. The Generator creates a data sample from scratch, i.e.



(a) Discriminator



(b) Generator

Figure 1.6: Illustration of how Generative Adversarial Networks are trained — a) the Discriminator is trained in a fully-supervised manner to decrease classification error for both real and fake examples; b) the Generator is trained using the Discriminator's feedback for the fake images and aims to increase its classification error for the fake images.

usually, the input to the Generator is just a vector of random numbers. Since it is hard to directly quantitatively evaluate the quality of the created sample, the Generator uses feedback from the Discriminator in order to improve the generation. The Generator aims to fool the Discriminator into classifying fake samples as real, so its goal is to minimise the Discriminator's accuracy for fake samples. This is graphically presented in Figure 1.6.

More formally, the Generator G and Discriminator D are differentiable functions represented by neural networks and parametrised by θ_{gen} and θ_{disc} respectfully. The Generator maps a noise z with a prior distribution p_z into a data space $G(z, \theta_{gen})$, forming a distribution of fake images, p_g. The Discriminator outputs a single scalar, representing the probability that given sample x comes from real data p_{data} rather than p_g. The Discriminator is trained to maximise the probability of assigning the correct label to examples from both distributions, while the Generator is trained to minimise the Discriminator's probability of assigning a correct label to a fake sample. Therefore, they play a two-player minimax game with a value function $V(G, D)$, formulated as follows:

$$\begin{aligned}\min_G \max_D V(G,D) &= \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D(x))] \\ &= \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \end{aligned} \tag{1.1}$$

Assuming that both models G and D have enough capacity, the Nash equilibrium of the game is achieved when $p_{data} = p_g$ and when D is not sure about the origin of a sample, so produces $\frac{1}{2}$ for both real and fake samples [5]. If binary-cross entropy is used as the Discriminator's loss function, the adversarial game approximates Jensen–Shannon divergence between p_g and p_{data} [5].

The theoretical background holds under the assumption that there is a sufficiently large sample size (dataset), the Generator and the Discriminator have enough capacity, and the training is performed adequately long [51]. In practice, these assumptions are often violated, which leads to well-known challenges involved in GAN training, such as:

- Mode collapse: the Generator maps diverse inputs to the same output, i.e. p_g captures only a few modes of p_{data}.

- Vanishing gradients: when the Discriminator is very confident (loss close to zero), the gradients provided to the Generator are small, resulting in very slow or no training of the Generator. Contrarily, if the Discriminator is not good enough, its feedback is also not valuable for the Generator's learning.

- Convergence: obtaining a global Nash equilibrium is not straightforward. Thus, learning usually oscillates or converges to a local Nash equilibrium which can be far from the global one.

To address the above problems, GAN modifications have been proposed related to architecture and/or objective function. Some of the extensive reviews of GANs architectures or objectives are [52–54]. The multitude of GAN architectures opens additional questions regarding their comparisons and evaluation. Thus, an important direction of research related to evaluation metrics has been rapidly developed [55].

Being able to generate samples from complex data distributions, GANs have great potential to overcome some of the limitations in applying deep learning-based solutions in digital histopathology. For example, to address the lack of data, GANs can be employed to synthesise artificial datasets [56]. In this way, many concerns related to data privacy can largely be reduced. Adversarial domain adaptation, on the other side, can be used to reduce domain shift between public and private datasets [57], which is a widely observed problem in digital histopathology. From a more clinical side, GANs can be used to enhance the diagnostic process [58]. Moreover, a fully virtual staining process [59] can reduce laborious tissue preparation time and decrease the number of artefacts introduced during staining. However, GAN applications are rather limited to research studies and rarely transferred to clinical practice despite such great opportunities. One crucial reason is that GANs are a very recent technique whose theoretical and practical basis is not yet well understood. This thesis contributes to a better understanding of the possibilities and limitations of GAN-based methods in digital histopathology.

1.3 Thesis Goals and Contributions

1.3.1 Objectives and Contributions

The main focus of this thesis is to investigate the opportunities for Generative Adversarial Networks (GANs) to be applied in the field of digital histopathology. Two main research directions can be identified in this thesis:

- Generative Adversarial Networks for stain transfer: this branch of research applies GANs for stain transfer, i.e. changing an image's appearance that was initially stained with stain A to look like as it has been stained with stain B. The obtained transfer needs to be plausible — the histological image, in the absence of patient-specific information such as the underlying disease, looks visually correct to a trained expert with regard to the staining characteristics and the morphological appearance of the tissue components. This thesis proposes several ways to obtain visually convincing translations and gives limitations of such approaches from both a diagnostic point-of-view and their application in the computer vision domain.

- Generative Adversarial Networks for stain invariance: this branch of the thesis investigates how GANs can be used to build better, more robust, deep learning models. For the problems where a stain invariant solution is feasible, i.e. the problems solvable across multiple stains, the thesis proposes approaches that result in stain invariant models. The learning is performed using a limited number of annotations from a single stain modality, and the goal of the obtained solution is to generalise across multiple stains, even those unseen during training.

This thesis, in parallel with other authors [19, 20], for the first time proposes the usage of a GAN-based image-to-image translation method for stain transfer between different staining modalities — the thesis' first contribution[3]. In the meantime, the approach based on CycleGAN has been established as a standard solution for virtual staining in general and is widely adopted in the field of digital histopathology. The most important works are classified and summarised in Chapter 2 of this thesis. From this summary it is evident that the main focus of the literature so far has been dedicated to stain normalisation, i.e. standardising histological image appearance inside one stain modality, where CycleGAN-based solutions are the dominant approaches. However, stain transfer between different staining modalities, on which this thesis focuses, is rarely addressed in the literature. The thesis reveals that stain transfer between different stains opens more intriguing questions and has specific limitations compared to stain normalisation.

This thesis identifies that CycleGAN-based translations contain imperceptible information related to stain differences whose manipulation can modify the resulting translation in a plausible way. As a consequence of this finding, the thesis proposes an unsupervised augmentation method that increases the robustness of deep learning-based solutions — the thesis' second contribution.

[3]The work [19] has been published at the same time as the preparation of the publication related to CycleGAN based results presented in this thesis. After the publication of Gadermayr et al. [19], the work was extended with additional analysis.

Moreover, this thesis discovers and demonstrates the sensitivity of CycleGAN-based solutions to small architectural modifications. These changes do not necessarily affect the visual quality of the obtained translations but influence the overall conclusion related to the usefulness of stain transfer from both diagnostic and application points of view — the thesis' third contribution.

Furthermore, the thesis takes advantage of the plausibility of the obtained translations to propose the first solution that encourages empirical stain invariance for the task of glomeruli segmentation. The obtained model is able to segment multiple stains and is also able to generalise to some unseen stains — the thesis' fourth contribution. The findings and conclusions of the research conducted in this thesis are furthermore used to propose an end-to-end trainable model that simultaneously performs stain transfer and stain invariant segmentation — thesis' fifth contribution. The proposed model is, for the first time, also able to simultaneously perform stain transfer between different stains, stain normalisation inside one stain modality and generalise the translation process to unseen stains. In addition to this, all the generated data (including the original and unseen stains) are correctly segmented via the model's stain-invariant segmentation module. These findings allow for the generation of the first artificially created, fully annotated dataset KIDNEY-ARTPATHOLOGY, which will soon be made available to the community for further advancement in digital histopathology — the thesis' sixth, and final contribution.

1.3.2 Data

This thesis is focused on renal pathology and the specific task of glomeruli segmentation in multiple stainings. Medical background information are provided in Appendix A. The dataset is composed of a private part, used for research conducted in this thesis, and a public part, adapted for testing purposes where applicable.

The private dataset contains tissue samples collected from a cohort of 10 patients who underwent tumour nephrectomy due to renal carcinoma. The kidney tissue was selected as distant as possible from the tumours to display largely normal renal glomeruli; some samples included variable degrees of pathological changes such as full or partial replacement of the functional tissue by fibrotic changes ("sclerosis") reflecting normal age-related changes or the renal consequences of general cardiovascular comorbidity (e.g. cardiac arrhythmia, hypertension, arteriosclerosis). The paraffin-embedded samples were cut into $3\mu m$ thin sections and stained with either Jones' basement membrane stain (Jones H&E), Periodic acid-Schiff reaction (PAS) or Sirius Red, in addition to two immunohistochemistry markers (CD34 highlighting blood vessel endothelium, CD68 for macrophages), using an automated staining instrument (Ventana Benchmark Ultra). Whole slide images were acquired using an Aperio AT2 scanner at $40\times$ magnification (each pixel corresponds to a square of $0.25\mu m$ in the slide). All the glomeruli in each WSI were annotated and validated by pathology experts by outlining them using Cytomine [60]. The dataset was divided into 4 training, 2 validation, and 4 test patients, and in this thesis, it is referred to as Hanover or private dataset. The number of glomeruli in each staining dataset is given in Table 1.1.

In addition to the above, for the specific analyses performed in Chapter 3 and Chapter 5 of this thesis, the publicly available dataset AIDPATH [18] is used. AID-

Table 1.1 The number of glomeruli in each staining used in the private dataset.

Staining	Training	Validation	Test
PAS	662	588	1092
Jones H&E	624	593	1043
Sirius Red	654	579	1049
CD34	568	598	1019
CD68	529	524	1046

PATH is a collection of five different datasets of human kidney tissue cohorts acquired and digitised from three European institutions: Castilla-La Mancha's Healthcare services (Spain), The Andalusian Health Service (Spain) and The Vilnius University Hospital Santaros Klinikos (Lithuania). Tissue samples were collected with a biopsy needle having an outer diameter between $100\mu m$ and $300\mu m$, and paraffin blocks were prepared using tissue sections $4\mu m$ thin, then stained using PAS [18, 61]. In total, the dataset contains 47 WSIs. The slides used in this thesis were manually annotated using the same system as for the private dataset.

Glomeruli segmentation is framed as a two-class problem: glomeruli (pixels that belong to a glomerulus) and tissue (pixels outside a glomerulus). The training set comprised all glomeruli from a given staining's training patients plus seven times more tissue (i.e. non-glomeruli) patches (to account for the variance observed in non-glomeruli tissue). In all experiments, patches of size 512×512 pixels (at $40\times$ magnification for private dataset and $20\times$ magnification for public dataset) are used since glomeruli and part of the surrounding tissue fit within this size of a patch at the level of detail used.

1.3.3 Thesis Outline

The remainder of this thesis is organised as follows (Figure 1.7 represents a condensed overview of the main thesis structure and contributions):

- Chapter 2 gives an extensive and systematic analysis of the literature regarding virtual staining and stain invariant solutions based on Generative Adversarial Networks. Several gaps are identified that are the focus of the contributions of this thesis.

- Chapter 3 proposes, in parallel with other authors [19, 20], the use of the CycleGAN to achieve plausible stain transfer between different stains. It will be demonstrated that such methods can significantly reduce domain shift caused by different stains. Moreover, this chapter gives several contributions regarding the critical analysis of the proposed method and identifies important limitations that are raised for the first time. Some of these findings are used to propose a new unsupervised augmentation strategy that, in a fully-supervised training setting, has a beneficial effect on model robustness. The results of this chapter are published in [62, 63] and partially in [64].

- Chapter 4 builds upon the findings of Chapter 3 to propose the first stain invariant solution, which is able to generalise across multiple stains, even to unseen stains. The results are primarily published in [64]. Moreover, this

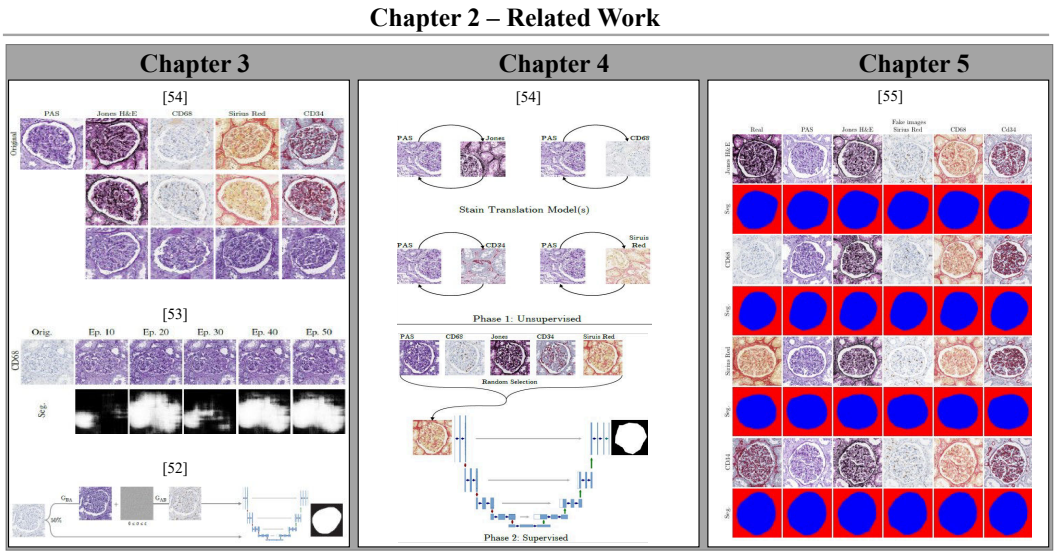


Figure 1.7: A visual overview of the structure and the main contributions presented in this thesis.

chapter demonstrates the benefits of stain transfer to feature space domain adaptation.

- Chapter 5 combines stain transfer and stain invariant segmentation into a single, end-to-end trainable model. Such a model demonstrates better generalisation to unseen stains than the previously proposed solution (Chapter 4). Moreover, for the first time, it is possible to translate unseen stains and perform stain normalisation and stain translation in a single forward pass of the model. These properties are exploited to generate the first artificially created, fully annotated kidney pathology dataset. The results of this chapter are under review [65].

- Chapter 6 gives the conclusions of the research presented in this thesis, and future research directions are identified.

Chapter 2

Related Work

With the introduction of whole-slide scanners, the amount of digitalised histological data has been dramatically increasing, which opens up the potential for automated analysis. However, a deep learning model's sensitivity to domain shift and the general scarcity of annotations in digital histopathology pose an important challenge to their effective application in the field. As previously mentioned in Section 1.1.1, a standard histological analysis needs to deal with the variation of a sample's appearance that occurs due to differences in tissue preparation and staining protocol. Therefore, two main sources of variation can be identified — intra-stain variation, which is the variation in the appearance of the same staining (e.g. due to different laboratory procedures), and inter-stain variation, which is the variation in the appearance of different stainings, as illustrated in Figure 2.1.

Most deep learning-based algorithms are sensitive to domain shift [9, 66], which is, in the context of digital histopathology, introduced by both inter-stain and intra-stain variation. For example, it is likely that models trained for a specific task on histological images of stain A exhibit a drop in performance when applied (for the same task) to histological images of stain B [17, 67, 68] or a variation of stain A (e.g. images from other laboratories) [69]. Typical solutions consider either fine-tuning existing models or training a new model for each possible variation, which requires additional labelled data. Nevertheless, medical image datasets are often characterised by their scarcity of annotations [70] and obtaining properly annotated images is time-consuming and costly as expert knowledge is required [42]. Thus, addressing stain variation becomes of great importance for the successful development and application of automated systems. The introduction of Generative Adversarial Networks (GANs) [5] triggered an exponential growth of approaches that employ them to tackle domain shift introduced by stain variation.

This chapter represents a review of papers that address intra-stain or inter-stain variation problems in digital histopathology by Generative Adversarial Networks(GANs). To start, Section 2.1 briefly describes GAN models and architectures. The main classes of approaches to GAN-based solutions in digital histopathology are then identified in Section 2.2, and the subsequent sections (2.3, 2.4, 2.5) discuss in more detail each of these approaches. A discussion regarding existing methods is presented in Section 2.6, which leads to the identification of the main points addressed in this thesis (Section 2.7).

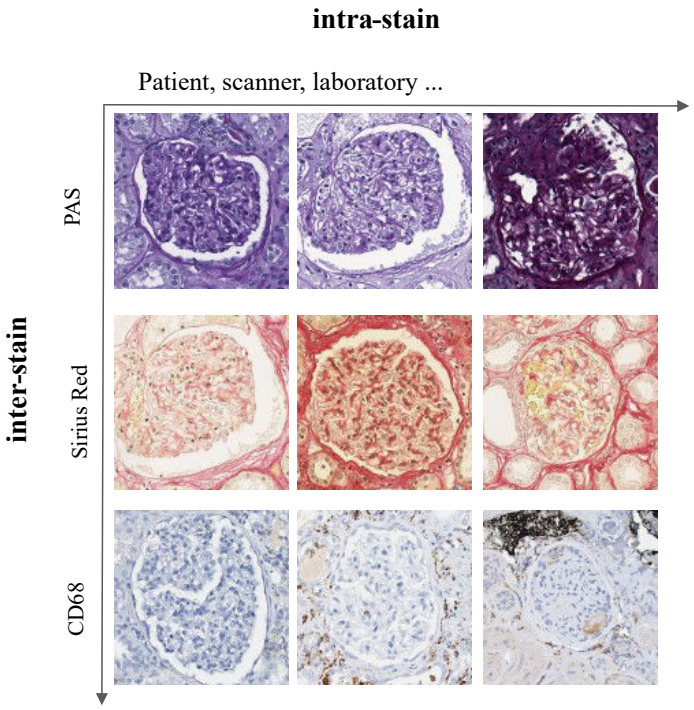


Figure 2.1: Stain variability: Intra-stain and inter-stain variation in the case of a kidney dataset. Each row contains samples stained with the same stain.

2.1 Overview of Generative Adversarial Networks

Generative Adversarial Networks (GANs) [5] have gained significant attention since their introduction in 2014. In practice, the idea of adversarial training of two neural networks facilitates sampling from very complex data distributions, which greatly increases the number of possible application areas. To be reminded, for a given dataset, the Generator is optimised to generate new samples coming from the same data distribution while the Discriminator learns to distinguish between real data samples and those generated by the Generator. The learning of these two models is a competitive (adversarial) game since the players' objectives are opposite to each other — the Discriminator aims to discriminate between the real and generated data as best as possible; in contrast, the Generator aims to create samples that are as close as possible to the real data. The optimal outcome of such a game is a Nash equilibrium, where the Generator produces samples indistinguishable (from the Discriminator's perspective) from the real data.

Specific GAN architectures have been developed with the expansion of GAN application areas. For example, the successful implementation by Radford et al. [73] of convolutional neural networks (CNN) for both the generator's and discriminator's architecture opens space for advancements in image synthesis applications. Involving GAN models in various tasks makes practical problems related to obtaining a Nash equilibrium (mode collapse, vanishing/exploding gradients and convergence) more evident. The approaches to address these issues usually involve modifying the objective function and/or model architecture. In Figure 2.2 some of the widely

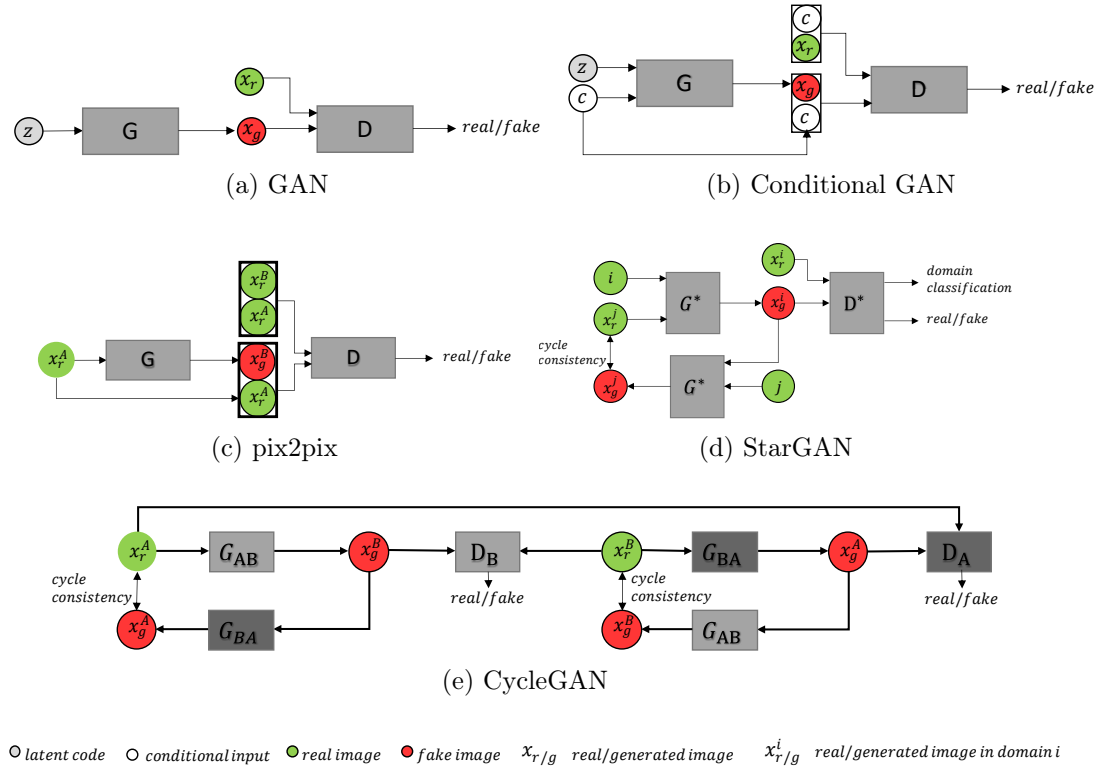


Figure 2.2: GAN-architectures widely adopted in the field of digital histopathology.

applied GAN models are graphically represented, while Table 2.1 summarises their objective functions and main characteristics.

The originally proposed GAN model provides unconditional generation, see Figure 2.2a, where there is no influence on the output, i.e. after training, the generated data are randomly sampled from the distribution that the generator learnt to approximate. Initially, the model used binary-cross entropy as the discriminator's loss, approximating the Jensen-Shannon divergence between the real and generated data distributions [5]. However, commonly reported issues associated with this loss are mode collapse, instability and uninformative loss values. Thus, there are multiple GAN variants that try to overcome these limitations. For example, Arjovsky et al. [74] propose a new cost function that measures the Wasserstein distance between real and generated data distributions. This distance metric has better theoretical properties than Jensen–Shannon divergence and widely reduces the probability of mode-collapse. Although the landscape of different GAN cost functions used in the literature grows daily, there is little evidence that one function is always better than another [75]. Consequently, there is no golden rule regarding the choice of objective function since each has its downsides. A recent review has studied the different cost functions in more detail [52].

Contrarily to the original GAN formulation, which does not allow control over the generation process, Mirza and Osindero [71] propose the Conditional GAN model, see Figure 2.2b. By providing additional input (e.g. class information) to both the generator and discriminator, control over the generation process is enabled,

Table 2.1 Different GAN architectures with corresponding objective functions.

Architecture	Objective function	Remark
GAN [5]	$\mathcal{L}_{\text{GAN}}(G,D) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$	Unconditional generation.
CGAN [71]	$\mathcal{L}_{\text{cGAN}}(G,D) =$ $\mathbb{E}_{x\sim p_{data}}[\log D(x\|y)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z\|y)))]$	Conditional generation.
pix2pix [72]	$\mathcal{L}_{\text{pix2pix}}(G,D) = \mathcal{L}_{cGAN}(G,D) + w\mathcal{L}_{L1}(G)$	Supervised image-to-image translation
CycleGAN [12]	$\mathcal{L}_{\text{cyc}}(G_{AB}, D_B, G_{BA}, D_A) =$ $\mathcal{L}_{\text{adv}}(G_{AB}, D_B) + \mathcal{L}_{\text{adv}}(G_{BA}, D_A) + w\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA})$	Unsupervised image-to-image translation
StarGAN [14]	$\mathcal{L}_{\text{Star}}(G,D) = \mathcal{L}_{\text{adv}}(G,D) + w_{\text{cyc}}\mathcal{L}_{cyc}(G) + w_{\text{cls}}\mathcal{L}_{cls}(G,D)$	Unsupervised multi-domain image-to-image translation.

e.g. by providing a class label, a random sample from a given class is generated. The additional input can represent any information such as a data label, an image attribute, or an image itself. Conditioning opens possibilities for different applications such as image-to-image translation [72], super-resolution and [76], image inpainting [77]. Some particularly interesting application areas are image-to-image translations.

Image-to-image translation can be defined as converting an image x_A from domain A into an image $\hat{x}_B$ in the domain B, taking the style of the domain B and preserving the content of image x_A. Conditional adversarial networks are widely adopted for this purpose, where pix2pix [72], CycleGAN [12], and StarGAN [14], are pioneers and have been widely applied and extended.

The pix2pix architecture [72], given in Figure 2.2c, is an extension of Conditional GANs with several modifications. The general idea is to enable translation between images from a domain A to domain B in a fully-supervised setting where pairs of translations are already known (i.e. scratched and original photo pairs). The model contains one Generator G, which takes an image from domain A, x_r^A and translates it to domain B, x_g^B. The Discriminator distinguishes in, an adversarial manner, groundtruth image pairs (x_r^A, x_r^B) and pairs which contain generated images, (x_r^A, x_g^B). Additionally, since the dataset is paired, the Generator is constrained with the L_1 loss to obtain translations as close as possible to the groundtruth. Pix2pix has inspired many works in supervised image-to-image translation [78, 79].

Contrarily to the pix2pix model which requires a paired dataset, the CycleGAN architecture [12] enables translation in an unpaired setting (see Figure 2.2e). The model contains two generators — G_{AB}, which translates an image from domain A to domain B; and G_{BA}, which translates an image from domain B to domain A. Generators play adversarial games with two discriminators — D_A, which distinguishes between real images from domain A and images which are translated from domain B; and D_B, which distinguishes between real images from domain B and images translated from domain A. To overcome the lack of supervised pairing and to

Table 2.2 Classification of virtual staining techniques.

Virtual staining	Original image	Target (artificial) image
Stain unstained tissue	Unstained	Stained with a target stain
Stain normalisation	Stained with a variant of stain A	Stained with a variant of stain A
Stain transfer	Stained with stain A	Stained with stain B

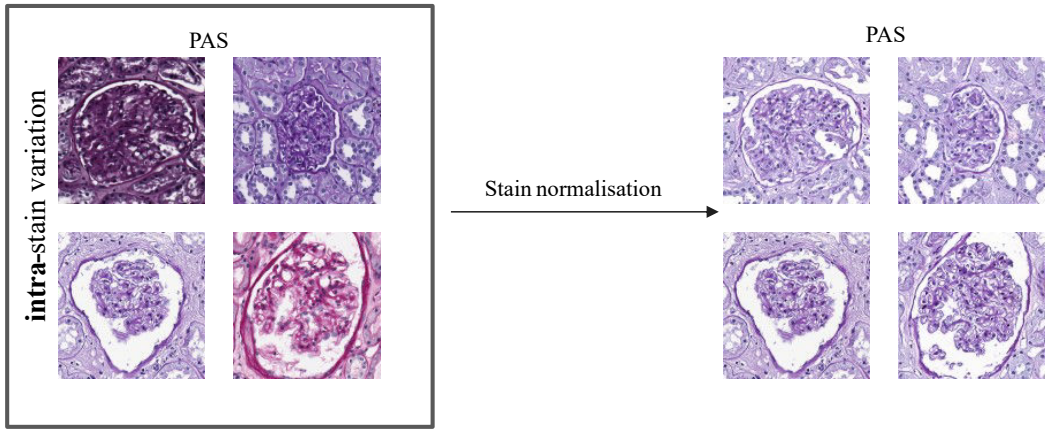
prevent structural changes during translation, the learning is constrained by a cycle-consistency loss. This is a requirement that the pixel-wise distance between a real image and its cyclic reconstruction is minimal (in Figure 2.2e the distance between x_r^A and x_g^A, and x_r^B and x_g^B, is minimised). Due to its ability to perform translations between unpaired domains, CycleGAN has been broadly adopted in medical imaging, where obtaining paired datasets is laborious (e.g. destaining an already stained tissue) and time-consuming. Thus, the model is able to learn translations without paired data, and this model has been extensively exploited.

CycleGAN provides bidirectional translations from domain A to domain B and vice versa. However, in the case of translations between multiple domains, a separate CycleGAN model needs to be trained for each pair of domains. The StarGAN architecture [14], given in Figure 2.2d, represents a model able to perform multi-domain translation. A specific discriminator design, which is extended with an auxiliary domain classification task, provides a multi-domain image-to-image translation framework. Similarly to CycleGAN, the model is able to learn translations between unpaired domains. The model contains one generator which, conditioned on a target domain label i, transforms an image from domain j, x_r^j, to look like coming from domain i. As with CycleGAN models, the cycle-consistency loss is employed to prevent structural changes during translation. Several works such as [80, 81] build upon the StarGAN idea to enable more fine-grained control over the translation process.

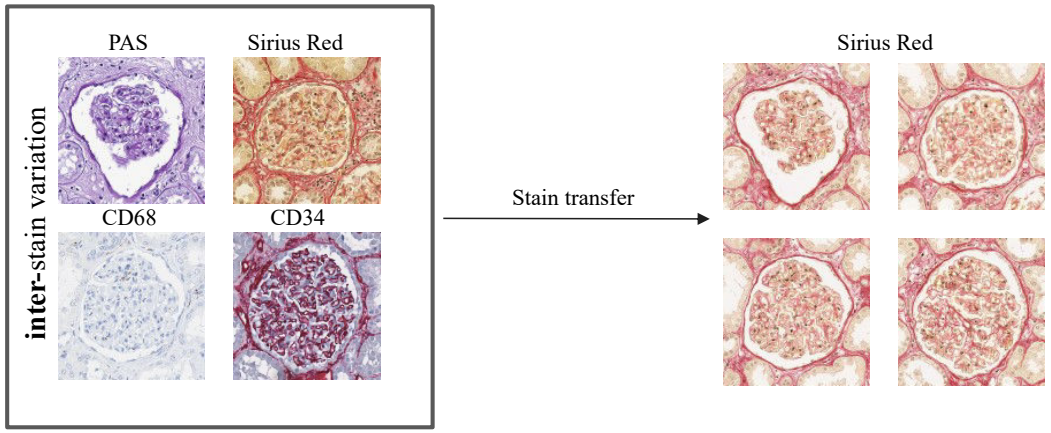
2.2 GAN-Based Solutions To Stain Variation

Generative Adversarial Networks (GANs) can be used for multiple purposes in Digital Histopathology from both clinical and computer vision perspectives.

From a clinical point of view, GANs can be used to obtain virtual staining between different stains, which is a progressively growing and significant application area. Since the difference between histological images stained differently is not just colour-based but also in the highlighted tissue components, multiple stainings of the same specimen provide clinicians with different information regarding a patient's health status. However, obtaining multiple stainings of the same specimen is not always possible. For example, Anglade et al. [82] indicate that cancer patients could have limited access to immunohistochemistry stainings, which are important for accurate diagnosis, depending on the resources available in a specific laboratory. Thus, a significant effort is invested in developing automated solutions that can exploit already stained tissue to conclude information typically obtained from multiple stainings. Recent approaches widely explore GAN's potential for the task of virtual staining. The GAN's architecture for image-to-image translation provides an effective way to obtain virtual staining — the artificial modification of a histopathological



(a) Pixel-space — Intra-stain (stain normalisation).



(b) Pixel-space — Inter-stain (stain transfer).

Figure 2.3: Illustration of the differences between intra-stain and inter-stain standardisation/augmentation approaches. Standardisation aims to normalise image appearance within one stain (intra-stain, here PAS) or between different stains (inter-stain). Augmentation approaches aim to increase data variability by simulating a wide range of possible image appearances.

image's appearance after its acquisition.

The final result of virtual staining approaches is visible as a histological image. Therefore, the pathologist can inspect it, which is an important consideration in medical imaging. Three main classes of such methods can be identified, as summarised in Table 2.2. The first class of approaches try to skip the physical staining process by enabling virtual staining of unstained tissue. The second class attempts to change the appearance of an image originally stained by stain A to look like as it had been stained by stain B. Depending on the relation between stain A and stain B, two types of such virtual staining can be identified. If stain A and B are the same stainings and the difference comes from intra-stain variation, the virtual staining process is referred to as stain normalisation. The process is referred to as stain transfer if the stains A and B are two different stainings. The difference between stain transfer and stain normalisation is also represented in Figure 2.3a and Figure 2.3b. Thus, stain normalisation methods can be considered a special case of stain

transfer. Sometimes, virtual staining and stain transfer are terms which are used interchangeably. However, for the remainder of this thesis the term stain transfer will refer to a virtual (re)staining where original and target stains are different.

Regardless of the diagnostically different information provided in various stainings, pathologists can detect the same tissue structures across multiple stainings, even though composing parts are highlighted differently in each. For example, as previously illustrated in Figure 2.3, glomeruli structures are observed in multiple stainings of kidney pathology. Being able to automatically detect such common structures across different stainings can be beneficial as it provides additional information during diagnosis, e.g. the distribution of immune cells around glomeruli [39]. However, from the computer-science point of view, stain variation can be regarded as a source of domain shift, representing an obstacle to the development of automated solutions. Generative Adversarial Networks (GANs) can be employed to reduce the domain shift in both pixel-space and feature space. Pixel-space adaptation is usually performed by virtual staining methods. Feature space adaptation is commonly inspired by adversarial learning and usually requires specifically designed GAN architectures. Each of these approaches will be discussed in the following.

Virtual staining for pixel-space adaptation is mainly used in two manners: standardisation, i.e. the reduction of a model's input variation and augmentation, i.e. the expansion of the model's input variation. Standardisation approaches aim to unify a model's input appearance. This can be achieved by modifying the properties of a target image in order to match the characteristics of images used during the model's training (source images), e.g. models trained on the source domain can be applied to the modified target images. This direction is illustrated in Figure 2.3. Conversely, the annotated domain can be translated to match the unannotated (in a way that annotations are preserved), which enables the training of a model on the transformed source images, and the resulting model is directly applicable in the target domain. Contrarily, augmentation-based approaches employ virtual staining to obtain more robust models by expanding the variation of the model's input space. That is the opposite direction of the illustrated in Figure 2.3.

When GANs are employed to reduce domain shift in feature space, the adaptation is performed by forcing feature-space alignment of the source and target data. The assumption is that if task-specific learning is based on domain invariant features, the model should be able to generalise across multiple domains (stains). Traditionally, feature-space-based methods explicitly impose feature distribution alignment via statistical measures such as Maximum mean discrepancy (MMD) [83, 84]. More detailed information about such methods can be found in [9, 85]. With the introduction of Generative Adversarial Networks, the state-of-the-art domain adaptation methods rely more on implicit feature distribution matching imposed by adversarial learning (see Figure 2.4). The task-related model is divided into a feature extractor and a task-specific module (e.g. classification, segmentation). On top of the feature extractor, the discriminator is attached with the aim of distinguishing the distributions of features from the annotated and unannotated stains (the top part of Figure 2.4). A feature representation is considered to be domain-invariant if features extracted from both domains follow the same distribution [86], i.e. the discriminator cannot predict the domain from which the samples originate. Suppose a task-related model is trained on a stain invariant feature representation extracted

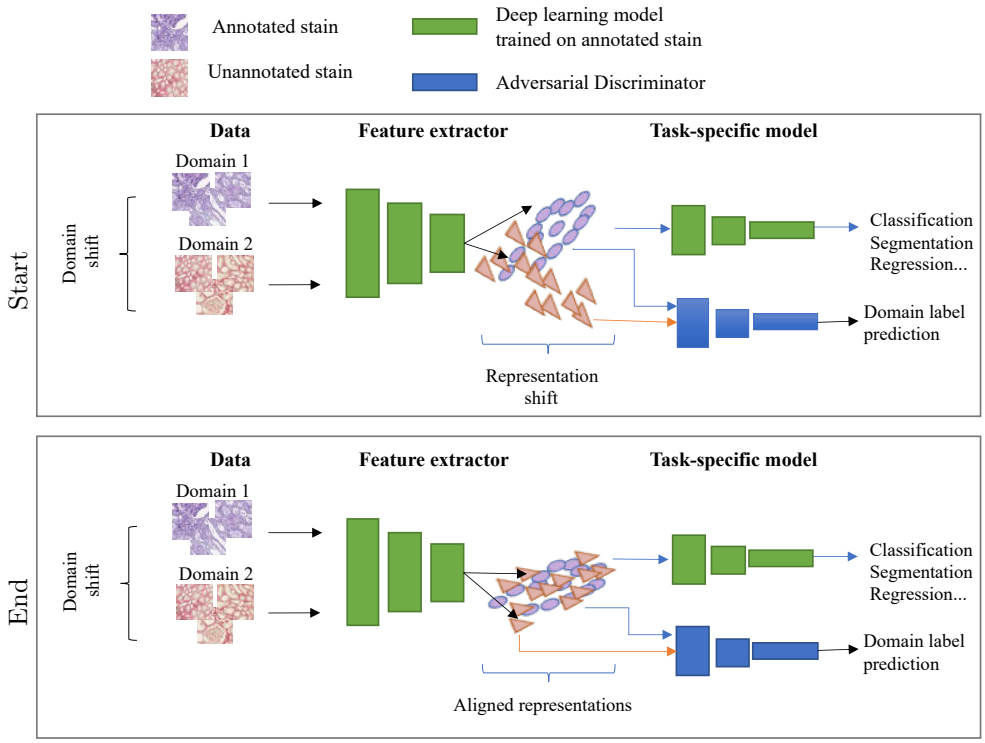


Figure 2.4: Feature-space domain adaptation approaches reduce domain shift in feature space, forcing a model to extract stain-invariant features.

from such feature extractor. In that case, it is expected that the model generalises to a target domain since features from both domains are indistinguishable (from the discriminator's perspective). During the training, usually, both source (annotated) and target (unannotated) data are passed through the feature extractor and discriminator, while only source data is passed to the task-specific (segmentation, classification, etc.) branch. The discriminator and feature extractor play an adversarial game that eventually leads to feature-space alignment between the source and target data (bottom part of Figure 2.4).

2.3 GANs for Staining Unstained Tissue

Obtaining histochemical staining of a specimen usually involves an irreversible multistep procedure which is destructive to the specimen. Thus, enabling diagnostically relevant virtual staining where the laborious and prone to variation staining process is bypassed can have great benefits. Table 2.3 summarises the representative GAN-based solutions for this task.

The proof-of-concept approach, which confirms that GANs can perform such a task, is done by Bayramoglu et al. [59] where unstained hyperspectral lung tissue images were converted into H&E staining using conditional GANs. Later, the staining process is bypassed by employing GANs to translate from an autofluorescence image of unstained tissue to a virtually stained histological image in multiple stain-

Table 2.3 GAN-based approaches to virtual staining of unstained tissue.

Source image	Representative work(s)	GAN model	Target Stain
Hyperspectral	Bayramoglu et al. 2017[59]	CGAN	H&E
Autofluorescence	Rivenson et al. 2019[32]	GAN model	Multiple
	Zhang et al. 2020[87]	CGAN	Multiple
	Liu et al. 2022[88]	DIRT [89]	Multiple
	Dimitrakopoulos et al. 2020[90]	Markov Random Field loss + GAN	Multiple
Unstained	Rana et al. 2020[91]	pix2pix	H&E
	Li et al. 2020[92]	pix2pix, StarGAN	H&E

ings [32, 87]. Following this line of research, other approaches also use GANs to translate microscopic images of unstained tissue to multiple stainings [91–93].

Existing approaches mainly rely on paired datasets to learn translations as accurate as possible. Obtaining a high-quality dataset usually involves a non-trivial slide registration task. Experts (experienced pathologists) are commonly involved in validating the obtained results [91]. The conclusions are relatively optimistic about the potential of GAN-based methods in these application areas. By bypassing the histological process, the problems arising from both inter and intra-stain variation can be reduced. Thus, multiple fields of digital histopathology can benefit from advances in these methods. However, current studies focus on the widely used H&E staining and/or particular tissue/disease. In order to make a universal virtual staining approach, there is a considerable consensus in the field that large-scale clinical studies need to be conducted to ensure the suitability and clinical applicability of proposed solutions [32, 87]. Moreover, using deep models in this context additionally raises the essential question of interpretability and explainability, which are so far rarely considered.

2.4 GANs for Stain Normalisation

Histopathological images of the same stain and tissue can take on a very different appearance due to variations in any of the multiple steps involved in the process of tissue preparation and staining. In Figure 2.1 such an effect can be observed row-wise, while the inter-PAS variation is presented in more detail in Figure 2.3b. It is known that such intra-stain variations can increase inter-observer disagreement [124, 125] and that it is also harmful to automatic analysis [8].

Historically, stain normalisation methods aim to standardise an image's appearance to match a selected reference image. The main principle is stain decomposition, where the image is represented by its stain concentration and stain colour matrices, which are further adjusted according to the reference image [126–130]. More recent approaches use machine learning or deep learning strategies to standardise image appearance [131, 132]. However, the outcome of these methods is known to be highly sensitive to the choice of reference image [106].

Table 2.4 GAN-based approaches to reduce the effects of intra-stain variation.

Approach	Representative work(s)	Considered stains
standardisation		
CycleGAN	StainGAN [94]	H&E
CycleGAN - modif.	Trans-Net [95]	H&E
	de Bel et al. [96]	PAS
	Residual CycleGAN de Bel et al. [97]	H&E,PAS
CycleGAN - extension	StainNet [98] (distillation learning)	H&E
	SAASN [99] (attention)	H&E
	Mahapatra et al. [100] (self-supervsion)	H&E
	Ke et al. [101] (contrastive learning)	H&E
InfoGAN	Zanjani et al. [102]	H&E
	Zanjani et al. [103]	H&E
pix2pix	STST [104]	H&E
CustomGAN	(style transfer) Nishar et al. [105]	H&E
	(auxiliary task) BenTaieb and Hamarneh [106]	H&E
	(auxiliary task) Liang et al. [107]	H&E
	(feature disentanglement) Moghadam et al. [108]	H&E
	SA-GAN [109]	H&E
StarGAN - extension	MultiPathGAN [110]	H&E
augmentation		
DIRT++	HistAuGAN [111]	H&E
Conditional GAN	HistoGAN [112]	H&E
	Li et al. [113]	H&E
	Histology CGAN [114]	H&E
Custom GAN	SDAE-GAN [115]	IHC
CycleGAN	Tsirikoglou et al. [116]	H&E
Progressive GAN	Levine et al. [117]	H&E
feature-space adaptation		
DANN	Lafarge et al. [118]	H&E
	Hashimoto et al. [119]	H&E
	Graziani et al. [120]	H&E
	Marini et al. [121]	H&E
	DA RetinaNet [122]	H&E
ADDA	Ren et al. [123]	H&E

Nowadays, GAN-based solutions are widely applied for stain normalisation. The majority of approaches consider stain normalisation as a problem of image-to-image translation, where standardisation and an augmentation-based group of approaches can be identified. Moreover, some authors consider stain normalisation as a domain adaptation problem, which forms the third group of adversarial domain adaptation-based approaches. The representative works from each group are summarised in Table 2.4 and will be discussed in the following.

Standardisation approaches: Image-to-image translation methods are usu-

ally employed in the context where one domain represents samples from a reference staining, and the other domain is translated to match the reference domain's characteristics. Some authors simulate paired datasets by discarding colour information (e.g. by using a greyscale image version or extracting a haematoxylin channel from the image) in order to employ image-to-image translations models such as pix2pix [72]. For example, Salehi and Chalechale [104] directly employ the pix2pix model while other authors design a specific GAN architecture [133, 134]. However, this might be an oversimplification of the problem, as discarding a majority of the colour information can also eliminate relevant diagnostic details [106, 108]. Thus, plenty of works adapt the idea of unpaired image-to-image translation methods, such as CycleGAN [12] since it does not require paired datasets.

StainGAN [94] is one of the first applications of the CycleGAN for stain normalisation, which shows superiour results to several classical approaches [126–128, 131]. Moreover, other works such as [95, 96] attest that modifications to the CycleGAN model's architecture or loss function could be beneficial in a given experimental setting. Generally, modification of the architecture or training strategy seems to have a significant effect on stain normalisation as several papers report an increase in performance compared to the original architecture by adjusting such parameters [96]. Plenty of works make modifications to the original CycleGAN, such as the loss function [94] or the architectural design, such as using a UNet [97] or a ResNet [94] generator. All of these modifications can lead to a final model that is large and therefore slow. Thus, Kang et al. [98] proposed the StainNet, based on distillation learning, to simplify the stain normalisation model and to make inference faster.

However, such an unpaired translation process can negatively affect the diagnostic quality of translations as the model is not constrained to keep such important information. Therefore, these models are extended with additional modules to enhance the translation process. For example, mechanisms such as self-attention [99] or self-supervised modules [100, 101] have been added.

CycleGAN-based models learn a mapping in two directions, although only one is usually used after training. Moreover, Moghadam et al. [108] criticise the use of such architectures since they could be prone to colour and structure alternation. As a result, specific GAN architectures have been developed to obtain better translations. Moghadam et al. [108] propose two solutions based on Garcia et al.'s model and StarGAN [14] to disentangle colour and structural features during the translation process. BenTaieb and Hamarneh [106] incorporate stain normalisation into a task-specific adversarial end-to-end framework. Furthermore, Liang et al. [107] propose a specific GAN architecture for normalising H&E images for the task of tumour classification where learning is guided by an auxiliary (pre-trained) classifier and specific losses are proposed in order to ensure that structural information is preserved. Nishar et al. [105] extended the idea of neural style transfer [135] with adversarial training to obtain stain normalisation. Zanjani et al. [102] adapt the idea of InfoGAN [136] to propose a framework for H&E stain normalisation. Kausar et al. [109] propose a new GAN architecture where two discriminators control stain normalisation to ensure correct colour translation and structure preservation.

Augmentation approaches: Some works try to overcome the need for normalisation by switching focus to learning a model robust to intra-stain variation. Traditional approaches are based on extensive stain-specific colour augmentation

[137–139]. For example, Tellez et al. [137] propose a specific H&E stain augmentation strategy based on deconvolution which artificially modifies the concentrations of haematoxylin and eosin in an image, generating a broad range of realistic H&E images. The idea of augmentation based on stain separation is further used by Faryna et al. [138] to adapt the RandAugment [140] technique for digital histopathology; or by Chang et al. [139] to propose the mix-up stain augmentation strategy. Moreover, Yamashita et al. [141] found that medically irrelevant style transfer used for augmentation is beneficial for a deep learning model's robustness. Recently, augmentation approaches have also been based on GANs [111–113, 115–117, 142] where their ability to generate high fidelity samples is used to augment a training set. Some methods [111, 113, 117] directly use generated samples for augmentation, while other approaches, such as [112, 142], consider that all virtually stained images are not equally beneficial for learning; thus they propose a specific schema to selectively use synthetic images for augmentation.

Adversarial domain adaptation: Attempts to reduce intra-stain variation in feature space via adversarial domain adaptation are also widely considered in the literature [118, 120, 121, 123, 143–145]. It is usually assumed that some annotated data are available, i.e. the source domain is annotated, while other stain variations are unannotated, i.e. the target domains. To be reminded, the basic model's architecture often contains a feature extractor, a task-specific branch trained on top of it using source data only, and a domain discriminator aiming to distinguish between the representations extracted by the feature extractor from the source and target domains. The discriminator is adversarially trained with a feature extractor to ensure that the domains from which the features originate are indistinguishable. The adversarial training is usually based on DANN [146] or ADDA [147] approaches. Frequently, even the basic models, sometimes adjusted to a specific architecture of feature extractor and discriminator, can already achieve satisfactory results, e.g. a solution based on the DANN approach [146] is used as a reference algorithm for the Mitosis Domain Generalisation Challenge 2021 [122]. Moreover, such ideas of adversarial training are also extended in the field of digital histopathology in several ways. For example, Hashimoto et al. [119] incorporated adversarial training with multi-scale multi-instance learning, while Graziani et al. [120] extended adversarial training with user-defined desired/undesired control targets. Although the final model in many of these approaches is still sensitive to stain variations not seen during training [138], the model trained in such a way, in a specific experimental design, can generalise better compared to standardisation and colour augmentation approaches [123]. Nevertheless, the conclusion can be different in different application areas, e.g. different tissues or tasks [118]. Thus, having a combination of adversarial training with stain normalisation or adversarial training with augmentation can have different effects on the model's robustness, depending on the task/tissue at hand.

2.5 GANs for Stain Transfer

One of the greatest challenges when performing stain transfer remains the availability of high-quality datasets. Therefore, it is not surprising that recent advances in stain transfer widely explore the potential of GANs. Two main groups of methods can be identified — methods that just aim to obtain a stain transfer between

different stainings without considering the reduction of domain shift introduced by intra-stain variation; and a group of methods that exploit stain transfer to reduce domain shift. The latter group can be subdivided into approaches that aim to develop a stain invariant model and those that result in a stain-specific model but provide a mechanism for them to be applied to other stains. The Table 2.5 summarises the most representative works for each category.

Stain-transfer oriented approaches: Several methods employ GANs-based stain transfer to obtain a virtual specimen having the appearance of multiple stains. The first attempts employ well-established GAN-based architectures to obtain the transfer. For example, Lahiani et al. [149] show that CycleGANs are able to translate between two immunohistochemical stainings; however, stain-specific expressions can be affected by the translation, which could interfere with diagnosis. Nevertheless, Lo et al. [155] have shown that in renal pathology the participating experts were unable to differentiate between real and CycleGAN-artificially produced microscopic kidney biopsy images, which confirms the visual plausibility of the obtained stain transfer. Similarly, an extensive Turing test study has been performed for liver pathology [163], which supports the visual plausibility of virtually stained images and even shows potential clinical application. Other approaches build on these findings to enhance the translation process, e.g. Lahiani et al. [164] incorporate a perceptual embedding loss function in the CycleGAN model to learn image embeddings that are less sensitive to colour, brightness and contrast variations in the input image. Several authors condition CycleGAN translations [151, 152] to force the preservation of diagnostically-relevant information. On the same line, Liu et al. [150] build upon the idea of CycleGAN to propose an architecture and training schema that ensures pathology consistency during the translation process. Nevertheless, Levy et al. [153] report superior results of pix2pix stain transfer compared to CycleGAN when virtually stained samples are used to identify melanocytic tissue in the subjective study involving both experts and non-experts. The recent work by de Haan et al. [58] show that their developed GAN-based framework is able to transform H&E samples into multiple special stains. Additionally, the study shows an improved diagnosis of several kidney diseases, which is promising for the potential clinical applications of these methods.

Stain-transfer for improving deep models: Bulten and Litjens [160] exploit the idea of adversarial training to propose an adversarial autoencoder able to perform unsupervised cancer detection. Recent approaches usually take advantage of well-established GAN-based architectures for image-to-image translation. Thus, Mercan et al. [156] demonstrated that synthetic images generated by a CycleGAN could be used to train deep models for mitosis detection and were able to obtain similar performance as when real samples are used. In parallel with work presented in Chapter 3 of this thesis, Gadermayr et al. [20] demonstrated for the case of multiple stainings that CycleGAN translations are able to effectively reduce domain shift introduced by inter-stain variation in kidney pathology. The proposed frameworks enable the application of stain-specific segmentation models to other stains by translating them to the source stain using CycleGAN models at test time. However, it is also noted that some translation directions are harder than others, and

Table 2.5 GAN-based approaches to reduce the effect of inter-stain variation.

Approach	Representative work(s)	Considered stains	Transfer Direction
Stain Transfer			
CustumGAN	de Haan et al. [58]	H&E, PAS, Masson's Trichrome, Jones silver	H&E to others
	UMDST [148]	H&E, MAS, PAS, PASM	multi-domain
CycleGAN	Lahiani et al. [149]	FAP-CK, Ki67-CD8	Ki67-CD8 to FAP-CK
CycleGAN - extension	PC-StainGAN [150]	H&E, Ki-67	H&E to Ki-67
	Xu et al. [151]	H&E, Ki-67	H&E to Ki-67
	cCGAN [152]	H&E, CK19/CK18	Bidirectional
pix2pix	Levy et al. [153]	H&E, Trichrome stains	H&E to Trichrome stains
DANN	Koga et al. [154]	H&E, CD20	CD20 to H&E
Better deep models			
stain specific-model			
CycleGAN	MDS, MDU [20]	PAS, AFOG, CD31, Col3	PAS to others, others to PAS
	Lo et al. [155]	H&E, PAS, Masson's Trichrome, Silver	H&E to others
	Mercan et al. [156]	H&E, PHH3	Bidirectional
CycleGAN-extension	DASGAN [157]	CK, PD-L1	CK to PD-L1
	Bouteldja et al. [158]	PAS, CD31, aSMA, Col3, NGAL	others to PAS
	Xing et al. [159]	H&E, Ki-67	Bidirectional stain transfer
adversarial training	CAAE [160]	H&E, CK8/18+P63	H&E to other
stain invariant-model			
CycleGAN	UDA-GAN [64]	PAS, H&E, Jones H&E, Sirius Red, CD68, CD34, CD3	PAS to others
adversarial training	SDA-sed, UDA-sed [161]	PASM, Masson	PASM to Masson
	DAPNet [162]	H&E, DAB-H	Bidirectional

consequently, domain shift reduction is stain-dependent. Similarly, Lo et al. [155] apply CycleGAN translation to obtain multi-stain glomeruli detection, while Wu et al. [165] propose fine-tuning a CycleGAN generator using a classification network. Moreover, Xing et al. [159] extend CycleGAN models with a task-specific module to enhance the translation process and obtain better nucleus quantification. In a similar manner, Kapil et al. [157] extend the CycleGAN model with an auxiliary

segmentation task while Bouteldja et al. [158] incorporate a pre-trained segmentation network to regularise CycleGAN training. Furthermore, this thesis in Chapter 4 proposes a combination of CycleGAN-based stain transfer and feature-space domain adaptation, which significantly enhances the adaptation process.

However, all of the previously mentioned methods tackle stain-specific models. Another line of research tries to build stain invariant solutions. From a computer-vision viewpoint, a stain invariant solution for a specific task can be obtained in a supervised manner by training a model on all available stains. However, obtaining annotations for each case is unfeasible. Thus, the primary constraint of obtaining a stain invariant solution is the limited amount of available annotations. Traditional methods use the standard practice in digital histopathology of staining consecutive slides differently. One of the slides is fully annotated by experts, while annotations in consecutive slides are obtained by mapping (registering) the annotated slide to the unannotated slides [68, 166]. Approaches that depend on consecutive slides are usually case-specific. Since they rely on a stain-specific source model, they are not applicable in the general case where the source slide's stain may vary. Alternative approaches exploit adversarial learning to obtain robust models. Mei et al. [161] build a specific GAN architecture for glomeruli segmentation in two different stains. Hou et al. [162] enhance adversarial training by using two discriminators that adversarially align features on different scales. This thesis in Chapter 4 demonstrates that feature space adaptation for particular inter-stain variation can result in a model able to segment in two stainings. The main drawback of such approaches is that the resulting feature extractor is biased toward domains seen during training. Therefore, the model is likely to fail when applied to stains not seen during training. Alternatively, inspired by the success of augmentation approaches for reducing intra-stain variation [106, 167], augmentation-based solutions for stain-invariant models are proposed [17, 64]. In this context, image manipulations, such as scaling, rotation, adding Gaussian noise or blurring, are particularly beneficial for increasing the robustness of deep learning models to stain variation [69]. Moreover, geometrical modifications to an image, such as elastic distortion, are helpful when limited data is available [168]. However, the nature of these augmentations can be regarded as too linear and an oversimplification of the variations that occur in the natural staining process [97]. Augmentation based on visually unconvincing virtual staining can have limited success [17] and GAN-based stain transfer, being able to obtain visually plausible images, can be highly effective as an augmentation strategy. This thesis in Chapter 4 (publication [64]) shows that CycleGAN-based translations used to augment the annotated dataset result in a robust model that works across several stainings, even those unseen during training. Additionally, in Chapter 5 (publication [65]), this thesis demonstrates that diverse translations (contrary to a deterministic offered by CycleGAN models) can furthermore improve the model's robustness. Contrarily to such augmentation approaches that result in (empirically) stain invariant models, alternative approaches exploit GAN-based stain transfer to integrate information from multiple stainings in order to enhance segmentation [41], or classification [165] performances. Although such approaches may also lead to stain invariant models that can generalise to unseen stains, the benefits of such methods are justified only in a particular experimental setting.

2.6 Discussion: Virtual Staining Perspectives

Two main areas of virtual staining application have been identified — increasing the diagnostic quality of the available histological data and building better deep learning-based solutions. Each of these aspects will be discussed separately.

2.6.1 Diagnostic Applicability

As pointed out by Van der Laak et al. [23], very few deep learning-based solutions reach the clinical application. When it comes to virtual staining, it can be expected that this number is even smaller.

One of the greatest challenges when generating artificial histological images is their evaluation in terms of plausibility and clinical quality. Contrarily to natural images, medical images have a more complex structure and details can have more significant meaning; therefore, important diagnostic information can be easily overlooked. As such, these types of approaches can only be applied in areas in which such changes are acceptable. Suppose the diagnostic purpose is the detection (e.g. counting) of morphologically consistent structures that are visible across multiple stain variations (such as glomeruli in case of kidney pathology). In that case, several GAN-based models can be used, among which the CycleGAN is the most common. In terms of visual plausibility, CycleGAN is robust to different architectural changes, as demonstrated in Chapter 3 (publication [63]) and also suitable for various task-related extensions, as attested by the numerous publications that modify it. Due to its limited capacity, the model preserves the shape and position of the global structures present in the image. However, as discussed in Chapter 3 (publication [62]) the model is able to change the appearance of stain-specific markers, which could interfere with a final diagnosis. From this perspective, some diagnostic applications are safe, e.g. glomeruli detection, counting or classification, but not recommended if the decision depends on cell position, which could be perturbed during the translation process. When the task at hand concerns specific structures/cell populations, task-related GAN-based methods can be more effective [111, 150].

Although the majority of GAN-based approaches consider virtual staining as a style transfer problem, the application of such approaches to the medical domain opens questions that go beyond the typical use of style transfer in natural images. Specifically, it is essential to address the possibility of misinterpretation of the images produced by GAN-based models in the medical domain [97, 156, 169]. The study of Xu et al. [152] gives an interesting perspective regarding the differences in the evaluation performed by medical experts (e.g. pathologists) and non-experts (e.g. computer vision researchers) — the medical experts were able to better notice errors in virtual staining compared to non-experts, to whom, virtually stained images look almost indistinguishable to real. Such findings additionally raise the importance of experts in the loop when developing and assessing the virtual staining methods.

Nevertheless, apart from diagnostically perturbed information, which trained experts can spot, GAN-based methods can produce human-imperceptible artefacts [62, 156, 158]. In Chapter 3 of this thesis (publication [62]) is shown that in the case of virtual staining between histochemical and immunohistochemical stainings, the translation could contain information embedded as imperceptible noise that encodes

the position of stain-specific markers. Moreover, this information can be perturbed in a way to vary the position and number of such markers in a plausible output image. This is just one type of hallucination for which it is known that GAN-based solutions in the medical domain are prone to [97, 156, 169] and it is possible that others are not yet known. These findings indicate that assessing the quality of translations using visual inspection, a widely adopted strategy in the field could be ill-advised.

Despite the success of the previously mentioned methods to achieve plausible virtual staining, it is important to note that the obtained translations cannot yet replace the real staining process. Stain transfer can affect the appearance of important diagnostic evaluation criteria, and thus, so far, artificially generated images cannot be truly relied upon for diagnosis purposes.

2.6.2 Effects on Deep Learning Models

Virtual staining has a wide application area in making better deep models. Most approaches in the literature consider stain variation as a domain shift that can be reduced in pixel space by virtual staining. The most frequent manner is to virtually re-stain the target stain (usually unannotated) during test time to look like it originates from the annotated domain on which the deep model is trained. Alternatively, virtual staining is used to build target stain-specific models by transforming the annotated domain to look like the unannotated target data during train time. Invariant approaches, able to work across multiple stains or stain variations, are considerably less studied. Such approaches usually rely on well-established GAN architectures, among which CycleGAN and pix2pix models are mostly adopted. However, due to its ability to work on unpaired datasets, CycleGAN-based approaches are predominant (see Table 2.5 and Table 2.4), despite the fact that the pix2pix model obtains biologically more justifiable translations [153, 156]. Although the original version of both architectures is already successful in many cases [19, 94, 104, 156], recent advances indicate that guiding learning by task-related modules (e.g. classification or segmentation networks) [157, 159] can have a positive effect on virtual staining and the underlying task at hand. These ideas are also extended by self-supervised and attention models [99–101].

Although many approaches rely on CycleGAN, there is little evidence and discussion about its limitations when applied to digital histopathology. A particularly significant concern, also raised by this thesis in Chapter 3, is hallucination effects [62, 150, 156], which can impact the performance of the underlying deep learning-based solution. As previously mentioned, it is known that CycleGAN-based models can introduce imperceptible noise during translations, which in specific contexts can be regarded as adversarial noise [63]. Recent attempts [158] try to isolate such noise from the translation; however, the obtained results suggest that some noise may still be present.

Most of the literature considers the problem of intra-stain variation, in which the H&E staining and its variation between different scanners or sites are mainly studied. Inter-stain variation is considerably less studied and rarely covers more than two stainings. This could be attributed to the lack of high-quality datasets. Due to privacy concerns, data are rarely publicly available, except for datasets

specifically released for public competitions. However, such data are usually related to specific tissue types or diseases, e.g. currently, cancer research is expanding, and thus the majority of available data collections contain H&E-staining samples. This could significantly shape the research advances towards specific studies [3] and could contribute to the scarcity of research related to other stainings and applications.

2.7 Conclusions and Research Opportunities

The general scarcity of annotation in digital histopathology and considerable sample variation introduced by the staining process pose an important challenge for deep learning-based solutions. Generative adversarial networks significantly contribute to the alleviation of these problems and open new application areas. GAN-based solutions used for virtual staining, under specific constraints and to some extent, enable the avoidance of the physical staining process. Extensive research has been made to exploit GAN-based stain transfer for more robust deep learning models, resulting in more accurate solutions. However, it also brings additional problems that open new research perspectives.

The primary challenge is how to assess the quality of the obtained translations. As attested in the literature, GAN-based solutions are able to obtain visually plausible results, but this is not sufficient to overcome the limitations that deep models face in practice. Chapter 3 (publication [63]) confirms this through extensive experiments that raise the importance of developing methods that are able to assess the usability of the obtained virtual staining. The first steps in this direction have been made by Nisar et al. [170].

Additionally, there is a noticeable lack of multi-domain stain transfer methods in the literature. This thesis fills the gap by 1) proposing the use of the StarGAN model in Chapter 3; and by introducing the HistoStarGAN model in Chapter 5, an end-to-end trainable framework for multi-domain stain transfer and stain invariant segmentation.

Furthermore, stain-invariant solutions are rare in the literature. The first such solution is proposed in Chapter 4 (publication [64]), and it is also the first to demonstrate generalisation across unseen stains. Additionally, a new model is introduced in Chapter 5 (publication [65]) that significantly improves stain invariance.

Moreover, a model proposed in Chapter 5 of this thesis (publication [65]) is able to generate vast collections of (artificial) multi-stain fully-annotated kidney pathology data, which will be realised for future use in the community upon validation by pathologists. Considering the general lack of annotated datasets in digital histopathology, this collection can contribute to kidney pathology research. Additionally, the proposed solution is general and has the potential to be extended for other applications, which can lead to the generation of diverse artificially-created datasets.

Chapter 3

Stain Transfer

Digital histopathology has become a rich area of innovation in both clinical application and research, where deep learning-based solutions achieve remarkable results [22]. However, nowadays, many state-of-the-art deep learning methods are data-hungry approaches which require huge collections of annotated data to be trained. Nevertheless, collecting medical data falls under strict law regulations, while obtaining high-quality annotations can be effectively performed only by trained experts [42]. All of that poses important constraints for the development of automated solutions. Moreover, already available data and annotations can be reused with limited success since the differences in tissue preparation and staining protocol highly affect the result of the staining process [7]. Figure 3.1 illustrates the result of staining kidney tissue sections using three different staining protocols. Since each staining provides specific information about the underlying tissue, general structures visible under multiple stains appear differently, e.g. glomeruli in the red circle in Figure 3.1. These differences represent a source of domain shift [9, 66] and therefore affect automatic systems [8, 19, 69]. Thus, exploring ways to address such variation becomes essential for successfully developing and applying automated systems.

One way to deal with stain variation is stain transfer — artificially changing the appearance of an image after its acquisition. The aim is to change the appearance of an image stained with stain A to look like as it has been stained with stain B. In Figure 3.1 the idea of stain transfer is illustrated by blue arrows on the right side of the image. The resulting B-stain like images are artificially generated and can be used, from a computer vision perspective, to reduce the domain shift and/or make models more robust to the stain variation. The hypothesis is that if stain transfer achieves visually convincing results, translations should be able to reduce domain shift between different stainings, e.g. deep models trained on real images from staining B should be able to extract similar features from B-stain-like images (translations from stain A).

This chapter investigates GAN-based methods for achieving stain transfer. In the Section 3.1 the scope and aims of the transfer are given. Section 3.2 proposes two GAN-based image-to-image stain transfer methods. Although these solutions have

[4]The images used for this illustration come from consecutive slides of the same kidney tissue. Thus, they largely represent the same anatomical structures.

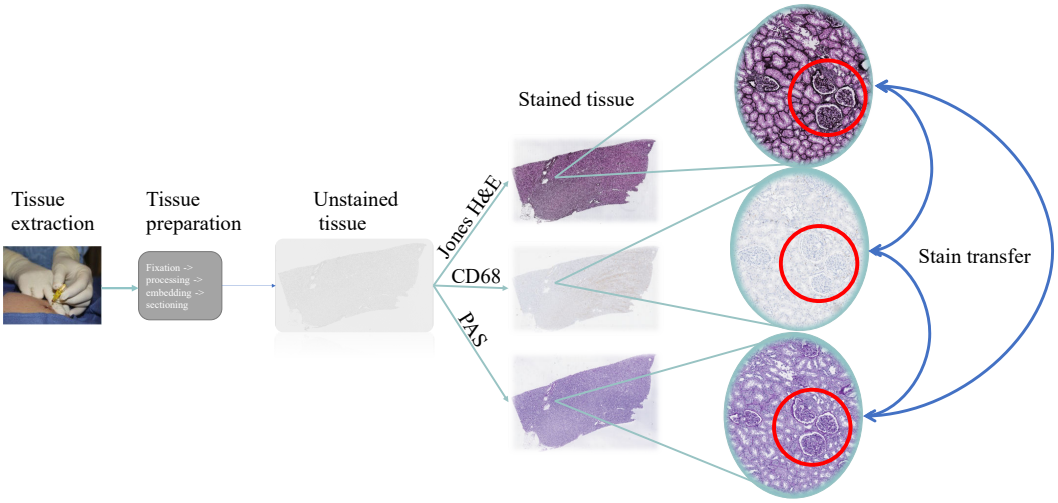


Figure 3.1: Example of a kidney tissue stained with different stains. The same tissue part [4]appears differently in different stains, e.g. glomeruli in the red circle.

become a standard way to approach stain variation in the field of digital histopathology, their limitations are rarely addressed. This chapter fills such gap and gives an in-depth discussion of some of the limitations of such solutions in Section 3.4 and Section 3.5. Furthermore, the findings presented in this chapter are used to propose a new augmentation strategy for supervised learning in digital histopathology which is presented in Section 3.6.

3.1 Scope and Aims

Stain transfer is considered, from a computer vision viewpoint, as a strategy to reduce domain shift for the tasks solvable across multiple stainings. It can be formulated as an image-to-image translation problem, where images stained by stain A are transformed to look like as they have been stained by stain B in a plausible way.

The term 'plausible' refers to the fact that an isolated histological image, without knowledge of adjacent sections processed with other staining modalities and in the absence of patient-specific information such as the underlying disease, looks visually correct to a trained expert with regard to the staining characteristics and the morphological appearance of the tissue components.

In this context, stain transfer is a many-to-many mapping between two stainings due to the differences in microscopic structures visible under different stains, as illustrated in Figure 3.2. The cardinality of these mappings primarily depends on the biological differences between stains, which particularly holds for translations between different groups of stainings. For example, immunohistochemistry stain CD68 marks a protein exclusively produced by macrophages while PAS, as a chemical reaction staining glycosylated proteins in general, can only highlight some parts of macrophages (co-located but not overlapping with CD68). During translation from PAS to CD68, there is no information about exact macrophage positions, and thus,

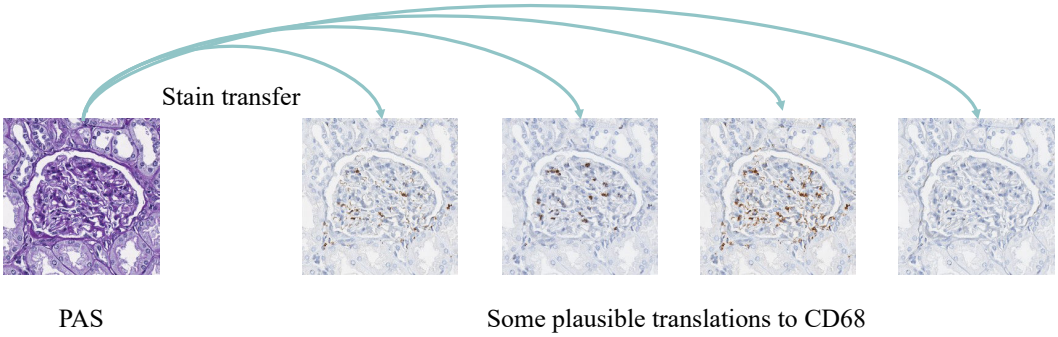


Figure 3.2: Illustration of many-to-many translations between PAS and CD68 — a real PAS-stained image can be translated to CD68 in many different plausible ways, e.g. by varying amount and position of macrophages (marked in brown).

the model is free to "invent" them as long as the overall image looks plausible (see Figure 3.2). Although this freedom in translation is not acceptable from a medical viewpoint (and consequently, their usage should be very limited), strictly speaking, from the mathematical side, there could be many possible ways to translate one PAS image into CD68 in terms of the amount and position of macrophages. Moreover, an additional factor which increases the cardinality of translations is the multiple variations in the staining process, such as contrast or colour intensity. Thus, an image stained with stain A can be validly translated into multiple images in stain B.

Defining stain transfer as a deep learning problem leads to the question of groundtruth. In the case of transfer between different stainings (e.g. PAS and CD68), it would be beneficial to obtain paired samples where the same image is stained with both stainings. That way, a model can learn to perform the translation in a more biologically acceptable way. However, this procedure can be very complicated to perform since de-staining can affect the underlying tissue and induce additional artefacts. An alternative approach is to use consecutive slides stained with different stains [164], but such comparisons are limited since the tissue structures vary between the slides. Thus, obtaining groundtruth for stain transfer tasks is not straightforward, and solutions which do not require such pairings are preferable. Furthermore, stain transfer considered from a computer vision viewpoint as a domain shift reduction strategy can be effectively applied only for the tasks solvable across multiple stainings — e.g. glomeruli segmentation as they can be observed in multiple stainings. That imposes additional constraints on the stain transfer problem that the task-related characteristics should be persevered during the translation.

Taking all that into account, the stain transfer aims to:

1. produce plausible results, in the sense of the definition given on page 34;
2. preserve task-related characteristics;
3. be unsupervised.

Evaluating the quality of the obtained stain transfer can be made by measuring their ability to reduce domain shift. It is important to note that this validation

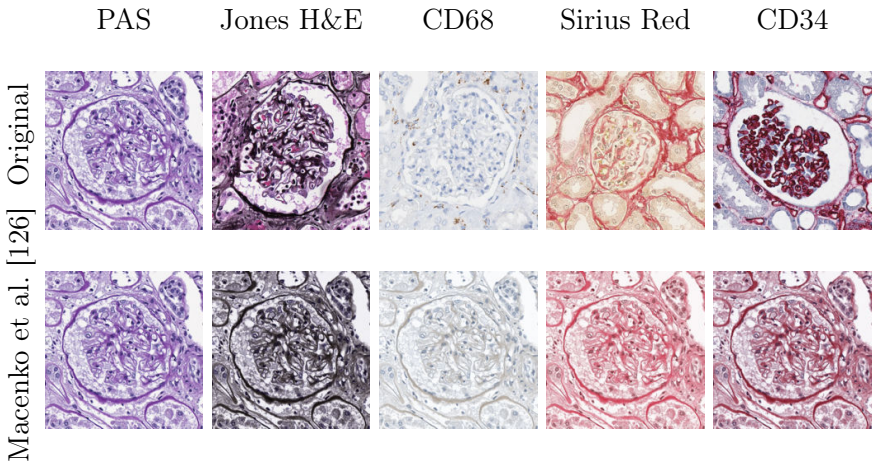


Figure 3.3: Macenko stain normalisation method used for stain transfer — first row: original glomeruli images from corresponding stains; second row: PAS image translated to other stains (image credits to Lampert et al. [17]).

strategy can be influenced by short-cut learning [171]. However, in the absence of better criteria, such a technique is well-established in practice [19]. In this chapter, it will also be used as an indication of the obtained stain transfer quality with respect to the task of glomeruli segmentation.

3.2 GAN-Based Stain Transfer

Given two stains A and B, the goal is to obtain a transfer model that is able to perform translation from a stain A to a stain B. In the case when B is a variation of stain A, this approach is called stain normalisation since it standardises appearance inside one particular stain. Historically, stain normalisation has been a wide field of research from both the diagnostic and automated system development point-of-view, which led to numerous classical and machine-learning-based approaches, as illustrated in Chapter 2 of this thesis. However, when stain B is not a variation of stain A, these approaches are not effective since the basic mechanisms on which they rely, such as colour deconvolution and standardisation, may not be relevant. Thus, such methods usually do not result in plausible outputs, as demonstrated in Figure 3.3 where well-known Macenko stain normalisation [126] is applied to translate images from PAS to other stains. The colour transfer is achieved by deconvolving the PAS image and applying its stain concentrations to stain vectors taken from another staining [17]. Obtaining plausible translations between different stains becomes possible with the introduction of Generative Adversarial Networks (GANs) [5, 53]. For the case of stain transfer, the generator can be used to modify an image from stain A in a way that it looks like as it has been stained with stain B such that the discriminator cannot make a differentiation between real B-stain images and translated A-stain (B-like) images. Moreover, since the main mechanism is distribution matching, the dataset does not necessarily need to be paired.

In the following, two such approaches will be investigated for the task of stain transfer. Subsection 3.2.1 introduces bi-directional stain transfer by CycleGAN

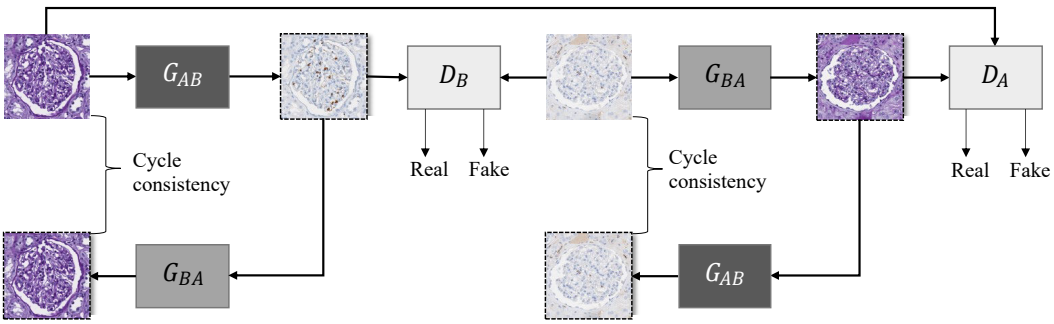


Figure 3.4: CycleGAN architecture for stain transfer.

[12], while Subsection 3.2.2 introduces multi-domain stain transfer by StarGAN [14]. Both methods will be evaluated from the perspective of their ability to reduce domain shift between different stain protocols in Section 3.3. Moreover, specific properties of these methods will be discussed in Section 3.4 and limitations will be given in Section 3.5.

3.2.1 CycleGAN for Stain Transfer

The CycleGAN [12] architecture is given in Figure 3.4. The model consists of two generators which perform translation between stains: $G_{AB} : A \to B$ to translate from stain A to B and $G_{BA} : B \to A$ to translate from stain B to A; in addition to two discriminators D_A and D_B. The aim of D_A is to distinguish between real A-stain images and those translated from B-stain to A-stain; while D_B aims to distinguish between real B-stain images and those translated from the A-stain to B-stain. These are trained using adversarial least-squared objective, such that

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G_{AB}, D_B, G_{BA}, D_A) = \mathbb{E}_{s\sim A}[(D_A(s)-1)^2] + \mathbb{E}_{t\sim B}[D_A(G_{BA}(t))^2] \\ + \mathbb{E}_{t\sim B}[(D_B(t)-1)^2] + \mathbb{E}_{s\sim A}[D_B(G_{AB}(s))^2].\end{aligned} \tag{3.1}$$

Moreover, training is constrained by the cycle-consistency and identity cost functions, which are formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) = \mathbb{E}_{s\sim A}[\|G_{BA}(G_{AB}(s)) - s\|_1] \\ + \mathbb{E}_{t\sim B}[\|G_{AB}(G_{BA}(t)) - t\|_1],\end{aligned} \tag{3.2}$$

and

$$\begin{aligned}\mathcal{L}_{\text{identity}}(G_{AB}, G_{BA}) = \mathbb{E}_{s\sim A}[\|G_{BA}(s) - s\|_1] \\ + \mathbb{E}_{t\sim B}[\|G_{AB}(t) - t\|_1].\end{aligned} \tag{3.3}$$

Thus, the full objective is

$$\begin{aligned}\mathcal{L}_{\text{CycleGAN}}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{\text{adv}}(G_{AB}, D_B, G_{BA}, D_A) \\ + w_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) \\ + w_{\text{id}}\mathcal{L}_{\text{identity}}(G_{AB}, G_{BA}),\end{aligned} \tag{3.4}$$

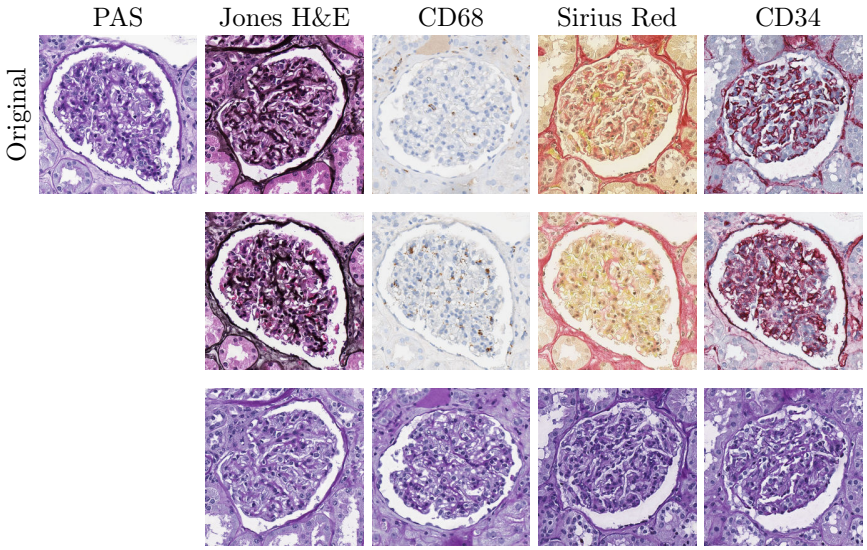


Figure 3.5: Illustration of stain transfer obtained with CycleGAN models. The first row contains real images from each staining. The second row represents the translations $T_{PAS \to X}$ of a PAS image to the target staining. The last row represents translations $T_{X \to PAS}$ of the real target image to the PAS staining.

where w_{cyc} and w_{id} control the relative importance of the cycle-consistency and identity losses, respectively. Training details are given in the Appendix B.2.

Once trained, the CycleGAN model is able to perform translations $T_{A \to B}$ and $T_{B \to A}$ between two stains A and B using the corresponding generators. The result of CycleGAN translations $T_{PAS \to X}$ and $T_{X \to PAS}$, where $X \in \{\text{Jones H\&E, Sirius Red, CD68, CD34}\}$ are given in Figure 3.5. All translations look plausible, as confirmed by pathologists.

In more detail, given an input image from stain A, generator G_{AB} translates it to look like stain B. The input image is the only information based on which this generator performs the translation. At this stage, the generator could translate the input image freely as long as the result is a plausible image in stain B. Thus, a trivial valid translation would be to translate a given image from stain A into a single image which looks like stain B. However, the choice of loss function and training procedure prevents such an outcome. The cycle-consistency term forces reversibility of the translation processes, which imposes additional constraints on the generator. More specifically, cycle-consistency requires that the opposite generator G_{BA} is able to reconstruct the exact image that was input to G_{AB} (pixel-space distance). This imposes indirect limitations on the translation process since the opposite generator also has a translated image from staining A as the only input, based on which it needs to recover the original image accurately. Thus, it would be easier for both generators to keep as much common information as possible (e.g. overall structure) than to perform more complicated translations that need to be reversed by the opposite generator. In practice, the architectural design and training procedure of this method enables stain transfer in a way that the internal structures are not

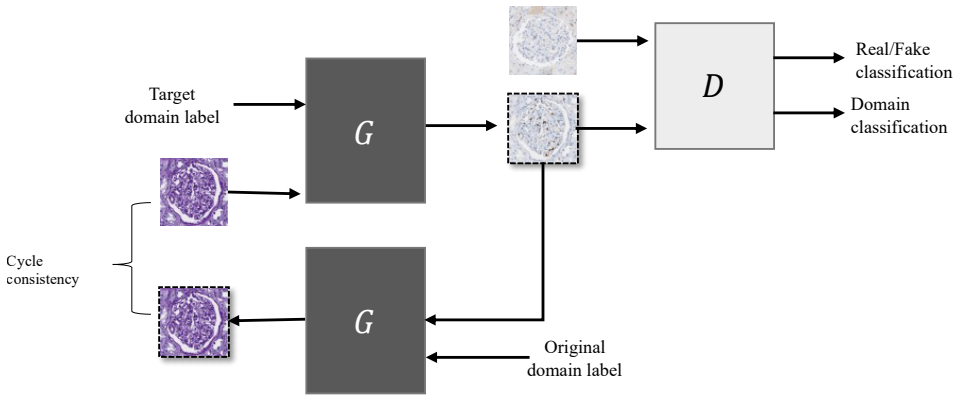


Figure 3.6: StarGAN architecture for stain transfer.

affected, e.g. glomeruli stay in the same position (shape, orientation) before and after translation. The method is general, not related to a specific stain combination (contrarily to methods specifically dedicated to stain normalisation), and thus it can be applied for translation between any pair of stains.

3.2.2 StarGAN for Stain Transfer

When it comes to translations between multiple stainings, training a model to translate between each pair quickly becomes impractical as the number of stains increases. It would be beneficial to develop a single multi-domain stain transfer model, i.e. a model capable of translating between multiple stains upon training. As the pioneer in multi-domain image-to-image translation, the StarGAN [14] model can be employed to achieve such multi-domain stain transfer. The model architecture is presented in Figure 3.6. It contains one conditional generator G_*, conditioned on the domain label (stain), which translates an input image x_i from stain i to image that have the characteristics of stain j, x_j, i.e. $G_*(x_i, j) \rightarrow x_j$; and one multi-task discriminator D_* that simultaneously distinguishes between real and generated samples (D_{adv}) and classifies each image to a domain (D_{stain}), i.e. $D_*(x) \rightarrow (D_{stain}(x), D_{adv}(x))$. The Generator G and D_{adv} play an adversarial game, making the Generator produce samples indistinguishable from real images. The classification branch D_{stain} guides the Discriminator to recognise the real image's domains (stain) correctly. Regarding the Generator's optimisation, D_{stain} forces it to produce fake samples indistinguishable from real samples of that domain (stain). As such, a single discriminator controls the translation to multiple stains. The overall objective function is:

$$\mathcal{L}_{\text{StarGAN}}(G_*, D_*) = \mathcal{L}_{\text{adv}}(G_*, D_*) + w_{\text{cyc}}\mathcal{L}_{cyc}(G_*) + w_{\text{cls}}\mathcal{L}_{cls}(G_*, D_*). \tag{3.5}$$

In order to obtain adversarial training stability, StarGAN [14] uses Wasserstein objective with a gradient penalty instead of the original negative log-likelihood, so

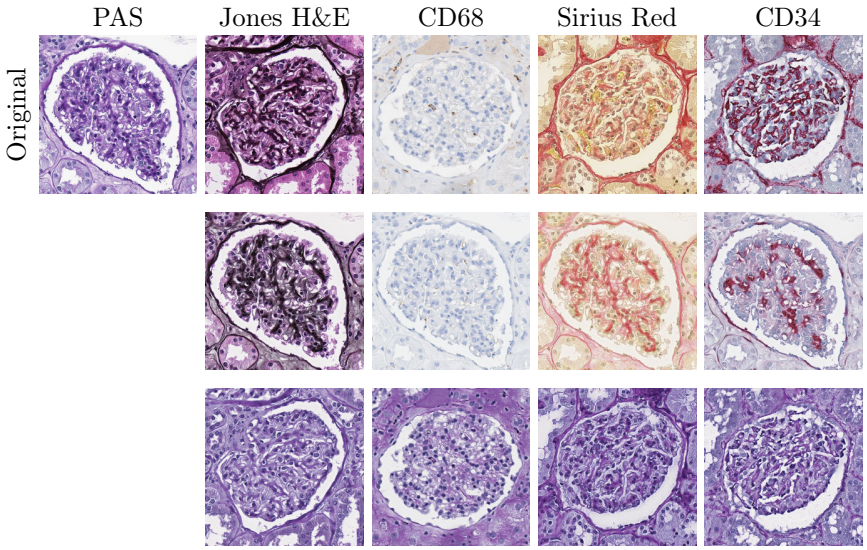


Figure 3.7: Illustration of stain transfer obtained with the StarGAN model. The first row contains real images from each staining. The second row represents the translations of a PAS image to the target staining. The last row represents the translation of the real target image to the PAS. As in Figure 3.5, the obtained translations look plausible.

the following individual objectives are used:

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G_*, D_*) &= \mathbb{E}_{x\sim P_j(x)}[D_{\text{adv}}(x)] - \mathbb{E}_{x\sim P_i(x),j}[D_{\text{adv}}(G_*(x,j)))] \\ &\quad + \lambda_{gp}\mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D_{adv}(\hat{x})\|_2 - 1)^2], \qquad (3.6)\\ \mathcal{L}_{\text{cyc}}(G_*) &= \mathbb{E}_{x\sim P_i(x),i,j}[\|G_*(G_*(x,j),i) - x\|_1], \qquad (3.7)\\ \mathcal{L}_{\text{cls}}(G_*, D_*) &= \mathbb{E}_{x\sim P_i(x),j}[\log D_{\text{stain}}(j|G_*(x,j))] \\ &\quad + \mathbb{E}_{x\sim P_i(x)}[\log D_{\text{stain}}(i|x)], \qquad (3.8)\end{aligned}$$

where $\hat{x}$ is sampled uniformly between the real and generated images [14]. Once trained, StarGAN is able to translate in multiple directions — between any pair of training stains. Training details are given in Appendix B.

The same translations as with CycleGAN obtained by this model are presented in Figure 3.7. Since the obtained model is multi-domain, it is possible to translate between any pair of stains seen during training, as illustrated in Figure 3.8. The first column contains real images from each staining. Other columns represent their translations to other stainings using the same StarGAN model. Similarly, as for the case of CycleGAN, obtained translations look plausible.

Although there are plenty of other methods for multi-domain unpaired image-to-image translation which outperform StarGAN in the domain of natural images, such as StarGANv2 [15] or TUINT, [16], their application for stain transfer is not straightforward (as discussed in Chapter 5 of this thesis). The majority of these build upon CycleGAN's principle of constraining the training with cycle-consistency. However, the way in which it is imposed depends on the purpose and can directly

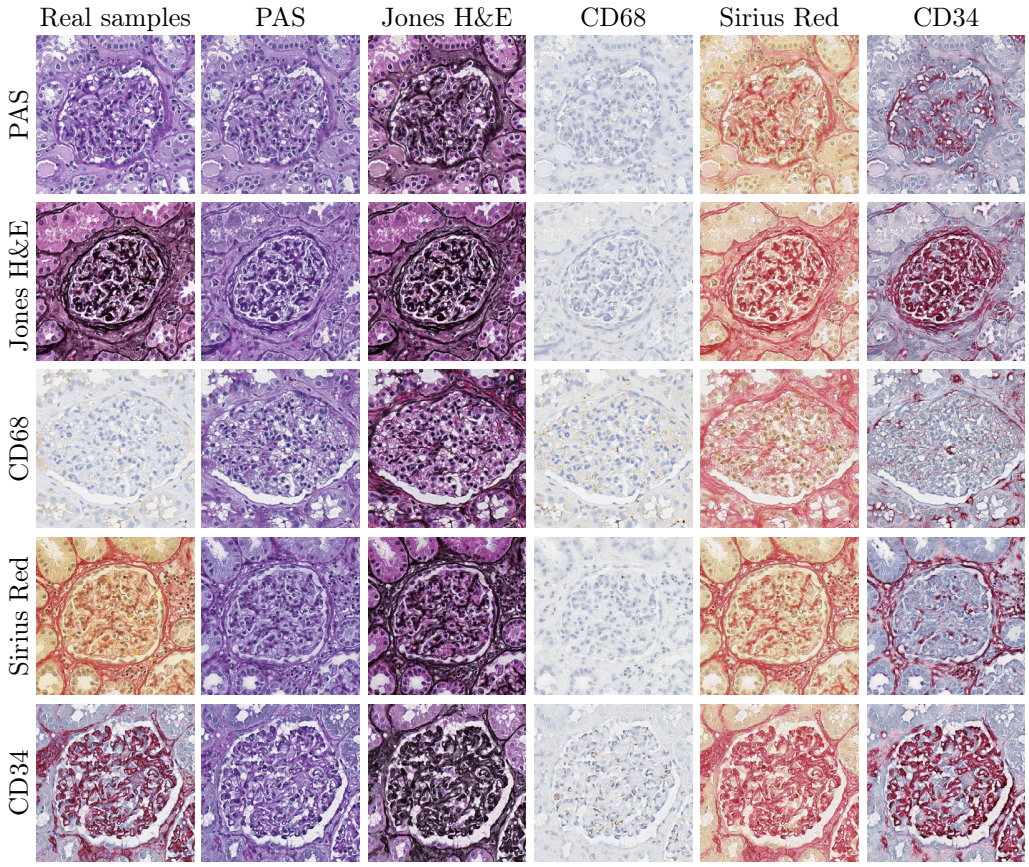


Figure 3.8: Illustration of multi-domain stain transfer obtained with the StarGAN model. The first column contains real images from each staining. Each row contains translations of given images to the corresponding staining.

influence translations. The differences between the aim of performing translation in natural images and stain translation become more important as more advanced architectures are developed. For example, in order to obtain a realistic output when translating a dog to look like a cat, strong geometrical changes are required (like changing a shape of a head). Thus, to produce realistic output, advanced architectures such as StarGANv2 [15] build additional modules that enable the generator to perform large changes to an image. However, when it comes to stain transfer, these changes could lead to the removal/invention of important tissue structures, limiting the application of the translation from both a medical and a computer vision point-of-view. Thus, creating a multi-domain stain transfer model is a challenging task. Contrarily to such methods, the StarGAN has limited capacity to perform structural changes during the translation process. The translation process is controlled by a single generator that receives the image and fixed domain label as the only inputs. Similar to CycleGAN translations, it is 'easier' to keep the structure during the translation process than to invent it during reconstruction.

Table 3.1 F_1-scores for the glomeruli segmentation baseline results (standard deviations are in parentheses).

	PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
F_1	0.907 (0.009)	0.864 (0.011)	0.853 (0.018)	0.867 (0.016)	0.888 (0.015)	0.876 (0.022)
Precision	0.885 (0.023)	0.821 (0.019)	0.849 (0.024)	0.786 (0.036)	0.849 (0.033)	0.838 (0.037)
Recall	0.932 (0.014)	0.912 (0.005)	0.858 (0.020)	0.963 (0.015)	0.931 (0.010)	0.919 (0.039)

3.3 Results — Domain Shift Reduction

The ability of CycleGAN/StarGAN-based stain transfer models to reduce domain shift is evaluated for the task of glomeruli segmentation. For each stain, a baseline model is obtained in a fully supervised way. Baseline performance for each stain is determined and presented in Table 3.1. These results indicate that glomeruli segmentation is possible in each of the considered stainings. However, each baseline model makes a prediction based on a different set of features extracted from a training dataset. If stain transfer gives a plausible result in a targeted stain, one can expect that the baseline model from that stain will be able to recognise the same set (or subset) of features in the translated images. In this regard, two directions of translations are considered. First, translations from other stains to PAS is evaluated on PAS pre-trained models, Table 3.2. Second, translation from PAS to all other stains is evaluated on stain specific pre-trained models, Table 3.3.

The baseline models are evaluated on test images from target stains prior to stain transfer — Table 3.2, the vPAS row represents the application of PAS models to other stains; Table 3.3, the vStain row represents the application of pre-trained models from other stains to PAS images. Since the baseline models are trained to be applied to one particular stain (used for training), the models fail to recognise glomeruli in other stainings due to domain shift introduced by different stainings. However, when stain transfer is applied, a significant increase in performance is observed, which confirms that the introduced stain transfer methods can reduce domain shift, see CycleGAN/StarGAN rows in Table 3.2 and Table 3.3.

Although the obtained translations look plausible, stain transfer is not equally successful in reducing a domain shift in all stain combinations. Even though the translations obtained using both CycleGAN and StarGAN look plausible (see Figure 3.5 and Figure 3.7 for comparison), it can be observed that the translation model (StarGAN vs CycleGAN) greatly influence the quantitative results. For example, in Table 3.3 the same set of pre-trained PAS models are applied to the same test images from target stainings translated to PAS using different stain transfer models. Thus, the difference observed in quantitative results can be attributed to the stain transfer model. A similar conclusion can be drawn by comparing the results in Table 3.3.

Regardless of its superiority over StarGAN, CycleGAN-based stain transfer was also not equally successful in reducing domain shift across all the tested stainings.

Table 3.2 Stain transfer using CycleGAN/StarGAN to reduce domain shift between different stains. vPAS represents the direct application of the pre-trained PAS model to other stains without translating data; CycleGAN/StarGAN represents the results obtained by translating PAS to a given stain during test time using CycleGAN/StarGAN models.

Training Strategy	Score	Test Staining					
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
vPAS	F_1	0.907 (0.009)	0.085 (0.034)	0.001 (0.002)	0.016 (0.018)	0.071 (0.063)	0.043 (0.041)
	Precision	0.885 (0.023)	0.055 (0.021)	0.097 (0.129)	0.034 (0.034)	0.257 (0.243)	0.111 (0.101)
	Recall	0.932 (0.014)	0.418 (0.316)	0.001 (0.001)	0.073 (0.101)	0.058 (0.039)	0.137 (0.190)
CycleGAN	F_1	-	**0.866 (0.017)**	**0.637 (0.034)**	**0.880 (0.015)**	**0.754 (0.033)**	**0.789 (0.223)**
	Precision	-	0.842 (0.035)	0.846 (0.050)	0.846 (0.031)	0.879 (0.027)	0.853 (0.018)
	Recall	-	0.894 (0.020)	0.516 (0.058)	0.918 (0.008)	0.662 (0.059)	0.747 (0.193)
StarGAN	F_1	-	0.756 (0.086)	0.092 (0.055)	0.599 (0.108)	0.751 (0.033)	0.550 (0.314)
	Precision	-	0.675 (0.136)	0.242 (0.116)	0.496 (0.123)	0.742 (0.092)	0.539 (0.223)
	Recall	-	0.881 (0.029)	0.061 (0.044)	0.780 (0.099)	0.774 (0.070)	0.624 (0.379)

For the case of Jones H&E and Sirius Red, translation to PAS obtains close to baseline results on pre-trained PAS models, which indicates that domain shift is greatly reduced. In the case of CD68, despite the visual quality of the obtained translations, performance gain by stain transfer is significantly worse. Since all results are determined using the same set of PAS pre-trained models, it can be concluded that the success of stain transfer also depends on the differences between stains. PAS, Jones H&E and Sirius Red stainings all mark general tissue structure, and thus mapping between stains are less complicated since the difference in visible structures is not huge. This is contrary to the immunohistochemistry stains CD68 and CD34, which specifically mark macrophages and blood vessel endothelium respectfully. In these cases, the position of stain-specific markers needs to be deduced. Furthermore, a similar variance between stains is observed in the opposite translation direction, Table 3.3. Therefore, the stain difference seems crucial for the reduction of domain shift as the translations between biologically closer stains result in overall better domain shift reduction.

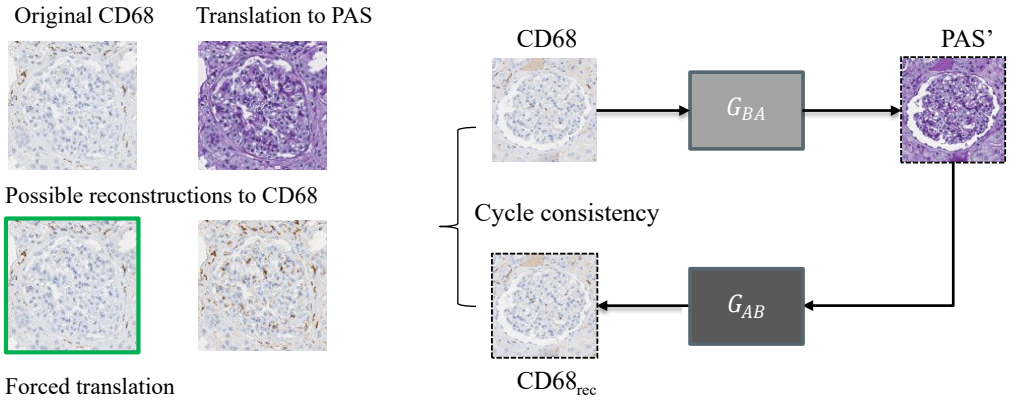
3.4 Discussion

The previously presented results show that architectural design and stain combinations can affect the quality of stain transfer when used for domain shift reduction. The potential reason can be that stain transfer is naturally non-deterministic, a many-to-many mapping that is reduced by CycleGAN/StarGAN architectures to a

Table 3.3 Stain transfer using CycleGAN/StarGAN to reduce domain shift between different stains. vStain represents the direct application of pre-trained models from the corresponding stain to PAS without translation; CycleGAN/StarGAN represents the results obtained by translating PAS to a given stain during test time using CycleGAN/StarGAN models.

Training Strategy	Score	Test Staining				
		Jones H&E	CD68	Sirius Red	CD34	Overall
vStain	F_1	0.029 (0.039)	0.006 (0.006)	0.000 (0.000)	0.029 (0.021)	0.016 (0.015)
	Precision	0.026 (0.027)	0.385 (0.258)	0.001 (0.002)	0.481 (0.164)	0.223 (0.246)
	Recall	0.062 (0.106)	0.003 (0.003)	0.000 (0.000)	0.015 (0.011)	0.020 (0.029)
CycleGAN	F_1	**0.891 (0.003)**	**0.608 (0.083)**	**0.813 (0.025)**	**0.590 (0.063)**	**0.725 (0.150)**
	Precision	0.877 (0.003)	0.834 (0.033)	0.846 (0.044)	0.848 (0.023)	0.851 (0.018)
	Recall	0.905 (0.007)	0.485 (0.096)	0.788 (0.072)	0.456 (0.080)	0.659 (0.223)
StarGAN	F_1	0.777 (0.037)	0.568 (0.077)	0.661 (0.110)	0.024 (0.026)	0.508 (0.334)
	Precision	0.899 (0.014)	0.500 (0.066)	0.752 (0.076)	0.299 (0.121)	0.613 (0.266)
	Recall	0.669 (0.052)	0.706 (0.114)	0.610 (0.201)	0.006 (0.003)	0.498 (0.330)

deterministic, i.e. one-to-one, mapping between stains. As an illustration, lets take again as an example a translation between PAS and CD68, see Figure 3.9. As a reminder, CD68 marks a protein exclusively produced by macrophages, while PAS, as more general staining, can only highlight some parts of macrophages (co-located but not overlapping with CD68). When performing translation from PAS to CD68, the generator $G_{PAS-CD68}$ should produce some macrophages as they appear in the CD68 image distribution. Since PAS does not contain information specifically related to macrophages, the model is free to invent them. Thus, there are multiple possible translations to CD68, and all of them can differ in terms of the appearance of macrophages. However, taking an image from CD68 that contains macrophages at specific positions, $G_{CD68-PAS}$ can ignore macrophages since their appearance is not related to the PAS staining. Nevertheless, cycle-consistency enforces that the reconstruction of this image performed by $G_{PAS-CD68}$ recovers exactly the original image, meaning that it requires macrophages in the same positions as in the original. That way, the generator $G_{PAS-CD68}$ is forced to perform a one-to-one mapping. Similarly, the opposite direction of translation is also forced to be a one-to-one mapping. The same conclusion holds for the StarGAN architecture. The fact that the generator G_* receives an image and a fixed domain label as the only inputs defines the deterministic nature of the mapping between stains in a similar way to the CycleGAN model, i.e. the model needs to reconstruct exactly the same image in CD68 given its translation to any other stain. However, the StarGAN can be more constrained in the translation process compared to CycleGAN. Taking into account that the model needs to learn mappings between several different stains at the same



(a) Multiple possible translations from PAS to CD68

(b) Cycle-consistency requires deterministic reconstruction

Figure 3.9: CycleGAN — implicit deterministic mapping.

time, it is forced to learn deterministic mappings between any pair of stainings seen during training. With the number of stains increasing, this can lead to improper translations between some pairs of stainings.

This brings specific limitations related to practical application and evaluation, such as:

Translation quality: the obtained translations might encode additional information in order to ensure a deterministic mapping. This can affect the ability of translations to reduce a domain shift between stainings.

Training stability and reproducibility: since there are no explicit stopping criteria, one can stop training when plausible translations are obtained. However, different deterministic mappings can be obtained in different training stages, which could lead to misleading conclusions about the quality of the obtained translations.

Generalisability: different stain combinations could encode different information, giving ambiguous conclusions related to the application of stain transfer models.

Bashkirova et al. [172] indicate that the cycle-consistency loss in CycleGAN-based models $\mathcal{L}_{\text{cyc}}$, Eq. (3.2), implicitly forces them to hide information necessary for proper reconstruction of B_{rec} and A_{rec} in translations A' and B' in the form of imperceptible low amplitude, high frequency noise. It is reasonable to assume that such noise introduces a domain shift when pre-trained models are applied to the translated images, which can explain the differences in the results given in Section 3.3. Given the StarGAN model, which is constrained by cycle-consistency in the same manner as CycleGAN, it can be assumed that StarGAN stain transfer also injects imperceptible noise. Bashkirova et al. [172] suggest that such hidden information can be perturbed by additive Gaussian noise.

To explore this phenomena for digital histopathology, the composition of translations PAS $\rightarrow$ Target $+ \mathcal{N}(0, \sigma) \rightarrow$ PAS is used, where $\mathcal{N}(0, \sigma)$ is a zero-mean Gaussian distribution with standard deviation σ. Figure 3.10 presents CycleGAN

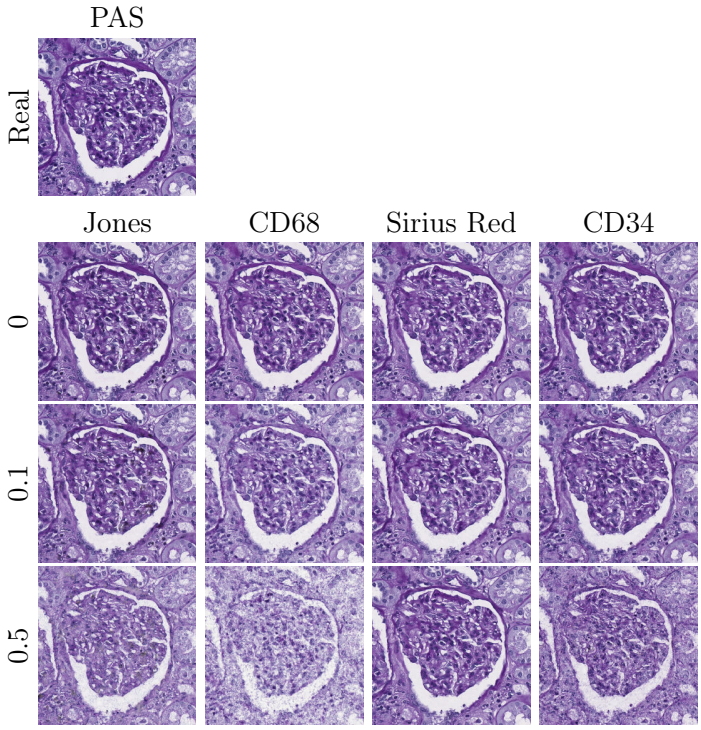


Figure 3.10: CycleGAN — the effects of additive zero-mean Gaussian noise added to intermediate representations of PAS images on their reconstructions.

reconstructions of the same PAS image after translation to each target stain, with different standard deviations of additive noise. This confirms that not all stain translators encode information in the same way. For example, adding noise with a standard deviation of 0.5 to the CD68 intermediate stain results in a higher reconstruction error than adding the same noise in the Sirius Red intermediate stain. It is hypothesised that the noise level in each target staining correlates with the difficulty of translation, i.e. more complex translations require more noise. Compared to the CycleGAN, the StarGAN appears to be more sensitive, as illustrated in Figure 3.11 for the reconstructions of a PAS image when intermediate translations to a given target stain are corrupted. This is probably because StarGAN performs a much harder task of multi-domain stain translation — to properly reconstruct its input, more information needs to be hidden. In this regard the lower performances for StarGAN translations observed in Table 3.2 and Table 3.3 can be understand as well. A similar conclusion holds for the other direction of CycleGAN/StarGAN translations, whose results are provided in Appendix B.5.

These observations, and recent studies on the adversarial nature of the CycleGAN model [173, 174], lead to the hypothesis that translations suffer from invisible artefacts produced during translation. The extent and type of these artefacts could be related to the differences between stainings. Stain pairs with a more significant difference in the highlighted structures require more complicated translation, forcing hallucination of specific features, as also confirmed by Mercan et al. [156]. Such finding is further exploited in Section 3.6 to propose an augmentation strategy for

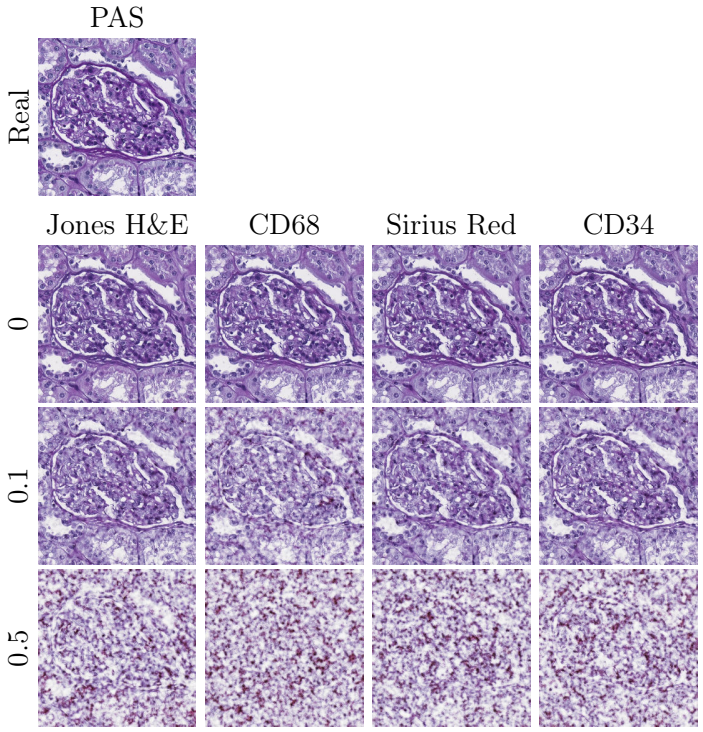


Figure 3.11: StarGAN — the effects of additive zero-mean Gaussian noise added to intermediate representations of PAS images on their reconstructions.

supervised training. Nevertheless, all of that emphasises the care and consideration needed for the direct application of stain transfer in the medical domain. When used in clinical practice, the implications should be held up to even greater scrutiny.

3.5 Limitations — Is Seeing Really Believing?

The findings presented in the previous sections, that CycleGAN-based stain transfer can achieve plausible results and therefore reduce domain shift introduced by stain variation, have been confirmed by numerous works in the literature [20, 64, 67, 94, 99]. Many works propose a modification to the original CycleGAN architecture [20, 95], its loss function [94] or, with respect to a specific task, extension with additional modules [99, 100]. However, due to lack of proper groundtruth, evaluation of the obtained translations is visual [98, 99]. Assuming that the translation results in high fidelity, these methods are more often used in the computer vision domain to reduce domain shift [19, 94]; or as a domain augmentation strategy to reduce the need for additional annotations [64, 111]. Since these approaches are becoming more commonplace, and new possibilities are being explored, such as multi-stain segmentation [64] or improving tumour classification [151], it is of great importance to raise awareness of the sensitivity of such methods to some common, and rather small, changes. In the following, it will be demonstrated that even the most simple architectural choice in CycleGAN-based models can play an important role in the

ability of the obtained models to reduce domain shift, even though visual appearance is not affected. Although most models produce plausible translations, i.e. visually indistinguishable from real samples, the huge performance difference observed in pre-trained models when applied to translated images confirms that the quality of translations differs. In order to limit the number of experimental degrees of freedom, the modifications to the original CycleGAN architecture are restricted to the normalisation layer. In the original architecture Instance normalisation is used, and in this study this is varied to other approaches commonly found in the literature: Batch, Layer, and Group. It is shown that the translations obtained by varying the normalisation layer belong to different data distributions, distinct from these of real samples, causing pre-trained models to perform badly. Furthermore, since manual visual inspection cannot determine a difference in quality between the translations, it follows that visual inspection cannot be used as a validation criterion for virtual staining.

3.5.1 Experimental Setup

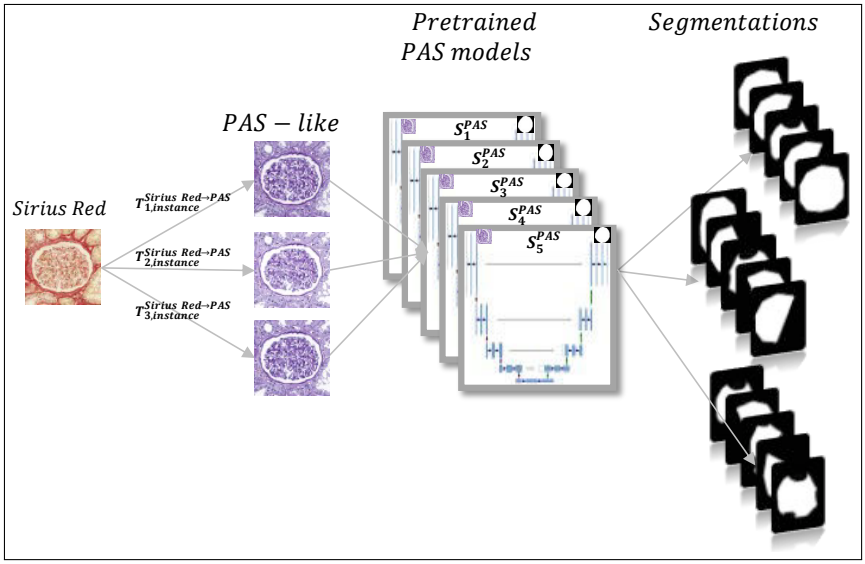
To demonstrate the CycleGAN-based model's sensitivity to the underlying architecture, the original CycleGAN architecture is taken and altered by replacing the normalisation layers in both the discriminators and the generators. As previously, to quantitatively measure the quality of obtained translations, their ability to reduce domain shift introduced by stain variation is determined for the task of glomeruli segmentation. For ease of reading, the stain on which the segmentation model is trained in a supervised manner is referred to as the source stain, and the stain that is translated to the source stain during application is the target stain.

Two sets of complementary hypotheses concerning what can affect the performance measured in this setting are identified:

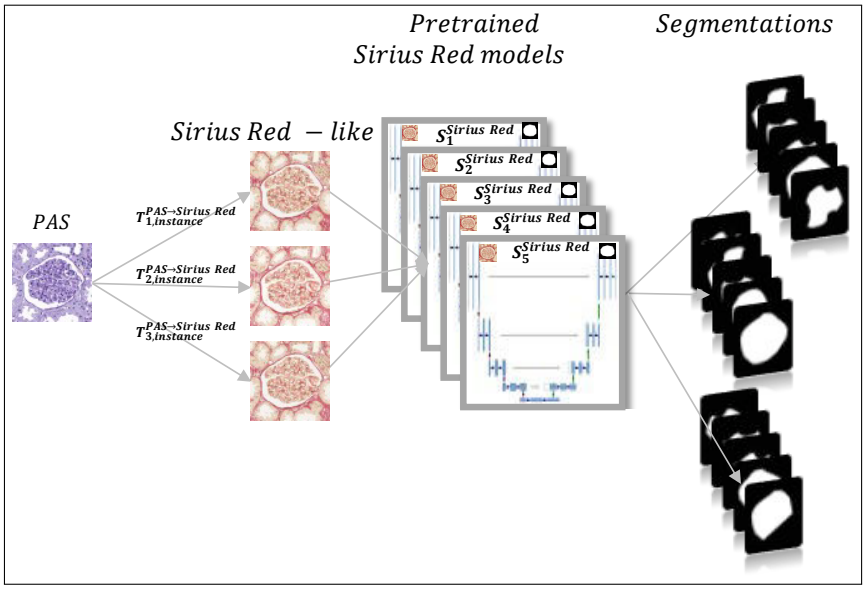
1. Pre-trained model
    (a) Short-cut learning [171] in pre-trained models: a model makes a decision based on some source dataset characteristics that are not necessarily related to the given problem. Thus, if the translated images do not contain the shortcut characteristics, the pre-trained model will not perform well.
2. Stain transfer
    (a) Stain transfer model: the model's ability to produce accurate translations between the target and source stains should impact downstream task performance.
    (b) Direction of translation: some stain translation directions may be harder (e.g. translation from a general purpose stain to a specific stain).

In order to test these hypotheses, several experiments are conducted, as illustrated in Figure 3.12 (stains taken for illustration are PAS and Sirius red, translation model has Instance normalisation layer. The same experiments are performed for other combinations).

In the case of inter-stain variability, this analysis is performed from two perspectives:



(a) Target to PAS experiments illustration.



(b) PAS to target experiments illustration.

Figure 3.12: Experiments design illustration: (a) Target to PAS — models trained on PAS data are evaluated on different target stains and translation models; (b) PAS to Target — models trained on target stain are evaluated on different translations from PAS stain.

- Target to PAS (Figure 3.12a): Five PAS segmentation models are trained (S_1^{PAS}, S_2^{PAS}, ..., S_5^{PAS}) and their performance is evaluated on translations from four other stainings. For each normalisation layer and each stain, three translation models are trained, i.e. $T_{1,n}^{x \to \text{PAS}}$, $T_{2,n}^{x \to \text{PAS}}$, $T_{3,n}^{x \to \text{PAS}}$, where, $x \in \{\text{Jones H\&E, Sirius Red, CD68, CD34}\}$ and $n \in \{\text{Instance, Batch, Layer, Group}_8, \text{Group}_{16}, \text{Group}_{32}, \text{None}\}$. In this way,

one pre-trained segmentation model, e.g. S_1^{PAS}, is applied to the translations from all stains, $T_{i,n}^{x\to\text{PAS}}$, allowing analysis of the effects of a stain transfer model (Hypothesis 2.1) and target stain (Hypothesis 2.2). The three stain translation models that are obtained for each combination of target stain and normalisation layer allow the measurement of short-cut learning (Hypothesis 1.1). The standard deviation in the performances using different stain translation models obtained in the same experimental setting can be attributed to a pre-trained model's bias.

- PAS to target (Figure 3.12b): Five segmentation models are trained for each of the target stains (S_1^x, S_2^x, ..., S_5^x), where $x \in$ {Jones H&E, Sirius Red, CD68, CD34}. These are evaluated on translations from PAS to each stain, using different stain translation models $T_{1,n}^{\text{PAS}\to x}$, where $n \in$ {Instance, Batch, Layer, Group_8, Group_{16}, Group_{32}, None}. As such, the test images and the segmentation models within one target stain remain constant, and therefore the variation in results within stains can be attributed to translation quality (Hypothesis 2.1). Moreover, by comparing the results from the previous experiment, the influence of translation direction can be investigated (Hypothesis 2.2). Similarly, as previously stated, the standard deviation within several runs of the same experimental setting can be related to short-cut learning (Hypothesis 1.1).

In the case of intra-stain variability, PAS pre-trained models' sensitivity is measured to the translation (stain normalisation) from the publicly available AIDPATH dataset [18]. From this perspective, hypotheses 1.1 and 2.1 can be investigated.

Normalization layers: In the case of 2D images, a feature computed by a model's layer, x, is a 4D tensor $x = (N, C, H, W)$ where N denotes the batch size, C is the number of channels and H and W are spatial height and width. A normalisation layer performs normalisation of x such that

$$\hat{x} = \frac{x - \mu_{norm}}{\sigma_{norm}}, \tag{3.9}$$

where μ_{norm} and σ_{norm} are the mean and standard deviation computed over different axes depending on the normalisation technique used.

In the case of Batch Normalisation (BN) [175], μ_{norm} and σ_{norm} are computed channel-wise, along the (N, H, W) axes, thus normalising all feature elements that share the same channel across a batch. Layer Normalisation (LN) [176], calculates μ_{norm} and σ_{norm} over the (C, H, W) axes, normalising features for each sample in a batch separately. Instance Normalisation (IN) [177] computes μ_{norm} and σ_{norm} across the (H, W) axes, thus normalising features for each sample and each channel separately. Similarly to Layer Normalisation, Group Normalisation [178] computes μ_{norm} and σ_{norm} over the (H, W) axes, but instead of normalisation over all channels, a specific number of groups of adjacent channels is chosen. Thus, when the number of groups is equal to 1, GN becomes LN, and it reduces to IN when the number of groups is equal to the number of channels. Therefore, the number of groups is a hyperparameter of this layer. In the literature, it is usually chosen to be a factor of 2, and herein groups of 8, 16 and 32 are tested (32 being the maximum

Table 3.4 F_1-scores with different CycleGAN normalisation layers (target stain translated to PAS). The values represent the average of 5 pre-trained segmentation models (S_1^{PAS}, S_2^{PAS}, ..., S_5^{PAS}), each applied to 3 repetitions of the translation model training ($T_{1,n}^{\text{y}\to\text{PAS}}$, $T_{2,n}^{\text{y}\to\text{PAS}}$, $T_{3,n}^{\text{y}\to\text{PAS}}$), therefore the average and standard deviations (in parenthesis) of 15 repetitions in total. The last row represents row-wise averages, excluding Batch normalisation results, since translations obtained by these models are often not plausible.

	Test Staining				
Normalisation Layer	Jones H&E → PAS	CD68 → PAS	Sirius Red → PAS	CD34 → PAS	Average
Instance	0.849 (0.017)	**0.684 (0.043)**	**0.870 (0.009)**	**0.754 (0.008)**	**0.789 (0.087)**
Batch	0.339 (0.059)	0.002 (0.001)	0.508 (0.041)	0.400 (0.067)	0.312 (0.218)
Layer	0.816 (0.014)	0.167 (0.046)	0.832 (0.005)	0.754 (0.024)	0.642 (0.319)
Group_8	0.848 (0.011)	0.308 (0.101)	0.810 (0.006)	0.628 (0.040)	0.649 (0.246)
Group_{16}	**0.849 (0.011)**	0.486 (0.060)	0.800 (0.036)	0.650 (0.039)	0.696 (0.163)
Group_{32}	0.815 (0.007)	0.546 (0.049)	0.807 (0.017)	0.737 (0.015)	0.726 (0.125)
None	0.770 (0.003)	0.250 (0.028)	0.730 (0.035)	0.747 (0.047)	0.624 (0.250)
Average (excl. BN)	0.824 (0.031)	0.407 (0.197)	0.808 (0.046)	0.712 (0.057)	

possible due to the minimal number of filters used in the CycleGAN convolutional layers).

3.5.2 Results

3.5.2.1 Inter-Stain Variability

The translations obtained by many of the stain transfer models are plausible (in the sense of definition given on page 34), as will be discussed in more detail in Subsection 3.5.3.1. Nevertheless, the quantitative analysis performed using pre-trained models shows that there are significant differences in their ability to reduce domain shift. Here, two directions are taken: by evaluating the PAS model's performance on translations from the target stains to PAS (see Table 3.4); and by testing the models pre-trained on each target stain to translations of PAS images (see Table 3.5). The results presented in each table are the averages over three separate CycleGAN models, each applied to five pre-trained baseline models. For ease of reading, the performances of the baseline models are once more given in Table 3.6 to remind that the problem is solvable with high accuracy in all considered stainings.

Since all the results in Table 3.4 are calculated using the same PAS pre-trained models, they can be used to determine the sensitivity of such models to: (column-wise) different types of normalisation (in which the translated stain, and therefore test images, in addition to the pre-trained models are fixed); and (row-wise) different translation models having the same normalisation strategies. As is established in the style-transfer literature, Instance normalisation achieves the best overall performance, although in some cases other normalisation strategies achieve similar performance. For example, with CD34, Instance, Layer, Group_{32} and None (without

Table 3.5 F_1-scores with different CycleGAN normalisation layers (PAS translated to target stains). The values represent the average of 5 pre-trained segmentation models($(S_1^x, S_2^x, \ldots, S_5^x)$, where $x \in \{\text{Jones H\&E}, \text{Sirius Red}, \text{CD68}, \text{CD34}\}$), each applied to 3 repetitions of the translation model training ($T_{1,n}^{\text{PAS}\to x}$, $T_{2,n}^{\text{PAS}\to x}$, $T_{3,n}^{\text{PAS}\to x}$), therefore the average and standard deviations (in parenthesis) of 15 repetitions in total. ↑ indicates improved performance compared to the reverse translation, see Table 3.4, and a ↓ a decrease in performance.

Normalisation Layer	Test Staining				Average
	PAS → Jones H&E	PAS → CD68	PAS → Sirius Red	PAS → CD34	
Instance	**0.891↑ (0.001)**	**0.630↓ (0.019)**	**0.744↓ (0.079)**	**0.641↓ (0.087)**	**0.726↓ (0.121)**
Batch	0.134↓ (0.022)	0.133↑ (0.087)	0.002↓ (0.001)	0.049↓ (0.008)	0.079↓ (0.066)
Layer	0.879↑ (0.002)	0.459↑ (0.111)	0.172↓ (0.080)	0.524↓ (0.106)	0.509↓ (0.291)
Group_8	0.873↑ (0.008)	0.444↑ (0.053)	0.470↓ (0.387)	0.373↓ (0.078)	0.540↓ (0.226)
Group_{16}	0.876↑ (0.002)	0.423↓ (0.121)	0.118↓ (0.025)	0.503↓ (0.106)	0.480↓ (0.312)
Group_{32}	0.883↑ (0.006)	0.577↑ (0.068)	0.320↓ (0.198)	0.377↓ (0.269)	0.539↓ (0.255)
None	0.862↑ (0.009)	0.568↑ (0.078)	0.075↓ (0.055)	0.483↓ (0.115)	0.497↓ (0.325)
Average (excl. BN)	0.877↑ (0.010)	0.517↑ (0.085)	0.316↓ (0.255)	0.483↓ (0.100)	

Table 3.6 F_1-scores for the baseline results (standard deviations are in parentheses).

PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
0.907 (0.009)	0.864 (0.011)	0.853 (0.018)	0.867 (0.016)	0.888 (0.015)	0.876 (0.022)

a normalisation layer) all achieve similar results, whereas in CD68 Instance norm is the clear winner. This indicates that the choice of architecture is dependent on the stain, and most likely, therefore, the complexity of the translation required. However, the fact that none of the pre-trained models applied to CD34 and CD68 translations can achieve baseline results indicates that either the pre-trained PAS models are sensitive to some features not captured by the translation models, and/or the translation models induce a domain-shift.

This can be explained, to some extent, by the difference between histochemical (HC) and immunohistochemical (IHC) stains. Since HC stainings PAS, Jones H&E and Sirius Red use chemicals that interact with several tissue components, multiple normalisation strategies are able to approach baseline performance. On the other hand, IHC stainings CD34 and CD68 are designed to detect specific proteins and here, performance varies greatly.

The results in each column of Table 3.5 are calculated using the same pre-trained segmentation model but now on the target stains. Therefore, each column represents a different model tested on the same PAS data translated to each target stain. As such, they complement the conclusions from Table 3.4, that is from the target stain perspective, by representing the sensitivity of the pre-trained target models to different normalisation strategies. For example, it becomes clear that the normalisation strategy has very little effect when applying the Jones H&E segmentation models

to the PAS translations (except with Batch normalisation). The row-wise results are calculated, again using the same PAS images but now translated to different stains, and therefore different pre-trained models are used. As previously discussed, it seems that in this particular application differences in staining type (e.g. HC vs IHC) can play an important role regarding the sensitivity of the pre-trained model to translations obtained by different stain transfer models.

Comparing the performances between Tables 3.4 and 3.5 represents the two directions of the same translation (PAS $\rightarrow$ Target and Target $\rightarrow$ PAS). Overall, better results are obtained when translating in the Target $\rightarrow$ PAS direction, which could be related to the fact that the translation difficulty is not symmetrical. Even when accounting for the fact that segmentation is more difficult in non-PAS stains (see Table 3.6), more significant drops in performance are observed between Tables 3.4 and 3.5. The differences between performance are indicated by an up or down arrow in each cell of Table 3.5, representing an increase or decrease compared to Table 3.4. When translating from general staining such as PAS to more specific stainings, the translation model must 'invent' stain-specific markers since they are not specifically marked in the general-purpose stain. Thus, this direction of translation can be harder than the other way around and the translation model may fail to reconstruct the finer details that the pre-trained segmentation model relies on. Moreover, these pre-trained segmentation models could be biased toward stain-specific markers (e.g. due to short-cut learning), and thus its performance can be highly dependent on translation quality. Evidence for this is given by the large standard deviations observed when applying the same translation architecture to the same pre-trained segmentation models (e.g. Table 3.5, Sirius Red, Group_8 and Group_{32}). When the translations contain the specific features focused on by the pre-trained models, they perform well (e.g. the best performing translation model in Group_8 achieves an average segmentation score of 0.776), otherwise the translated images can be seen as out-of distribution examples in which the segmentation model fails (the worst translation model in results in Group_8 achieves an average segmentation score of 0.034), even though the translations appear plausible, see Figure 3.13.

Additional evidence for this will be given in Subsection 3.5.3.2 when the segmentation model's variance will be considered from the perspective of stain translation model training.

3.5.2.2 Intra-Stain Variability

In the case of intra-stain variability, the same pre-trained PAS segmentation models used previously are evaluated on the AIDPATH dataset containing PAS-stained WSIs from the Servicio de Salud de Castilla-La Mancha (SESCAM) (see Figure 3.14 for a visual comparison between the two datasets). Direct application of the segmentation models to this variation of PAS is not successful, missing the majority of glomeruli (see the vPAS column in Table 3.7), which confirms the need for a stain normalisation procedure. As in Subsection 3.5.2.1, CycleGAN models were trained to translate the AIDPATH dataset to the source PAS dataset, using different normalisation strategies.

Table 3.7 presents these results, in which it can be observed that the normalisation strategy also has an important role when performing stain normalisation and

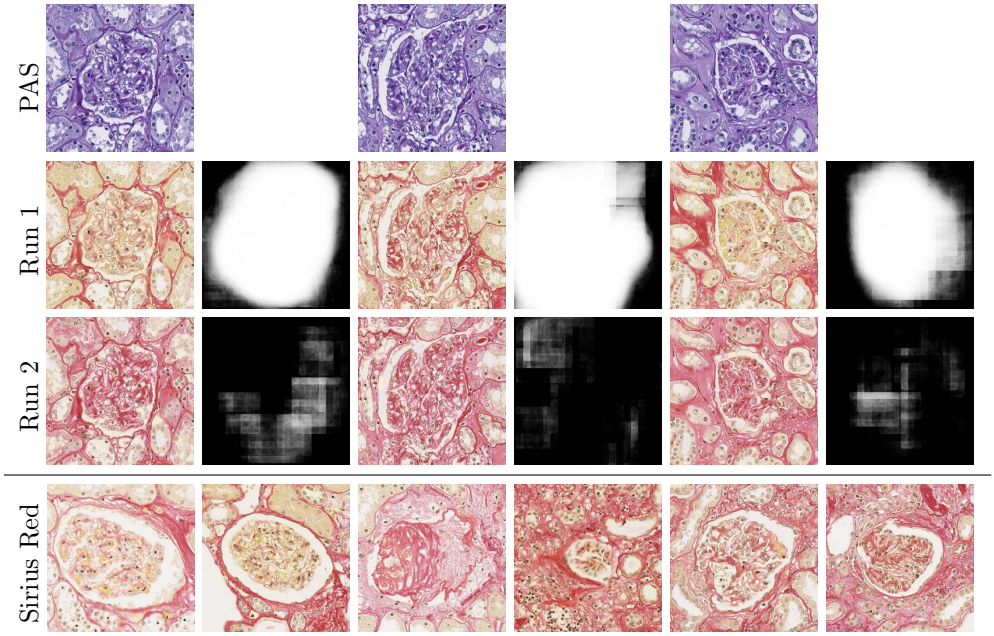


Figure 3.13: PAS patches translated to Sirius Red with two repetitions of the CycleGAN (Group_8) model alongside corresponding segmentations from a pre-trained Sirius Red model. The last row represents real patches from the Sirius Red domain.

segmentation performance does not correlate with visual quality, see Figure 3.15.

The results presented in Table 3.7 should be interpreted with caution. Since the AIDPATH dataset is composed of biopsies, the number of glomeruli in each image is smaller than in the Hanover dataset (private dataset used in this thesis). Thus, a small number of false positives (or negatives) has a big effect on the overall score. Also, the images contain a significant portion of sclerotic glomeruli which is not the case in the Hanover dataset and therefore lower segmentation performance should be expected due to dataset bias. For example, translation models with Batch normalisation obtain the best overall recall, i.e. the lowest rate of false negatives, which means that the segmentation masks predicted by pre-trained models cover the majority of glomeruli. However, its low precision indicates that there are more false positives, i.e. more structures are wrongly classified as glomeruli. Contrarily,

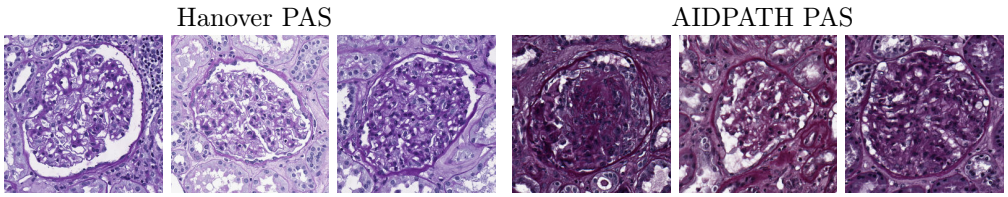


Figure 3.14: Glomeruli PAS variation between Servicio de Salud de Castilla-La Mancha (SESCAM) and Hanover.

Table 3.7 Stain normalisation, the effects of different CycleGAN normalisation layers on the F_1-scores of pre-trained PAS models.

Score	vPAS	Instance	Layer	Batch	$Group_8$	$Group_{16}$	$Group_{32}$
F_1	0.183 (0.091)	0.351 (0.042)	0.504 (0.029)	**0.532 (0.034)**	0.223 (0.053)	0.236 (0.046)	0.282 (0.019)
Precision	0.229 (0.175)	0.819 (0.028)	0.806 (0.024)	0.434 (0.047)	0.680 (0.119)	0.633 (0.105)	0.775 (0.044)
Recall	0.385 (0.256)	0.226 (0.035)	0.370 (0.033)	0.738 (0.012)	0.135 (0.034)	0.148 (0.031)	0.174 (0.015)

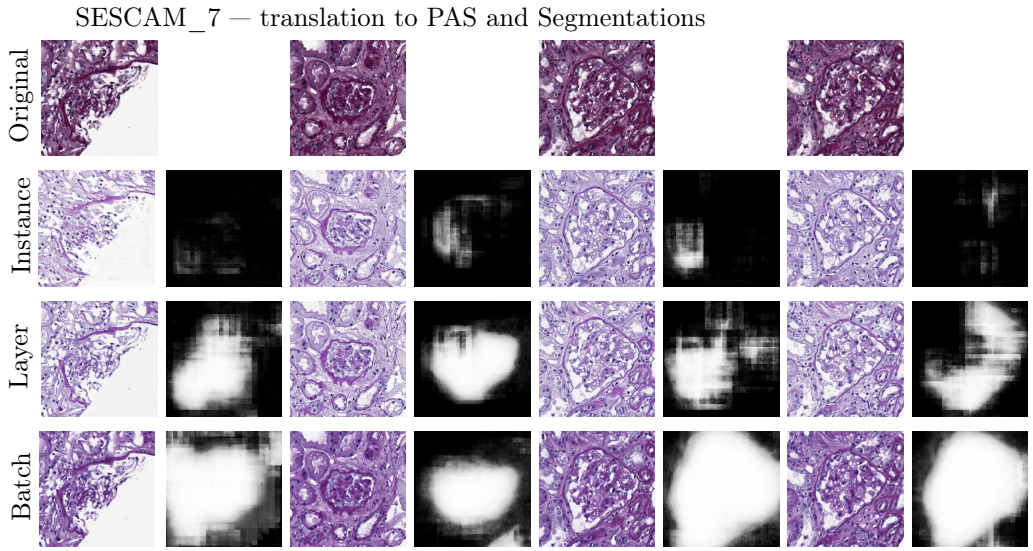


Figure 3.15: Glomeruli patches extracted from SESCAM_7 image (first row) and their translations to Hanover dataset PAS using different CycleGAN models trained on SESCAM_1 and SESCAM_3 images, with corresponding segmentations from a pre-trained segmentation model on Hanover dataset.

Instance normalisation has the best overall precision, meaning that the pre-trained models produce fewer false positives, but the detection is less robust, i.e. not all of the glomeruli structures are detected.

Nevertheless, this study is concerned with performance relative to each normalisation strategy, and since the same pre-trained models are used for these evaluations, the effect of the translation model is evident. Taken together with the results of inter-stain variability (Subsection 3.5.2.1), there is no 'golden' rule for the best choice of normalisation strategy, and it is rather dependent on the problem at hand.

3.5.3 Qualitative and Quantitative Analysis

In this section, qualitative and quantitative assessments of the stain transfer models will be presented. The qualitative analysis includes visual assessment, which is presented in Subsection 3.5.3.1. However, the findings in Subsection 3.5.2 give

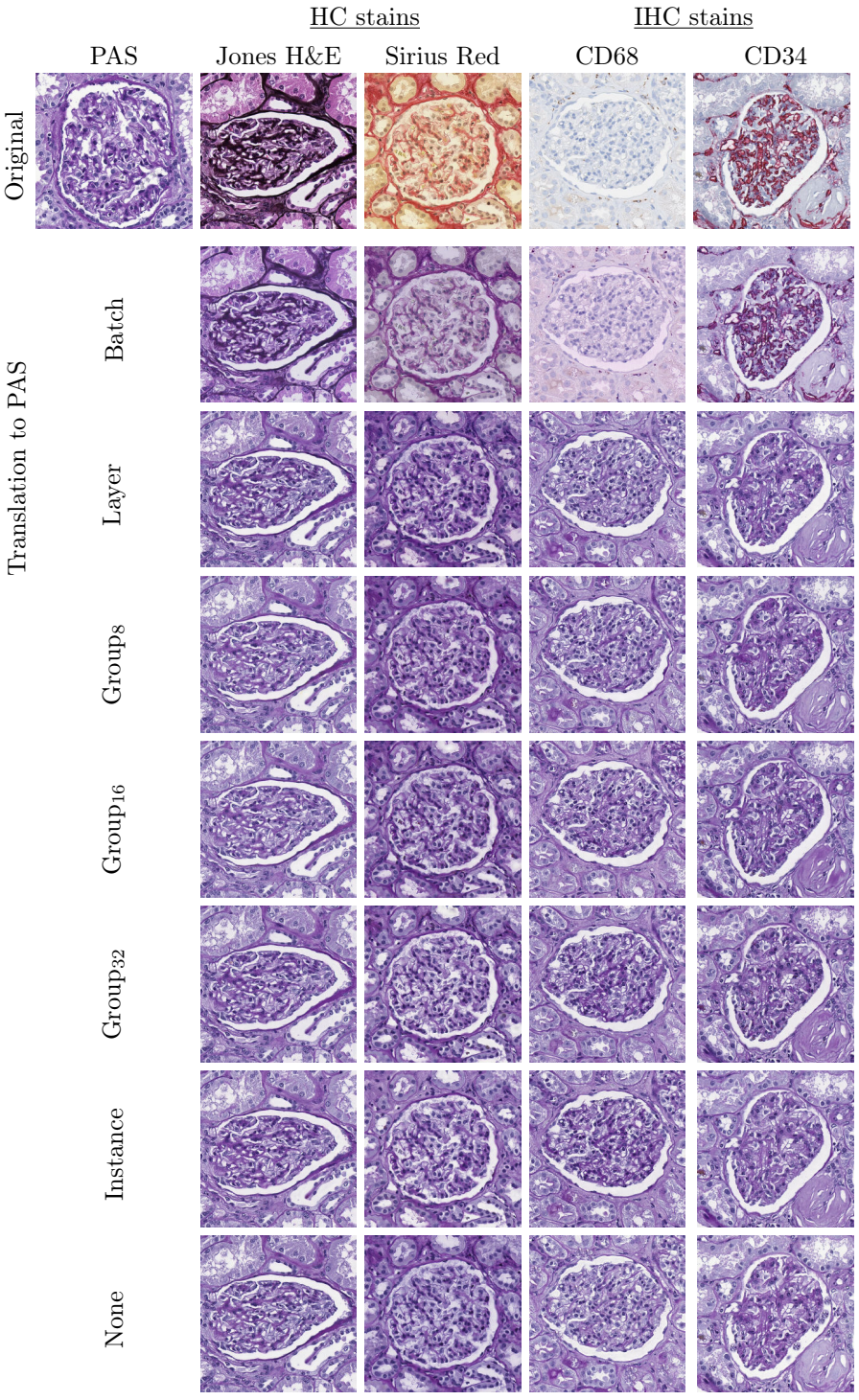


Figure 3.16: Target stain patches translated to PAS using CycleGAN models trained with different normalisation layers.

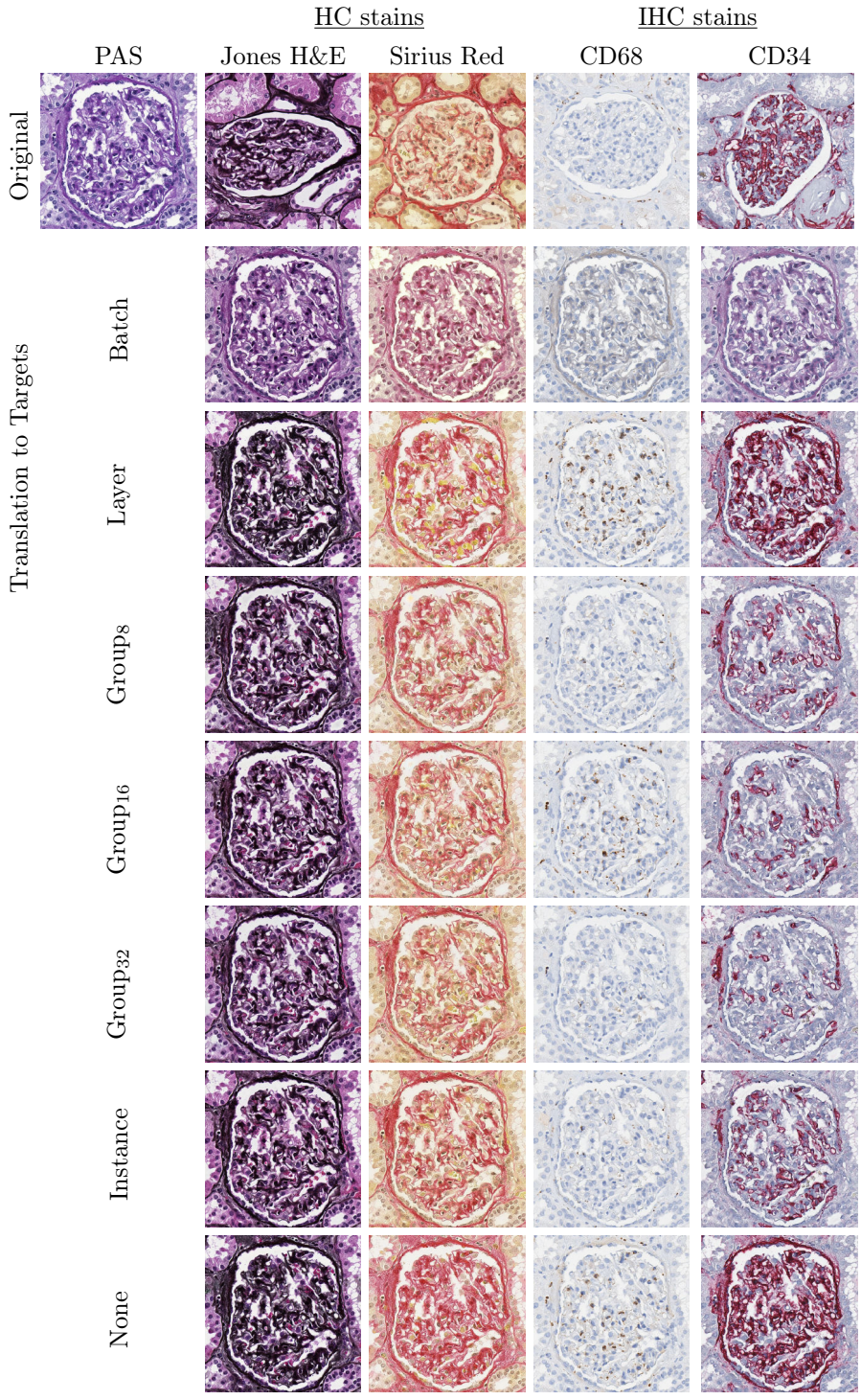


Figure 3.17: PAS patch translated to target stains using CycleGAN models trained with different normalisation layers.

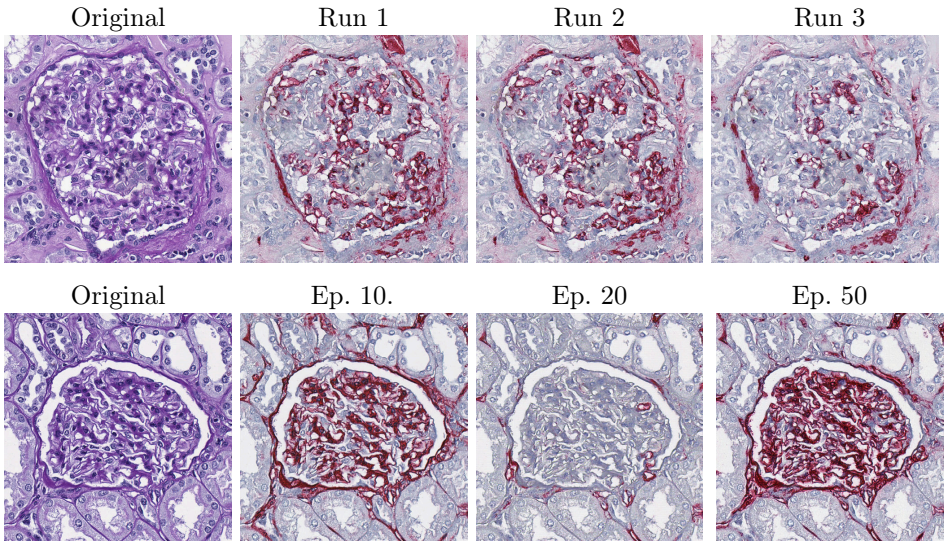


Figure 3.18: An illustration of inter-stain variance, PAS patches translated to the CD34 target stain using (first row) CycleGAN with instance norm from the 50th epochs in three separate training repetitions; (second row) CycleGAN with layer norm from different epochs of the same training run.

strong evidence that this cannot be relied upon. Subsection 3.5.3.2 will further demonstrate this by highlighting the model's instability during different training stages. Moreover, Subsection 3.5.3.3 presents some failure cases that can be easily overlooked by non-experts. The quantitative analysis includes assessment via evaluation approaches found in the literature [150, 155], which are given in Subsection 3.5.3.4, and a comparison of image distributions is presented in Subsection 3.5.3.5. Furthermore, some guidelines about the clinical usage of artificially stained images are presented in Subsection 3.5.4.

3.5.3.1 Visual Quality

Figure 3.16 illustrates the visual quality of the obtained translations, in which each staining has been translated to PAS using different CycleGAN models. Furthermore, Figure 3.17 presents the translations of a PAS patch to each of the target stainings. Visually, all translations (except Batch normalisation) look plausible (in the sense of definition given on page 34).

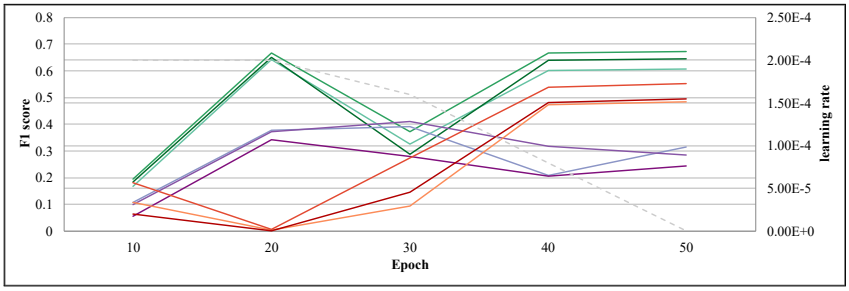
Note that it is not expected that every normalisation type produces the same output as the translation between stains is not a one-to-one mapping. This is more noticeable in stains CD34 or CD68, where translations from PAS can vary greatly in the amount of stain-related markers. Therefore, the general biological aspects of considered stains make a visual comparison between different stain transfer models, such as the one done in [150] (e.g. Figure 6 and 9 of [150]), more unreliable due to technical consideration. Nevertheless, these variations fall within the range of those that can occur naturally. Furthermore, the same variations can be observed for one translation model in different epochs or training repetition, as shown in Figure 3.18,

and thus it can not be strictly related to change in a model architecture. Therefore, drawing general conclusions about model capacity based on visual inspection and the presence/absence of stain-related markers, may lead to incorrect findings.

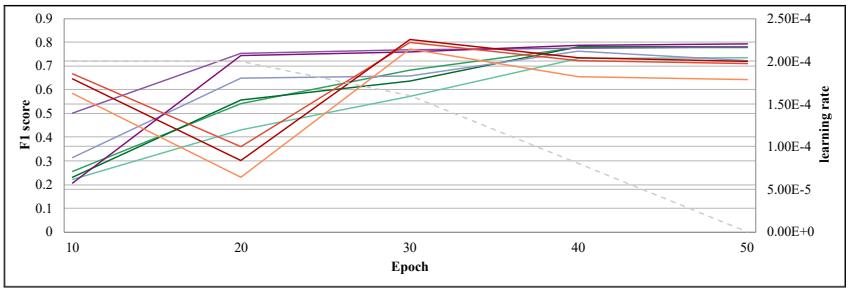
3.5.3.2 Training Stability

As previously noted by several authors, CycleGAN-based stain transfer is able to reach plausible translations early during training [64, 67]. Since there are no explicit stopping criteria in the training process, one can stop training at any moment when no obvious artefacts are produced and the translations are plausible. Taking into account that there is no groundtruth for stain translation (the staining process is irreversible) and that the staining process itself is prone to high variation (particularly between labs), many possible translations are valid. Thus, it is possible that for the same patch, a stain translation model produces different valid translations during training (as shown in Figure 3.18).

In order to investigate how the quality of translation varies during training, the test set (4 WSI images) is evaluated using CycleGAN models from five different epochs—10th, 20th, 30th, 40th and 50th using stains CD34 and CD68, since they are (biologically) the most different to PAS and perform the worst in the previous section (see Tables 3.4 and 3.5). It is assumed that translations between them and PAS are hard.



(a) CD68 → PAS



(b) CD34 → PAS

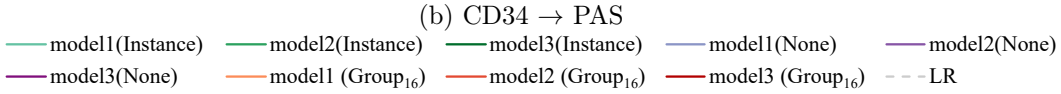


Figure 3.19: (PAS) Segmentation performance in different CycleGAN epochs.

The architectures with Instance normalisation, Group_{16} and without any normalisation (None) are used since they obtained respectively the best, average, and

worst overall scores of the models producing plausible translations in the previous section (i.e. Batch normalisation is excluded). To visualise this effect, three pre-trained PAS models were randomly selected and their segmentation scores are shown in Figure 3.19 over different epochs. Concerning CD34, better performance is generally obtained in a later epoch; however, this is not the case with CD68. In both cases, longer training does not necessarily correlate with better translations. Note that the learning rate (also included in the figures) decreases during training, explaining the stability obtained at later epochs.

Moreover, the ranking of the normalisation strategies is not constant in each epoch; for example in the case of CD34, the translations obtained using Group_{16} in the 30th epoch are better segmented than those obtained by the Instance norm in the final epoch. As such, the results presented in Tables 3.4 and 3.5 may vary depending on the experimental setup (training duration, etc.). Apart from visual differences, an additional cause of the variance of pre-trained model performance could be different levels or types of noise being injected into the translations at different epochs due to self-adversarial attacks to which CycleGAN-based architectures are prone [172]. Additional evidence for this is that all segmentation models are affected similarly at the same epoch (e.g. in the case of CD68 and Group_{16}, all models have almost 0 F_1-score), indicating that the problem originates in the translation, rather than short-cut learning. This also goes inline with the well-known phenomenon of transferability of adversarial examples [179].

To confirm that visual quality is not related to segmentation performance, Figure 3.20 presents translations to PAS at different epochs during training using Instance norm (since it is found to be the best strategy overall), along with their corresponding segmentations (using PAS model 2 from Figure 3.19). As can be seen, they are all plausible; however, the segmentations vary greatly.

3.5.3.3 CycleGAN Failure Cases

In addition to replacing Instance normalisation with other types of normalisation, the normalisation layer was removed entirely from the CycleGAN architecture. Although this modification can sometimes lead to more unstable training (in this case, the translations between PAS and CD68 or Sirius Red were more frequently unstable), the obtained results are still visually appealing and even better than with some normalisation strategies (e.g. Batch normalisation). Nevertheless, in this setting it was found that the CycleGAN models are more likely to produce artefacts. The model is prone to hallucinate features, such as these presented in Figure 3.21. This behaviour was observed particularly often when CD34 and CD68 were the target stains. Since the produced artefacts are visually in accordance with the overall image texture, these cases could be easily unnoticed by the untrained eye, highlighting the importance of including pathologists in the stain translation development process.

3.5.3.4 Reconstruction Assessment

Modifications to CycleGAN architecture can include specific modules or loss functions in order to ensure that the model preserves important structural information, which is sometimes quantified by PSNR and SSIM scores [150]. However, it has been

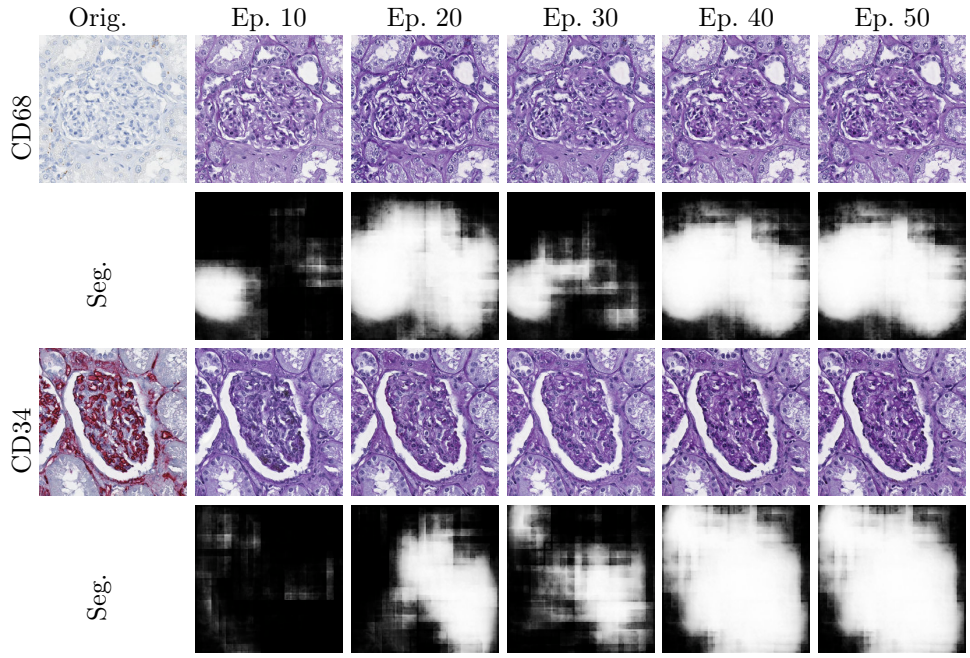


Figure 3.20: Glomeruli patches from CD68 and CD34 stainings translated to PAS using CycleGAN (Instance) models from different training epochs and their segmentation using pre-trained PAS model 2 from Figure 3.19).

shown that changing the normalisation layers of even the most basic CycleGAN architecture can cause differences in the preservation of structural information during translation. Figure 3.22 presents the SSIM and PSNR of PAS images reconstructed via translation to different target stains with different CycleGAN normalisation layers. These are calculated over 200 random patches (100 glomeruli and 100 negative). As it can be observed from these figures, significant variation in both metrics is present in all target stains. More importantly, the order of the metrics does not correlate with the ability of the architecture to reduce domain shift (see Table 3.4 and Table 3.5). This indicates that using these metrics in this setting may not accurately reflect the benefits of modifications to CycleGAN-based models.

3.5.3.5 Translation Distributions[5]

Despite the success of CycleGANs, they are prone to self-adversarial attacks [62, 172, 173]. The cycle-consistency constraint forces the generator to hide information necessary to reconstruct the input image as imperceptible noise and since it has been shown that the results appear plausible (in the sense of definition given on page 34), one possible hypothesis is that this imperceptible noise causes the domain shift observed in Subsection 3.5.2 [170]. Song et al. [180] show that a PixelCNN++ generative model can be used to detect adversarial attacks in images and it is therefore used here to detect the presence (or not) of adversarial noise in the

[5]This analysis was done in collaboration with Zeeshan Nisar (PhD student, SDC research team, ICube).

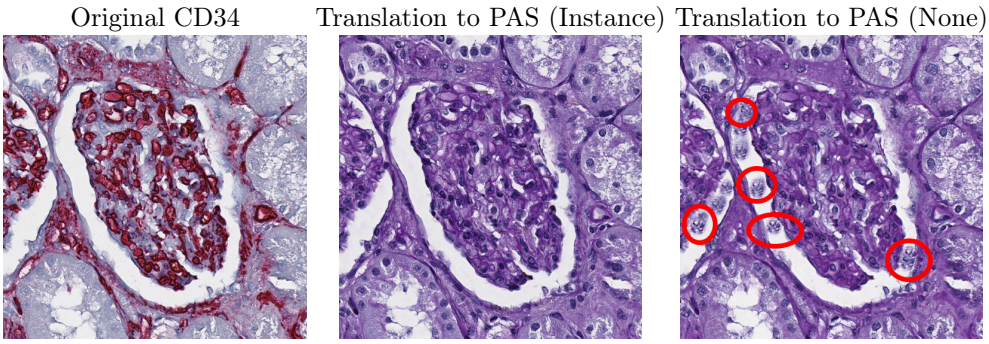


Figure 3.21: Hallucination effect of CycleGAN without normalisation.

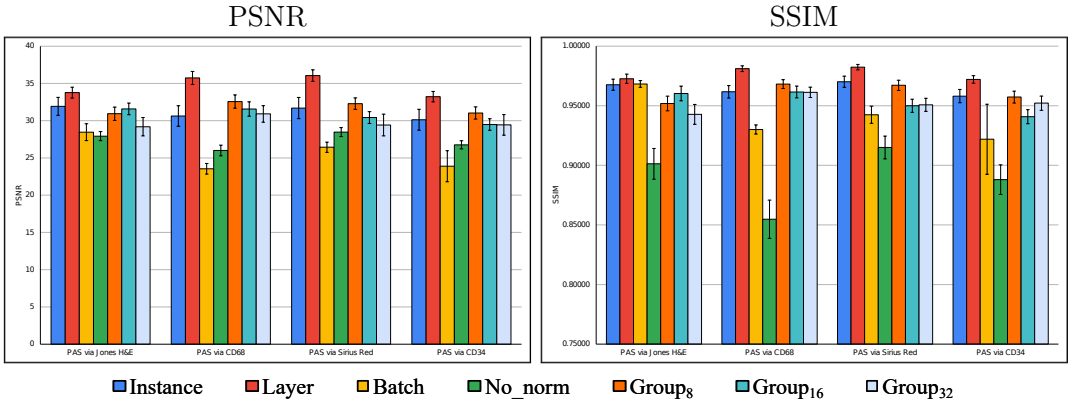


Figure 3.22: PSNR/SSIM scores of reconstructed PAS images using different CycleGAN models.

obtained translations.

PixelCNN++ [181] quantifies the pixels of an image x over all its sub-pixels as a product of conditional distributions, such that it learns to predict the next pixel value given all previously generated pixels, that is

$$p(x) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1}). \tag{3.10}$$

These conditional distributions are parameterised by a convolutional neural network (CNN) and hence shared across all pixel positions in the image. The PixelCNN++ [181] architecture is used to model the underlying distribution of each stain separately (training details are presented in Appendix B.6): PAS, Jones H&E, CD68, Sirius Red, and CD34. As such, the PixelCNN++ models are able to generate images that belong to the real data distribution. Figure 3.23 presents examples of several such patches. Due to memory limitations, the models are trained on 32×32 pixel patches (therefore, each 512×512 pixel patch is decomposed into non-overlapping patches), and therefore the models are able to generate only structures visible at this patch size. Visual evaluation can clearly identify cell nuclei, endothe-

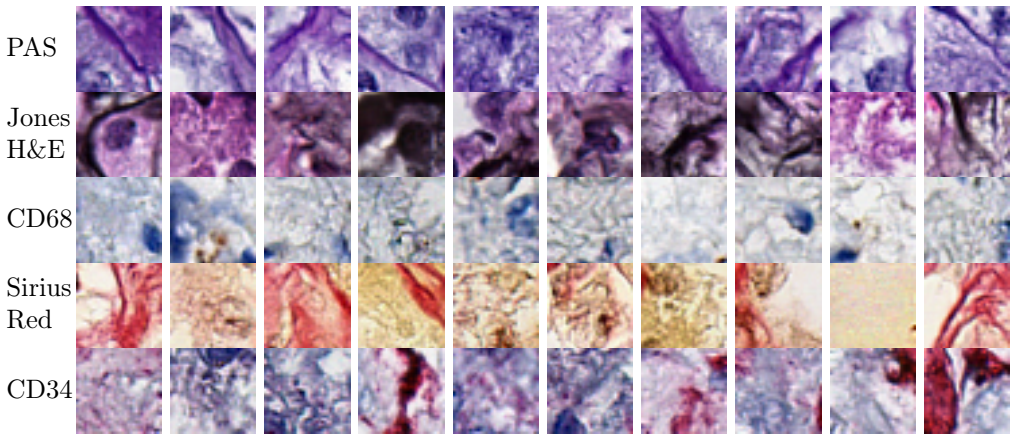


Figure 3.23: Samples generated from the trained PixelCNN++ for each stain (PAS, Jones H&E, CD68, Sirius Red, and CD34).

lial lining, a partially granular cytoplasmic texture, extracellular matrix components (such as collagen fibres), and even some cell borders are faintly outlined, recapitulating the cell membranes of some epithelial cells.

To further validate the efficacy of the PixelCNN++ models, the distributions of the training, validation, and test sets are plotted for all stains, see Figure 3.24. This confirms that the PixelCNN++ model is able to accurately estimate real data distributions, since there is an overlap between all three distributions in all stains. To investigate whether the drop in performance of the pre-trained models is caused by an imperceptible domain shift, all the test target stains are translated to PAS using the CycleGANs models with different normalisation layers. Figure 3.25 shows the distributions of the resulting images compared to the real PAS test set. It can be observed that the translated target-to-source stains have a different data distribution, confirming the existence of a domain shift, which causes the pre-trained models to fail. If the translation is performed in the opposite direction (from PAS to target), the same domain shift is found, see Figure 3.26. It is important to note when interpreting these figures, that the relative distance between the graphs of real and translated distributions does not necessarily correlate to the performance of pre-trained models [180].

These results confirm that, although plausible, the translations obtained with various stain translation models, actually generate data in a manner that slightly mismatches the real data distributions. Thus, the pre-trained models can exhibit variation in performance when applied to such data even though the output is visually plausible. This additionally confirms that stain transfer (Hypothesis 2.1 and 2.2) is the cause of segmentation's performance variability.

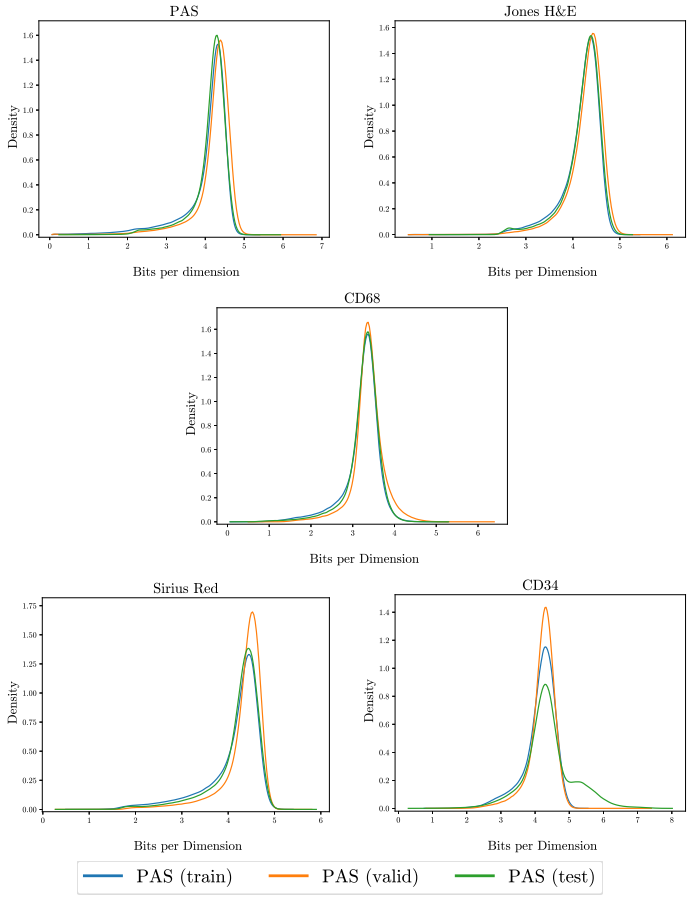


Figure 3.24: Visualisation of training, validation, and test data distributions for each stain under PixelCNN++.

3.5.4 Stain Transfer for Clinical Application[6]

Stain translations can be useful and hold great potential for the future development of digital histopathology. The potential risk that their results may be misleading under certain circumstances and can be mitigated by carefully considering the biological and image-related context and the intended use case.

It is significant to note that the translation process can greatly affect the appearance of stain-specific markers in immunohistochemical stains, such as CD68 and CD34. In the given examples, brown immunohistochemical staining (CD68) reflects the expression of a specific protein during macrophage differentiation and activation, whereas gradual enrichment of purple staining as a result of the chemical PAS reaction reflects the presence of carbohydrate macromolecules that are not specific for macrophages, but enriched in their phagocytic subset and is associated with protein degradation. Thus, both methods highlight slightly different populations of macrophages, illustrating the important caveat of translating histochemical (HC) to immunohistochemical (IHC) and vice versa: the translation looks "plausible" (see

[6]This analysis was done in collaboration with Prof. Dr. Friedrich Feuerhake (Institute of Pathology, Hannover Medical School, Germany; University Clinic, Freiburg, Germany).

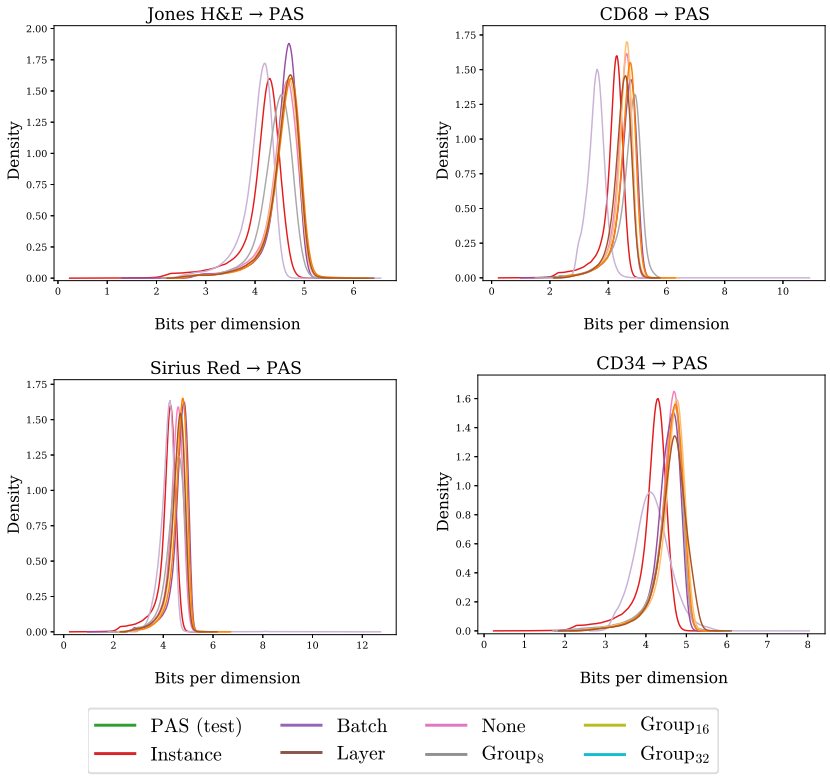


Figure 3.25: Qualitative comparison of the real PAS (test) distribution and translated target-to-PAS distributions using different CycleGAN normalisation layers. Each distribution is calculated using the test set for each stain.

definition on page 34). In this context this means that their visual appearance is consistent with the target staining and reflects regular morphological features of macrophages such as size and shape. However, this visual plausibility may not be accurate for biomedical evaluation. Examples include biological features such as "macrophage activation" (reflected by expression of the CD68 protein on the cell surface), or "protein digestion by macrophages" (reflected by PAS-positive granular substance within the macrophage cytoplasm). The same holds true for translations between CD34 and PAS: both methods can highlight blood vessels, CD34 (IHC) specifically the inner layer (endothelium) and PAS (HC) less specifically components of the entire vessel wall. Again, translations between stains in any direction are likely to look "plausible" and may even be useful for general visual detection of blood vessels, but they are clearly misleading for other purposes (e.g. specific evaluation of endothelial pathology). For example, the general detection of blood vessels (e.g. their quantification) is feasible for larger vessels and even down to the size of arterioles in PAS → CD34 translations and vice versa. However, the evaluation of microvessel density including very small vessels (capillaries) would be difficult, as they are strongly labelled by CD34 but not necessarily by PAS. Likewise, there are specifically PAS-positive structures, e.g. so-called granular osmiophilic material ("GOM", a diagnostic hallmark of a vascular disease abbreviated "CADASIL" [182, 183]) that

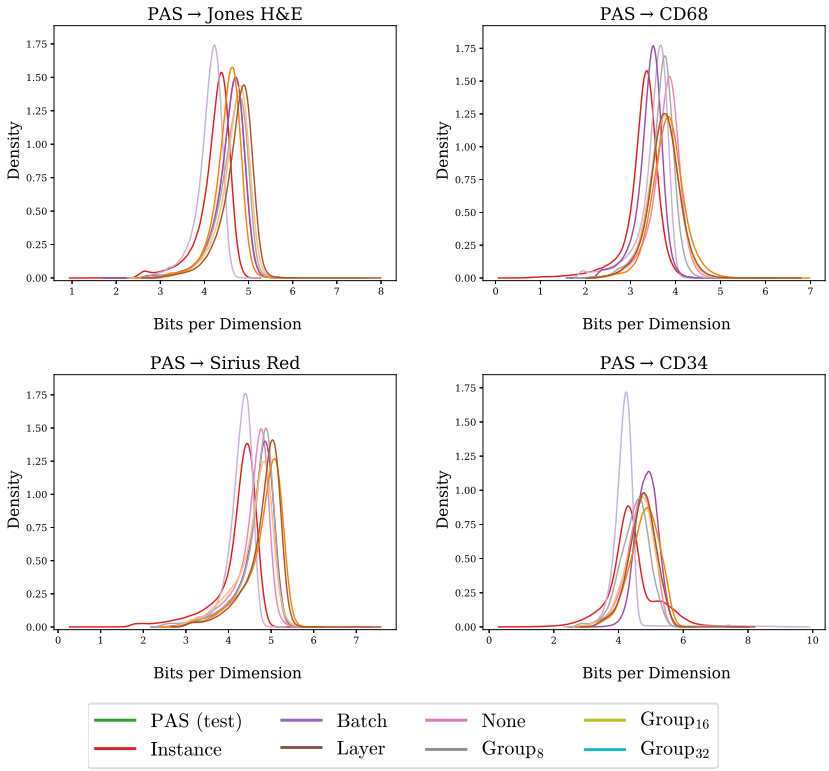


Figure 3.26: Qualitative comparison of real target stain distributions (test set) and translated PAS-to-target stain distribution (test set) for each type of CycleGAN normalisation layer.

would not be visible in CD34 staining. Translations between stains would be misleading in these cases. Other examples include the detection of glomeruli in kidney tissue. This would be possible in both stainings and in translations thereof if the aim is solely glomeruli quantification, but misleading if particular substructures are evaluated for diagnostic reasons.

Overall, the purpose of this study is to demonstrate the sensitivity of the stain transfer model to small modifications in architecture. However, as previously mentioned, the obtained translations represent artificially generated images and, in the current state, cannot replace real images for diagnosis purposes. The aim of such analysis is to give evidence that a strong comparison between stain transfer models cannot be easily determined, particularly solely based on visual inspection, since the overall conclusion depends on various hardly-controllable factors (such as training run, epoch, architecture, etc.).

3.6 Benefits to Supervised Methods

The findings presented in Section 3.4 and Section 3.5 give evidence that CycleGAN-base translations contain imperceptible noise that can be associated with stain characteristics. In this section it will be further demonstrated that such noise can be per-

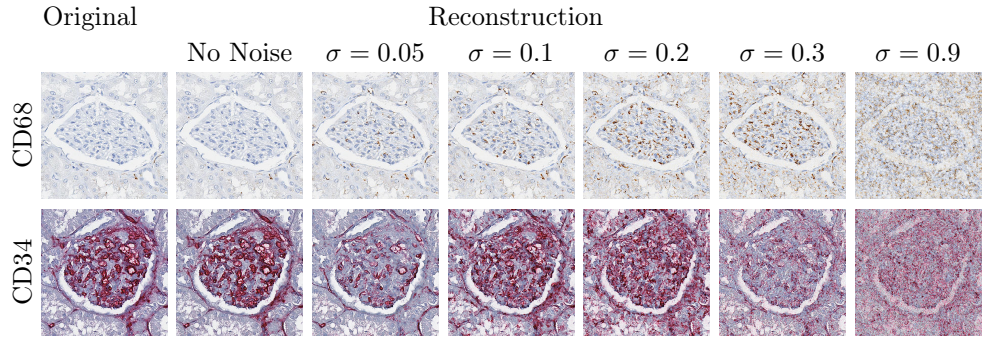


Figure 3.27: Generating variation by adding noise to intermediate PAS representations, the images are reconstructions of CD68/CD34 $\rightarrow$ PAS + $\mathcal{N}(0, \sigma)$ $\rightarrow$ CD68/CD34.

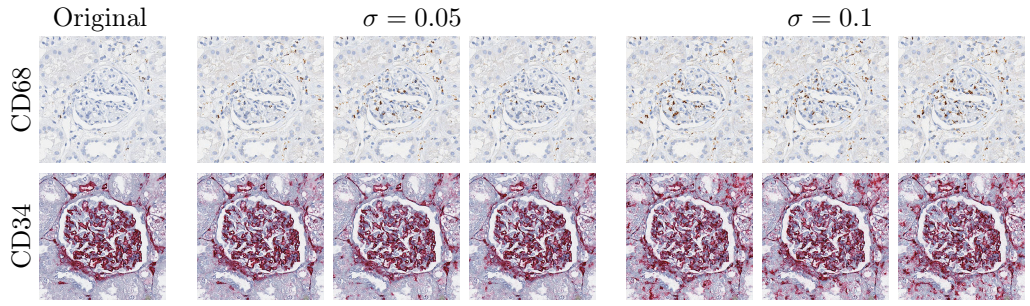


Figure 3.28: Effects of additive Gaussian noise with the same standard deviation, the images are reconstructions of CD68/CD34 $\rightarrow$ PAS + $\mathcal{N}(0, \sigma)$ $\rightarrow$ CD68/CD34.

turbed in a way to generate artificially created, diverse histopathological examples in given staining. Such ability is used to propose a new augmentation technique for supervised training in digital histopathology which positively affects a deep learning model's robustness.

3.6.1 Self-Adversarial Attack as an Augmentation Strategy

Given a CycleGAN translation between a general and a specific stain (e.g. immunohistochemical), the translation in the specific to general direction should encode stain-specific information, particularly that related to the position of stain-specific markers which are not visible in the general stain. If this is the case, then by perturbing the hidden noise, one can affect the appearance of stain-related markers in the reconstructed image. By doing this a new augmentation technique can be introduced that increases the variability of stain-specific markers in histopathological data. The aim of such augmentation would be to increase a model's robustness when trained for non-stain-related tasks.

A mapping between PAS (not specific staining) and two immunohistochemical (specific) stainings CD68 and CD34 is considered to explore this hypothesis. To remind, CD68 marks a protein exclusively produced by macrophages, and CD34 stains a protein specific to the endothelial cells of blood vessels. PAS, as a chem-

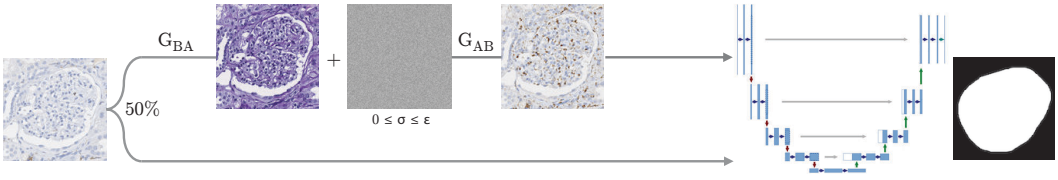


Figure 3.29: Proposed self-adversarial attack-based augmentation approach.

ical reaction staining glycolysated proteins in general, can highlight some parts of macrophages (co-located but not overlapping with CD68), the basal lamina of blood vessels (co-located with CD34), and other structures not highlighted by either CD68 nor CD34 that contain glycolysated proteins.

3.6.1.1 Method

To translate between PAS and immunohistochemical, separate CycleGAN models are trained. As discussed in Section 3.4, when translating between immunohistochemical and histochemical stains, imperceptible noise is present in the intermediate translation and this contains information about stain-related markers, in the direction from a specific stain to general. Thus, changing the encoded noise changes the reconstruction of stain-related markers. This noise can be perturbed by introducing additive zero-mean Gaussian noise to the intermediate translation [172]. The amount of stain-related characteristics can be controlled through the Gaussian's standard deviation. Figure 3.27 shows that translation output (i.e. reconstructed input, B_{rec}) variance is directly proportional to the level of additive noise and Fig. 3.28 shows that different translations result from varying noise of the same standard deviation.

The physical accuracy of the resulting stain-related markers remains an open question, but the fact that they are positioned in plausible locations opens the possibility of exploiting them to reduce a model's sensitivity to such stain-related markers. It should be noted that the amount of additive noise is stain-dependent: a standard deviation, σ, of 0.3 produces realistic CD68, but a noisy CD34, output. The proposed augmentation process is described in Fig. 3.29. Let denote PAS as A and an immunohistochemical stain as B. During supervised training of a model on B (e.g. for glomeruli segmentation), each sample b_i is first translated to PAS, A', using the trained CycleGAN generator G_{BA}, with a probability of 50%. Next, zero-mean Gaussian noise with standard deviation σ is added to the intermediate translation, which is translated back to B using G_{AB}, where $\sigma \in (0, \epsilon_{\text{stain}}]$ with uniform probability. The value ϵ_{stain} is determined for each staining separately. As such, the input is altered by the arbitrary appearance of stain-related markers and the supervised model is forced to be less sensitive to their appearance.

As the translation process likely hides non-overlapping inter-stain information, the intermediate stain potentially determines which information is encoded. As with all augmentation techniques, a parameter value must be chosen; in this case it is the noise level ϵ_{stain}. Since the problem being addressed is supervised, ϵ_{stain} can be optimised experimentally; however, it could be chosen by manually validating the reconstructions. A grid search was conducted on a separate dataset partition containing a

Table 3.8 Quantitative results, standard deviations are in parentheses, number of glomeruli training patches follow the data percentages.

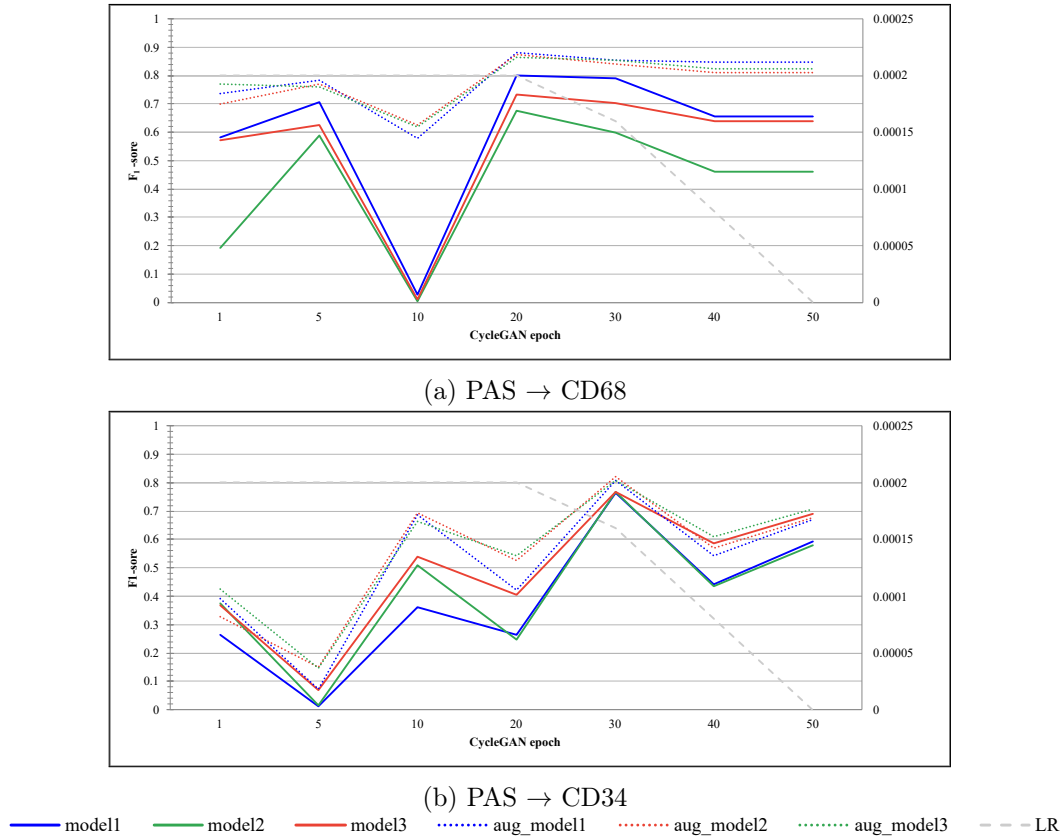
Stain		Baseline			Noise Augmented		
		F_1	Precision	Recall	F_1	Precision	Recall
CD68	10% (53)	0.739 (0.018)	0.754 (0.047)	**0.728** (0.034)	**0.767** (0.036)	**0.832** (0.053)	0.713 (0.044)
	30% (159)	0.812 (0.026)	0.839 (0.038)	0.788 (0.038)	**0.828** (0.026)	**0.848** (0.065)	**0.812** (0.017)
	60% (317)	0.831 (0.024)	0.812 (0.037)	**0.852** (0.014)	**0.856** (0.017)	**0.888** (0.026)	0.826 (0.021)
	100% (529)	0.853 (0.018)	0.849 (0.024)	**0.858** (0.020)	**0.878** (0.007)	**0.899** (0.023)	**0.858** (0.010)
CD34	10% (57)	0.837 (0.017)	0.770 (0.033)	**0.919** (0.009)	**0.839** (0.035)	**0.778** (0.061)	0.913 (0.008)
	30% (170)	0.877 (0.012)	0.841 (0.030)	**0.917** (0.012)	**0.890** (0.011)	**0.867** (0.023)	0.916 (0.009)
	60% (341)	0.882 (0.008)	0.840 (0.015)	**0.927** (0.005)	**0.901** (0.007)	**0.884** (0.019)	0.919 (0.010)
	100% (568)	0.888 (0.015)	0.849 (0.033)	**0.931** (0.010)	**0.903** (0.006)	**0.888** (0.014)	0.919 (0.009)

random 10% subset of each class. The range $\epsilon_{\text{stain}} \in [0.01, 0.05, 0.1, 0.3, 0.5, 0.9]$ was tested by averaging 3 repetitions. It was found that adding noise in the range that produces realistic output improves upon the baseline ($\epsilon_{\text{CD68}} \leq 0.3$ and $\epsilon_{\text{CD34}} \leq 0.1$), confirming that the parameter can be chosen manually as well. Nevertheless, the best value should be determined for each stain to maximise F_1 score and these were found to be $\epsilon_{\text{CD68}} = 0.05$ and $\epsilon_{\text{CD34}} = 0.01$.

To evaluate the augmentation's effect with a few data samples, each training set is split into 5 folds containing 10%, 30%, and 60% of each class taken at random. A separate random 10% subset of the training data is extracted to choose ϵ_{stain}. All models are trained for 250 epochs, the best performing model on the validation partition is kept, and tested on the 4 held-out test patients. The average F_1-score and standard deviation are reported.

3.6.1.2 Results

Table 3.8 presents the baseline and noise augmented results with varying amounts of data. The proposed augmentation improves F_1 scores unanimously due to increased precision. The recall does not improve since no new task-specific information is added, e.g. glomeruli shape or positional variance. Since stain-related markers are not indicative of glomeruli in general, the model should largely ignore them. However, fibrotic and sclerotic glomeruli are present, to which the model can wrongly associate a specific pattern or marker. For example, fibrotic changes are associated with CD68 positive macrophages [184] and a loss of CD34 positive vascular structures. Overemphasising immunohistochemical variations via augmentation biases the model to other properties, decreasing recall but disproportionately increasing precision.



(a) PAS → CD68

(b) PAS → CD34

Figure 3.30: Comparison of segmentation performance of pre-trained models trained with and without the proposed augmentation strategy on the whole PAS test set over different CycleGAN epochs (the CycleGAN model is different from that used for proposed augmentation).

3.6.1.3 Effects on a Model Robustness

As previously discussed in Section 3.5, models trained in a supervised way on a particular stain obtain fluctuating performance when applied to images from other stains translated using CycleGAN models in different epochs of CycleGAN training. The potential cause is noise encoded into the translations. However, pre-trained models trained with the proposed augmentation strategy should be more robust to such translation variability. Thus, the models obtained using the full training set (i.e. 100% of data) are applied to PAS images translated to CD68/CD34 using different CycleGAN epochs (and the model is different from that used for augmentation). The obtained results are presented in Figure 3.30. In the case of CD68, the augmentation strategy improves robustness dramatically, particularly in cases where models trained without this augmentation completely failed to recognise glomeruli (epoch 10). Overall, taking the last CycleGAN epochs for translation, augmented models obtain on average 24.1% of the increase in overall F_1-score. In the case of CD34, the improvements are not as significant, although the augmentation strategy does bring an average increase of 6.4%.

3.7 Conclusions

This chapter introduced two GAN-based methods for plausible stain transfer — CycleGAN based, to perform pair-wise stain transfer and StarGAN based, to enable multi-domain stain translation. The work related to the application of CycleGAN for stain transfer has been done in parallel with other authors [19, 20, 94], and now represents the standard approach for virtual staining. These results are partially presented in a journal article Vasiljević et al. [64]. However, concurrent works did not provide an in-depth analysis regarding the quality of the obtained translation, which has been addressed in this chapter.

Specifically, the study demonstrates that stain transfer using the most commonly used technique, CycleGAN, to reduce the domain shift introduced by both inter- and intra- stain variation, is highly sensitive to training settings such as the number of epochs or simple architectural modifications such as the normalisation layer. Since CycleGAN-based methods are widely adopted in the literature and different architectural modifications are introduced aiming for better translations, the experiments in this chapter compared different CycleGAN models. In order to control the architectural differences between stain translation models, the experiments focused on different normalisation layers in the CycleGAN architecture.

Surprisingly, the majority of architectures tested lead to visually plausible translations. However, by extensive experiments it was shown that these models generate data that belongs to different distributions, leading to unpredictable performance when using pre-trained segmentation models. Thus, it can be concluded that visual inspection is not sufficient in all situations and should be complemented by additional criteria for comparing and choosing stain transfer models. Specifically, this chapter shows that in both stain transfer and stain normalisation, pre-trained models exhibit huge performance fluctuations even when there are no obvious visual differences between the translations. These phenomena are attributed to the self-adversarial attack, a consequence of the natural many-to-many mapping that exists between different stains (or even the same stain in different labs) which is reduced to a deterministic mapping in CycleGAN-based architectures. This is confirmed by showing that there is a difference between the distributions of real and translated images using PixelCNN++ generative models. The findings also give strong evidence that the architectural choice affects the appearance of important diagnostic evaluation criteria (such as markers for macrophages) and thus artificially generated images using these models cannot be relied upon for diagnostic purposes. These findings are summarised in a journal publication, Vasiljević et al. [63].

Moreover, this chapter confirms that the deterministic nature of CycleGAN models forces noise to be injected into the translations. In the case of translations between histochemical and immunohistochemical stainings, it is shown that the noise encodes stain-specific features and these can be perturbed in a way to alter their appearance in a plausible manner. This is used to propose a new augmentation method for supervised training that increases the robustness of deep learning models. The method was presented as an oral presentation at IEEE International Symposium on Biomedical Imaging (ISBI) 2021, Vasiljević et al. [62].

The presented findings raise awareness about the clinical usage of stain transfer, which has been rarely addressed in the literature. Of particular concern is the fact

that the presented stain transfer approaches can affect the appearance of diagnostic evaluation criteria and thus such artificially generated images cannot be relied upon for diagnostic purposes. Moreover, different runs of the same model can result in diagnostically different translations, which can lead to misleading clinical conclusions. All of this indicates that experts are needed in the loop, both for the development and evaluation stages, as non-competent validation can easily neglect some of these limitations.

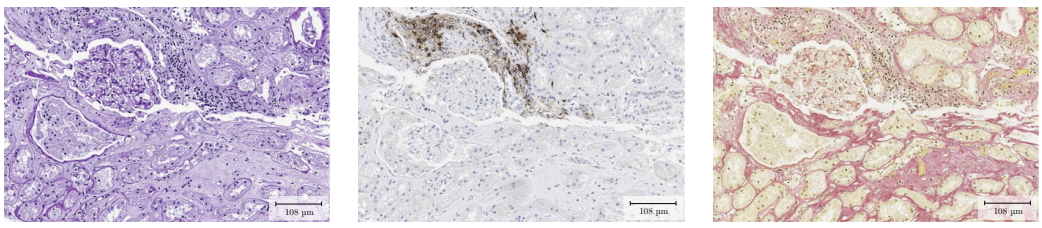
The work presented in this chapter indicates that stain transfer itself, at the given stage of development, might not be sufficient to deal with domain shift in digital histopathology. However, the large fidelity of the obtained transfer can be exploited to build more robust deep models, as confirmed in the proposed augmentation method [62]. Further investigation regarding building more robust models is presented in Chapter 4 and Chapter 5 of this thesis.

Chapter 4

Stain Invariance

A typical procedure in the histopathological analysis is the examination of consecutive tissue slices stained differently (see Figure 4.1). Each stain provides specific information about the underlying tissue, enabling pathologists to inspect various aspects of a specific organ, structure or pattern in the tissue. For example, to diagnose kidney allograft rejection, it is necessary to study the inflammatory microenvironment of the kidney, such as the distribution of immune cells (e.g. macrophages) in relation to the number of glomeruli and their health status [39].

To enable the automatic integration of information from differently stained images, the structure of interest, e.g. glomeruli, should be detected in each tissue slice, regardless of the staining. Assuming that annotations exist for each stain, such detection would be a typical application of deep learning. However, as discussed in Chapter 3, developing stain invariant solutions is not straightforward due to the scarcity of annotations in digital histopathology datasets and intra/inter stain-variation. Usually, for a particular task, only a limited amount of annotated data is available, e.g. for glomeruli segmentation, annotations exist for one stain obtained from one laboratory. Thus, the main obstacle in developing stain invariant [7] solu-



(a) Periodic Acid Schiff (PAS) (b) CD3 (c) Sirius Red

Figure 4.1: An example of three consecutive WSIs of a kidney nephrectomy sample with three common stains. Each staining provides different information on the tissue: general structural information in PAS, distribution of T lymphocytes in CD3, specific structures such as collagen or muscular fibres in Sirius Red.

[7] The term stain invariance, as taken in this thesis, refers to the ability of a solution to generalise across different stains. It should be noted that in the literature, this term is also used for solutions

tion is learning from a limited set of annotated data, which frequently contains just one staining.

Considering the problem of stain invariance from a domain adaptation perspective, given a source domain (annotated stain) and several target domains (unannotated stains), the goal is to obtain a model that is able to generalise across all domains. That can be achieved either by directly obtaining an invariant model or by adapting the already trained model to a given target stain in an unsupervised manner. In Chapter 3, it has been demonstrated that adaptation of already trained models by only pixel-space alignment can have limited success in reducing inter-stain domain shift. This chapter proposes two solutions to obtain more stain robust models. In Section 4.1 an augmentation strategy based on stain transfer is proposed, which yields a test-time stain invariant model able to generalise across several stains, including unseen stains. In Section 4.2 a solution which exploits stain transfer to enhance feature-space-based adaptation of already trained models is proposed.

As in the rest of this thesis, the proposed solutions aim to segment glomeruli, a highly relevant functional kidney unit, considering a dataset of five different stainings (a detailed description of the dataset is given in Chapter 1, Subsection 1.3.2). The PAS staining is taken as a source, annotated, domain and four other stainings as targets — two histochemical (Sirius Red and Jones H&E) and two immunohistochemical stainings (CD34 and CD68)— which are considered to be unannotated.

4.1 Unsupervised Domain Augmentation

Recalling the approaches that use virtual staining to tackle the domain shift problem in digital histopathology presented in Chapter 2 of this thesis, herein, the proposed method belongs to the augmentation class of approaches — increasing the variability of the input to a deep learning model, implicitly forcing more general features to be extracted. High fidelity of GAN-based unpaired image-to-image translation methods is used to synthesise plausible samples from different stainings that are used as an augmentation for annotated staining.

4.1.1 UDA-GAN

Unsupervised Domain Augmentation using Generative Adversarial Networks (UDA-GAN) is a general approach for training stain invariant Convolutional Neural Networks (CNNs) for a specific task. After training, the model is able to perform a given task in various stains, potentially unknown during training time. It is assumed that annotated WSIs are available for a stain A while WSIs of other stains $B_1, B_2, \dots B_N$ are unannotated. The aim is to increase the variability of the (annotated) training set through augmentation by randomly translating it to the unannotated domains (including the original, annotated domain). The overall architecture of the proposed method is presented in Figure 4.2. The training details are given in Appendix C.1.

The method contains two phases:

a) (Unsupervised) Stain Translation — in order to obtain realistic translations of the annotated stain A to unannotated stains $B_1, B_2, \dots, B_N$, a GAN-based

that are robust to variations in one particular stain (herein referred to as intra-stain variation).

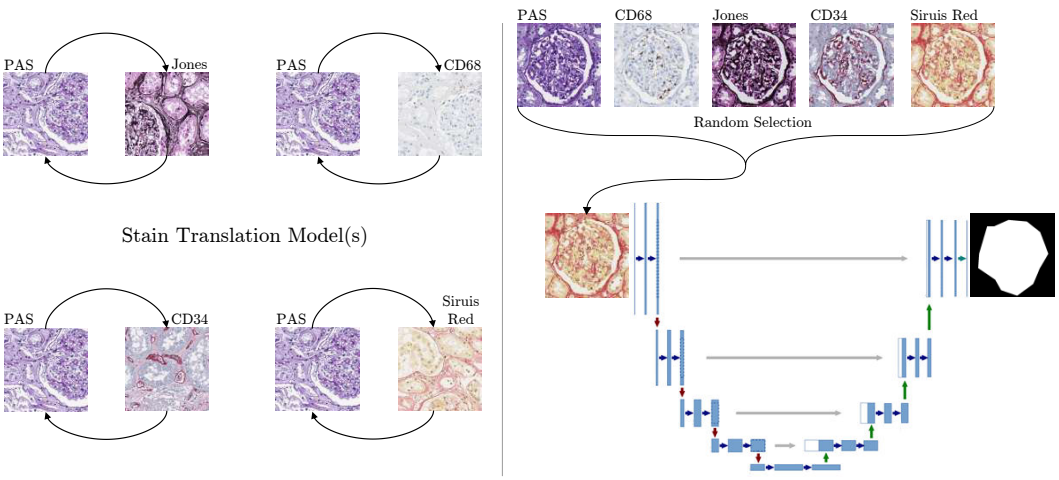


Figure 4.2: Overall diagram of the proposed approach. Phase 1, translation models are learnt to translate images from the source domain to the target domains; Phase 2, patches of the source domain are randomly translated to the target domains during training (U-Net image taken from [13]).

unsupervised, unpaired image-to-image translation model is employed. The translation model needs to guarantee that important structures are not changed during the translation process. The structures that need to be preserved are task-depended. In this study, the task of glomeruli segmentation is considered, and thus, their position, shape, size and orientation should not be changed in the translation process. Based on the results and analysis presented in Chapter 3, both CycleGAN [12] and StarGAN [14] methods ensure their preservation, and thus they are considered herein.

b) (Supervised) Task-related model (Segmentation Model) — this model is trained on the annotated data after being translated to a random unannotated stain. Since translation does not change the overall structure of the image (see Chapter 3), the original domain's label/groundtruth is still valid. Thus, various annotated samples of all available stainings are presented to the model during training, forcing it to learn stain invariant features. Once the segmentation model has been trained, it can be directly applied to the unannotated stains without any further translations.

4.1.2 Quantitative Results

The experimental results are presented in Table 4.1. The UDA-GAN is compared to the solutions proposed by Gadermayr et al. [20]: 1) train a segmentation model on source data and apply it to target data translated to the source domain, referred to as MultiDomain Supervised 1 (MDS1); 2) train on the source stain translated to the target, and apply the model directly on the target images, MDS2 [20, 67]. For MDS1 and MDS2, the translation models are trained according as given in the Appendix B. Variants of MDS1 and MDS2 using the StarGAN translators were evaluated and are referred to as MDS$*$1 and MDS$*$2. Moreover, the presented approach is

compared to the method proposed by Brieu et al. [67] which trains a segmentation model using translators taken from multiple epochs of CycleGAN training with the aim to increase augmentation variability. This approach is referred to as Multi UDA-CGAN, where translation models from each 5^{th} epoch are used, resulting in 40 translation models used for augmentation in total. The F_1-score, along with precision and recall, are used to measure performance.

The presented results are the averages of five independent training repetitions, with corresponding standard deviations. The highest F_1-scores for each staining are in bold (PAS is not included in the vPAS average since it is the training staining). Baseline performances (U-Net models trained and tested on the same stain using each stain's groundtruth) were determined for each staining and presented in Table 4.2 (repeated here from Chapter 3 for ease of reading).

Despite the fact that the translations obtained using both CycleGAN and StarGAN look plausible (see Chapter 3), it can be observed that the direction of translation (MDS1 vs MDS2) and translation model (StarGAN vs CycleGAN) influence the results. This is best illustrated with MDS1, in which a model trained on the original PAS data is applied to target data translated to PAS. It can be observed that in each target stain, the difference between CycleGAN and StarGAN translations is significant, although there appears to be no significant difference in the quality of the translations. This phenomenon has been discussed previously in Chapter 3 of this thesis[8]. The proposed UDA-GAN approaches show more stable performance while the best results are obtained using UDA-CGAN.

Furthermore, UDA-CGAN reaches almost baseline performance in three out of five test stainings. Even though the model has seen data from PAS stain only 20% of the time during training, the model has baseline performance on this (source) domain. The model also approaches baseline performance in target stains Jones and Sirius Red. For stains CD68 and CD34, the model reaches an F_1-score of 0.705 and 0.799, meaning that it gives an improvement of 11.9% and 6% respectively over the next best CycleGAN method (MDS2). The average performance over the five different stainings shows that UDA-CGAN reaches an average F_1-score of 0.827 (0.808 without including the PAS staining, in order to be fairly compared to the MDS approaches), while MDS2, as the next best method, reaches an F_1-score of 0.748. The biggest relative difference is observed in staining CD68, where the overall improvement is 55.8% compared to the original approach [17] and 11.9% compared to MDS2. Other than the baseline, UDA-CGAN is the only to achieve acceptable results in this staining. Multi UDA-CGAN does not improve upon this, possibly because it introduces too much variability.

Gadermayr et al. suggest that translation "should always be performed from the difficult-to-segment to the easy-to-segment domain" and that MDS1 is the preferred method [20]. Table 4.2 shows that the difficult-to-segment stain is CD68 and the easy is PAS, as it results in a more accurate baseline segmentation. It is also observed in Table 4.1 that MDS1 with PAS as the source domain is, in fact, surpassed

[8]The results reported in Chapter 3 correspond to the MDS1 approach. However, they are different due to the implementation differences related to the framework used (Keras vs Tensorflow). As previously discussed in Chapter 3, such differences can be expected due to multiple factors (training repetitions and duration). Appendix C.1.1 gives UDA-GAN results using the same models as in Chapter 3.

Table 4.1 UDA-GAN — Quantitative results for each strategy trained on PAS (source staining) and tested on different (target) stainings.

Training Strategy	Score	Test Staining					
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
vPAS	F_1	0.907 (0.009)	0.085 (0.034)	0.001 (0.002)	0.016 (0.018)	0.071 (0.063)	0.043 (0.041)
	Precision	0.885 (0.023)	0.055 (0.021)	0.097 (0.129)	0.034 (0.034)	0.257 (0.243)	0.111 (0.101)
	Recall	0.932 (0.014)	0.418 (0.316)	0.001 (0.001)	0.073 (0.101)	0.058 (0.039)	0.137 (0.190)
UDA-SD [17]	F_1	0.891 (0.007)	0.791 (0.079)	0.147 (0.048)	0.828 (0.046)	0.739 (0.026)	0.679 (0.303)
	Precision	0.840 (0.018)	0.699 (0.116)	0.365 (0.238)	0.778 (0.088)	0.695 (0.054)	0.675 (0.183)
	Recall	0.950 (0.007)	0.926 (0.015)	0.099 (0.032)	0.892 (0.024)	0.795 (0.059)	0.732 (0.359)
MDS1[20]	F_1	-	**0.872** (0.016)	0.395 (0.057)	0.828 (0.040)	0.673 (0.033)	0.692 (0.215)
	Precision	-	0.843 (0.036)	0.447 (0.092)	0.787 (0.071)	0.857 (0.033)	0.734 (0.193)
	Recall	-	0.904 (0.018)	0.364 (0.071)	0.877 (0.020)	0.556 (0.047)	0.675 (0.261)
MDS2[20]	F_1	-	0.869 (0.020)	0.586 (0.059)	0.797 (0.040)	0.739 (0.044)	0.748 (0.121)
	Precision	-	0.833 (0.049)	0.519 (0.108)	0.699 (0.061)	0.723 (0.051)	0.695 (0.132)
	Recall	-	0.909 (0.013)	0.697 (0.059)	0.929 (0.004)	0.765 (0.106)	0.825 (0.112)
UDA-CGAN	F_1	**0.901** (0.011)	0.856 (0.036)	**0.705** (0.031)	**0.873** (0.025)	0.799 (0.034)	**0.827** (0.078)
	Precision	0.869 (0.034)	0.800 (0.069)	0.690 (0.059)	0.830 (0.051)	0.754 (0.076)	0.789 (0.069)
	Recall	0.936 (0.014)	0.924 (0.012)	0.723 (0.034)	0.922 (0.009)	0.856 (0.036)	0.872 (0.089)
Multi UDA-CGAN	F_1	0.897 (0.010)	0.863 (0.030)	0.684 (0.046)	0.861 (0.021)	**0.808** (0.023)	0.822 (0.084)
	Precision	0.860 (0.021)	0.812 (0.057)	0.648 (0.098)	0.813 (0.043)	0.764 (0.061)	0.779 (0.081)
	Recall	0.937 (0.007)	0.922 (0.009)	0.736 (0.038)	0.917 (0.010)	0.862 (0.032)	0.875 (0.083)
MDS∗1	F_1	-	0.756 (0.086)	0.092 (0.055)	0.599 (0.108)	0.751 (0.033)	0.550 (0.314)
	Precision	-	0.675 (0.136)	0.242 (0.116)	0.496 (0.123)	0.742 (0.092)	0.539 (0.223)
	Recall	-	0.881 (0.029)	0.061 (0.044)	0.780 (0.099)	0.774 (0.070)	0.624 (0.379)
MDS∗2	F_1	-	0.816 (0.060)	0.525 (0.048)	0.837 (0.032)	0.766 (0.030)	0.736 (0.144)
	Precision	-	0.740 (0.096)	0.874 (0.037)	0.785 (0.059)	0.752 (0.030)	0.787 (0.061)
	Recall	-	0.918 (0.008)	0.376 (0.046)	0.901 (0.014)	0.785 (0.073)	0.745 (0.253)
UDA-∗GAN	F_1	0.890 (0.022)	0.807 (0.031)	0.549 (0.081)	0.792 (0.052)	0.758 (0.076)	0.759 (0.127)
	Precision	0.853 (0.043)	0.717 (0.050)	0.794 (0.044)	0.703 (0.085)	0.738 (0.082)	0.761 (0.062)
	Recall	0.933 (0.008)	0.926 (0.010)	0.426 (0.090)	0.913 (0.013)	0.796 (0.135)	0.799 (0.216)

Table 4.2 Quantitative baseline results (standard deviations are in parentheses).

	PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
F_1	0.907 (0.009)	0.864 (0.011)	0.853 (0.018)	0.867 (0.016)	0.888 (0.015)	0.876 (0.022)
Precision	0.885 (0.023)	0.824 (0.020)	0.846 (0.027)	0.801 (0.042)	0.862 (0.015)	0.844 (0.032)
Recall	0.932 (0.014)	0.911 (0.005)	0.856 (0.022)	0.957 (0.018)	0.929 (0.011)	0.917 (0.038)

by MDS2 in all but one target stainings (in which it is equal). These results suggest that the characteristics of "difficult-to-segment" stainings may vary, and the method/translation direction should be adjusted to the specific requirements of a given biological question, i.e. the panel of necessary staining methods.

In the case of staining CD68, neither MDS1, MDS∗1, MDS2, nor MDS∗2 perform well. Poor MDS1 (MDS∗1) performance could indicate that the CycleGAN (StarGAN) translation between the CD68 and PAS domains does not capture the features the PAS model uses for segmentation. On the other hand, poor MDS2 and MDS∗2 performance could indicate that the translation between PAS and CD68 contains features that are not present in real CD68. From a biological viewpoint, this most likely represents the fact that immunohistochemistry for CD68 highlights just one specific, migratory cell population (macrophages) that is not part of the pre-existing tissue architecture, with a brown chromogen, while the anatomical structures are only faintly stained (blue "counterstain" using hemalaun). Strikingly, immunohistochemistry for CD34, a marker for vascular endothelial cells, labelled with a red chromogen, performs much better. This can probably be explained by specific immunohistochemical labelling of anatomical structures (blood vessels) in addition to the blueish counterstain, containing more features that are also covered in the other staining methods (PAS, Jones H&E). This is particularly evident with StarGAN, which exploits common inter-stain characteristics.

Between MDS1 and MDS2 (both CycleGAN and StarGAN translations), the largest difference is seen in stain CD68, which marks a protein exclusively produced by macrophages. PAS, as a chemical reaction staining glycolysated proteins in general, highlights a part of macrophages (co-located, but not overlapping, with CD68). Thus the translation from PAS to CD68 (MDS2, MDS∗2) is easier than the reverse (MDS1, MDS∗1) since PAS contains some (but not all) of the information exposed by CD68.

The fact that UDA-CGAN outperforms both MDS1 and MDS2 using the same translation functions indicates that it is capable of extracting more general (stain invariant) features, i.e. it avoids learning stain-specific and 'false' features introduced by the CycleGAN. This is also the case for UDA-∗GAN but to a lesser extent. UDA-∗GAN uses the same generator for all translations, so it is likely to extract similar features between the source stain and all target stains. This reduces the impact of the multi-stain augmentation, which becomes evident when comparing UDA-∗GAN to MDS∗2.

4.1.3 Feature Distributions

In order to visualise the distributions of each model's extracted features, Figure 4.3 presents UMAP (Uniform Manifold Approximation and Projection) embeddings [185] of the penultimate convolutional layer's activations in the best performing model (over all stainings) for two hundred random glomeruli and two hundred random tissue patches of each staining using MDS1, MDS∗1, UDA-CGAN, and UDA-∗GAN (MDS2 and MDS∗2 models are stain specific, therefore cannot be applied to all stainings). For MDS1 and MDS∗1, the translations to the PAS stain are achieved using CycleGAN and StarGAN, respectively, and UDA-CGAN and UDA-∗GAN are applied to original data without any modification.

In order to quantitatively measure these distributions, silhouette scores [186] have been calculated between:

- each stain's glomeruli class and the union of the glomeruli samples from all other stains;
- PAS glomeruli and each target stain's glomeruli;
- each stain's glomeruli and negative samples.

These are presented in Table 4.3. The first should favour the UDA-GAN approaches since their goal is to learn a stain invariant representation, and the second should favour MDS1 and MDS∗1, since their objective is to translate the target stainings to the PAS distribution. At the same time, the third should be the goal of all approaches.

In the first case, the UDA-GAN approaches exhibit larger (or equal) overlap between the glomeruli in all stainings, indicating greater clustering, which is reflected in the fact that these models are able to segment all stainings. Higher scores for the MDS1 approaches indicate less concentrated clustering (e.g. Jones H&E, which appears to be concentrated in one part of the glomeruli space, separate from the other stains).

Since the MDS1 approaches are trained on PAS, they rely on accurate translation models, which must result in a direct overlap with the PAS distribution. In comparison, the UDA-GAN approaches can tolerate more translation variance (that does not result in glomeruli-tissue overlap) since they are trained on the translated data. Interestingly the scores for MDS1 and MDS∗1 for PAS are relatively high, indicating that this approach fails to completely overlap the target stains with PAS.

In the second case, it can be observed that UDA-GAN approaches have better (or equal) overlap with the PAS glomeruli in all stainings. The score for MDS∗1 in CD68 is much higher than for any other stain, and Figure 4.3 shows that the CD68 glomeruli class has been merged with the negative class, explaining the very low recall in Table 4.1.

In the third case, it can be observed that the UDA-GAN approaches better separate the glomeruli and tissue classes. This is especially illustrated in the case of CD68 in which UDA-∗GAN learns to separate the glomeruli and tissue classes, whereas MDS∗1 fails. As mentioned, MDS∗1 relies on accurate translations, whereas the UDA-GAN approaches are able to correct for weak translations during training.

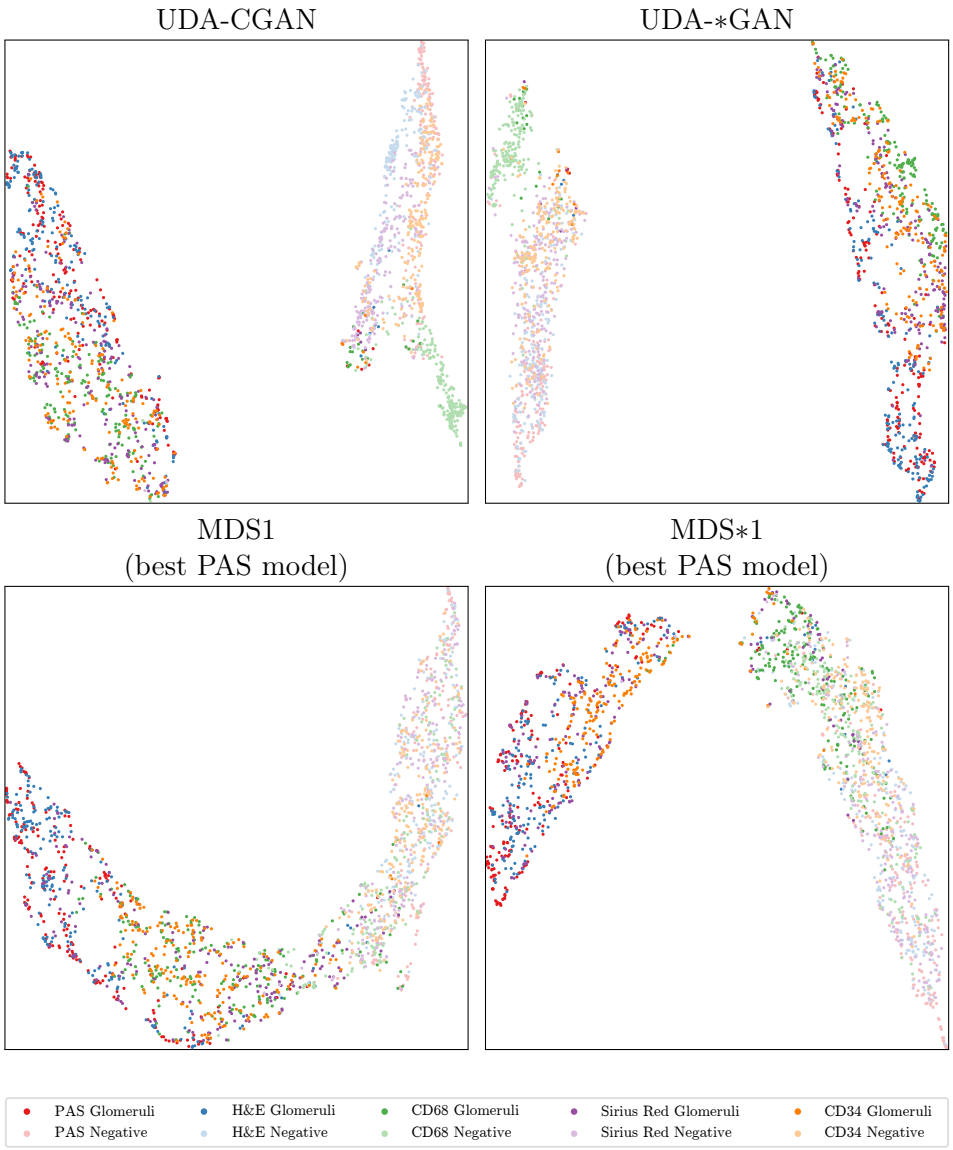


Figure 4.3: Two-dimensional UMAP embeddings of the representation learnt, sampled from the penultimate convolutional layer using 200 patches per stain per class from the overall best performing PAS and UDA-CGAN models. Each point represents a patch from the respective class and staining (glomeruli patches are centred on a glomerulus).

Table 4.3 Silhouette scores measuring (averaged over three different random samplings): the separation between each stain's glomeruli class and the glomeruli class formed from all other stains (Combined), the separation between each stain's glomeruli class and the PAS glomeruli class (PAS), and the separation between each stain's glomeruli class and its negative class. Calculated using the features of 200 random patches per stain per class, extracted from the penultimate convolutional layer: 0 means total overlap, 1 means total separation, -1 means that samples are more similar to the other class than their own, therefore values closer to 0 are better (in bold) for Glomeruli-Combined and Glomeruli-PAS Glomeruli, and 1 is better (in bold) for Glomeruli-Tissue.

	Training Strategy	PAS	Jones H&E	CD68	Sirius Red	CD34
Glomeruli-Combined	UDA-CGAN	0.069 (0.004)	0.106 (0.008)	**0.000** (0.007)	-0.044 (0.003)	**-0.019** (0.003)
	UDA-∗GAN	**0.051** (0.008)	**0.094** (0.005)	0.069 (0.010)	-0.041 (0.003)	-0.024 (0.003)
	MDS1	0.198 (0.011)	0.176 (0.014)	-0.057 (0.002)	-0.038 (0.004)	-0.070 (0.005)
	MDS∗1	0.219 (0.006)	0.048 (0.011)	0.166 (0.004)	**-0.027** (0.003)	-0.063 (0.004)
Glomeruli-PAS Glomeruli	UDA-CGAN	-	0.004 (0.002)	**0.123** (0.008)	0.078 (0.003)	**0.071** (0.006)
	UDA-∗GAN	-	0.004 (0.001)	0.179 (0.003)	**0.070** (0.002)	0.090 (0.003)
	MDS1	-	**0.002** (0.001)	0.253 (0.007)	0.175 (0.009)	0.255 (0.003)
	MDS∗1	-	0.037 (0.005)	0.477 (0.007)	0.098 (0.003)	0.186 (0.010)
Glomeruli-Tissue	UDA-CGAN	0.594 (0.009)	0.567 (0.017)	**0.533** (0.009)	0.554 (0.012)	**0.551** (0.010)
	UDA-∗GAN	**0.625** (0.016)	**0.584** (0.015)	0.475 (0.014)	**0.581** (0.012)	0.530 (0.005)
	MDS1	0.489 (0.007)	0.481 (0.017)	0.300 (0.009)	0.300 (0.016)	0.424 (0.006)
	MDS∗1	0.489 (0.007)	0.456 (0.019)	0.051 (0.004)	0.354 (0.023)	0.449 (0.005)

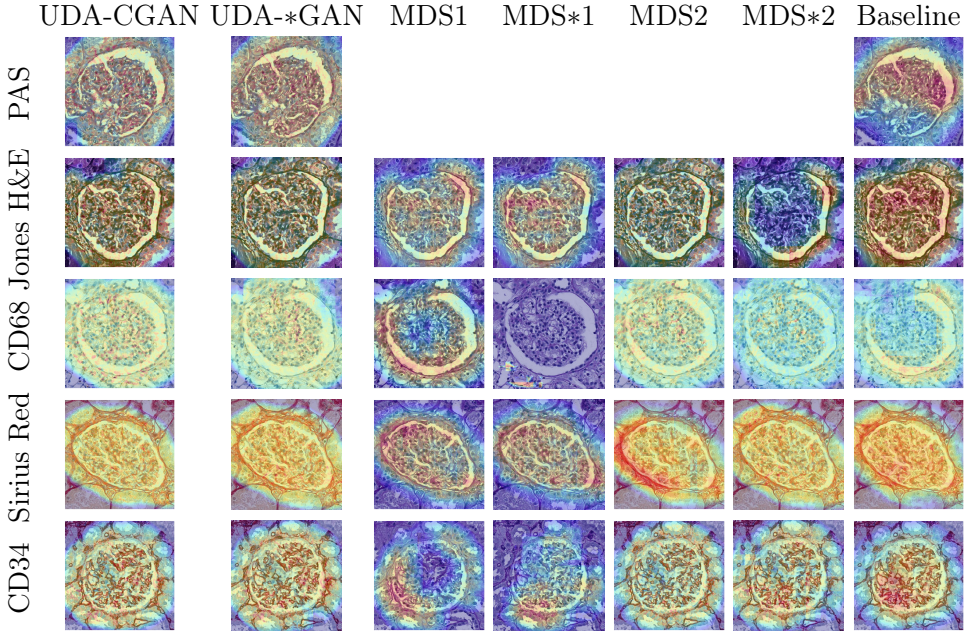


Figure 4.4: GradCAM visualisation of (column-wise) 1) overall best performing UDA-CGAN model; 2) best MDS1 model in each particular stain; 2) best MDS2 model in each particular stain; 3) best baseline model in each particular stain. N.B. the first column represents the attention of the same model over different stainings.

4.1.4 Attention

Another approach to illustrate the representations learned is to use attention visualisations. Grad-CAM [187] is used to visualise the attention of the penultimate convolutional layer, see Figure 4.4. For a fair comparison, the best performing MDS1, MDS∗1, MDS2, and MDS∗2 model for each staining, and the baseline approaches are used. For UDA-CGAN and UDA-∗GAN, the overall best-performing model is taken.

The attention of the best-performing MDS1 (and MDS∗1, except CD68 for which MDS∗1 does not work) models in all stainings is focused on border-like features, which the model uses in the original PAS domain (on which it was trained). Poor MDS1 and MDS∗1 performance can be explained by an absence of the specific features on which the trained model is focused on, and which are not necessarily present in the target domains nor relevant for the detection of the structures in general. When comparing attention in all the baseline models, it can be observed that in each staining, the models have a tendency to focus on stain-specific features. The attention of the MDS2 models is more general and close to the stain-invariant model's attention. According to the results presented in Table 4.1, the stain-invariant approach (UDA-CGAN) gives an improvement in precision, while the recalls of both models are similar (in all stainings except CD34), meaning that the stain-invariant model reduces false positives and, to some extent, false negatives (while both models detect true positives similarly). Thus, attention in the true positive class is expected to be similar. Both the MDS∗2 and UDA-∗GAN use a common translation model

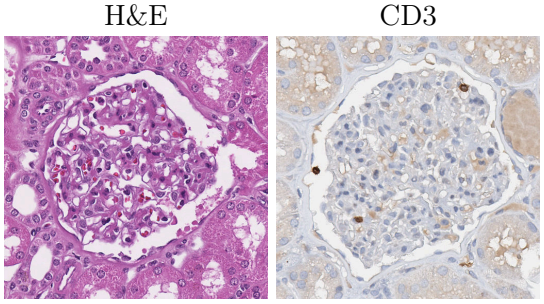


Figure 4.5: Examples of glomeruli from the unseen stains.

(StarGAN), which, as previously mentioned, is biased towards common features between the stains. Thus both models are likely to focus on these common features, explaining their similar attention (this is also confirmed in Table 4.1). In addition to common features, the MDS2 models can also use stain-specific features because they are only exposed to one type of image (as is the case with the presented Jones H&E example).

The main advantage of the UDA-GAN approaches is their ability to properly generalise across different stainings, as the presented attentions represent the features learnt by one model, while in all other cases a domain (stain) dependent model is obtained.

4.1.5 Unseen Stains

In order to further evaluate the stain invariance of the UDA-GAN approaches, they are tested on two unseen stains (unseen to both the CycleGAN/StarGAN and UDA-GAN models), see Figure 4.5: histological stain H&E (a general overview staining not specific for a protein) and immunohistochemical stain CD3 (T cell marker). Even though they highlight similar structures to the "virtually" seen stains, they are visually very different in appearance. For each stain, images are taken from 3 patients containing 1151 (H&E) and 1083 (CD3) glomeruli. Table 4.4 presents the results, averaged over the previously trained UDA-GAN models.

Although the results are lower than those obtained using stains virtually seen during training, i.e. stain translation targets during augmentation, they confirm the network's capacity for stain invariant segmentation. On average, both UDA-CGAN and UDA-∗GAN perform equally well on unseen stains. When taken in the context of Table 4.1, although they generally achieve lower results, UDA-∗GAN remains within the range previously seen, and UDA-CGAN exhibits more variance.

A high F_1-score is achieved when facing a completely new stain colour profile (H&E). When faced with a similar stain profile to one virtually seen (CD3, similar to CD68), the corresponding UDA-CGAN F_1-score is similar, and UDA-∗GAN improves, likely because CD3 has more contrast (due to an unspecific reaction of the primary antibody) when compared to CD68, with which it struggles. It is worth noting that the models with the best PAS performance (which can be determined since annotations exist) are also those that perform best on unseen stains.

Table 4.4 Quantitative results of UDA-GAN models on unseen stains (standard deviations are in parentheses).

Training strategy	Score	H&E	CD3	Average
vPAS	F_1	0.126 (0.066)	0.000 (0.000)	0.063
	Precision	0.070 (0.041)	0.018 (0.018)	0.044
	Recall	0.854 (0.109)	0.000 (0.000)	0.427
UDA-CGAN	F_1	0.731 (0.100)	0.658 (0.075)	0.694
	Precision	0.781 (0.122)	0.569 (0.124)	0.675
	Recall	0.697 (0.123)	0.802 (0.042)	0.750
UDA-∗GAN	F_1	0.752 (0.087)	0.650 (0.030)	0.701
	Precision	0.824 (0.057)	0.853 (0.065)	0.839
	Recall	0.706 (0.147)	0.531 (0.058)	0.618

4.1.6 Multi vs Single Stain Translation

The fact that StarGAN has a single translator may force it to preserve common features between stains, features that UDA-∗GAN is likely to focus on. It is also likely that these features are general and present in unseen stains, therefore UDA-∗GAN's performance is similar within the virtually seen and unseen stains. Nevertheless, the single translator may force StarGAN to ignore hard-to-translate stains, e.g. CD68, and to minimise its loss upon the other stains. This would cause the general features to be extracted from the remaining stains. Low recall for CD68 and CD3 indicates that the model struggles to identify glomeruli, i.e. the general features do not exist in these stains. This is confirmed in Figure 4.3, in which the CD68 glomeruli class overlaps the negative class. UDA-CGAN, on the other hand, is trained using the translators and has more sources of variation since each augmentation translator is independent. It can therefore leverage the additional information present in the augmented translations to achieve higher accuracy during application to them, while still learning a stain invariant representation.

The obtained results can be interpreted in light of the findings presented in Chapter 3 related to the noise injected into an image during translation. These analyses, and recent studies on the adversarial nature of CycleGANs [173, 174], lead to the hypothesis that the translations suffer from invisible artefacts produced by the translation models. The extent and type of these artefacts could be related to stain differences. As previously discussed in Chapter 3, stains with a greater difference in highlighted tissue components require more complicated translation, forcing the translators to hallucinate specific features. The relatively high sensitivity to noise (therefore potentially high levels of noise present in the translation) offers an explanation for UDA-CGAN's relatively low precision in CD68 and CD34, and UDA-∗GAN's relatively low precision overall. From this perspective, it can also be understood why UDA-CGAN outperforms MDS1 and MDS2 as the model is forced to be robust to artefacts produced by different translations. In this sense, the lack of improvement offered by Multi UDA-CGAN model (see Table 4.1) over UDA-CGAN can be understood. In MDS1, these artefacts hamper the performances as translated images could act as adversarial examples, while in MDS2, these artefacts

could be considered by the model as features. The hypothesis is that a mechanism for assessing the quality of the translation, such as FID [188] in natural images, would offer a further understanding of these phenomenons and possibly improvement in developing better translations and stain invariant models.

4.2 Adversarial Domain Adaptation

Instead of putting a constraint directly into the input data (pixel-space methods), such as reducing their variability as presented in Chapter 3, or increasing their variability as presented in Section 4.1, a model can be guided by constraints imposed in feature space (feature-space methods). The most common strategy in feature-space-based methods is to enforce the extraction of domain agnostic and task-related features. Contrarily to pixel-space methods, these approaches are mathematically better defined since it is assumed that all domains contain relevant information for solving a particular task, and thus a domain-invariant set of features can be found. However, obtaining such a set is not always straightforward.

One widely adopted approach to feature-space alignment in digital histopathology is Domain-Adversarial Training of Neural Networks (DANN) [189]. Although this method has been surpassed as the state-of-the-art in general domain adaptation, due to its simplicity, it has been widely applied in digital histopathology [118–121, 143, 145]. As an illustration, in the recent Mitosis domain generalisation challenge 2022, the reference model and two submitted solutions used DANN [190], demonstrating its usefulness even nowadays. However, DANN is primarily used in stain normalisation tasks, to improve classification [119–121, 143, 145] or segmentation [118] performance and it has not yet been studied in the more challenging task of reducing the domain shift induced by inter-stain variation.

4.2.1 DANN for Stain Invariant Segmentation

Domain-Adversarial Training of Neural Networks (DANN) [189] is an unsupervised domain adaptation strategy which forces domain-invariant feature extraction via a *Gradient Reversal Layer (GRL)*. As previously discussed in Chapter 2 (Section 2.2), when performing a feature-space adaptation, the model is divided into feature extractor and task-specific branch and extended with the discriminator. In DANN approach, a GRL connects the feature extractor and the discriminator, and acts as an identity function during the forward pass, while during the backward pass, it negates the gradients. Feeding both source (annotated data) and target (unannotated data) to such model's composition, makes that the discriminator's loss is minimised by itself and maximised by the feature extractor. The training is performed end-to-end, i.e. both models are updated in a single iteration. Optimising the network in this way is a special type of adversarial game [146] that eventually trains the feature extractor to extract domain invariant features. Since the task-specific branch is trained using a domain invariant representation extracted by feature extractor, the model should be able to generalise to the target data.

The DANN approach has been widely adopted in the area of digital histopathology, mainly for classification tasks. A classification model is usually split into a feature-extractor up to the last convolutional layer and the classification part

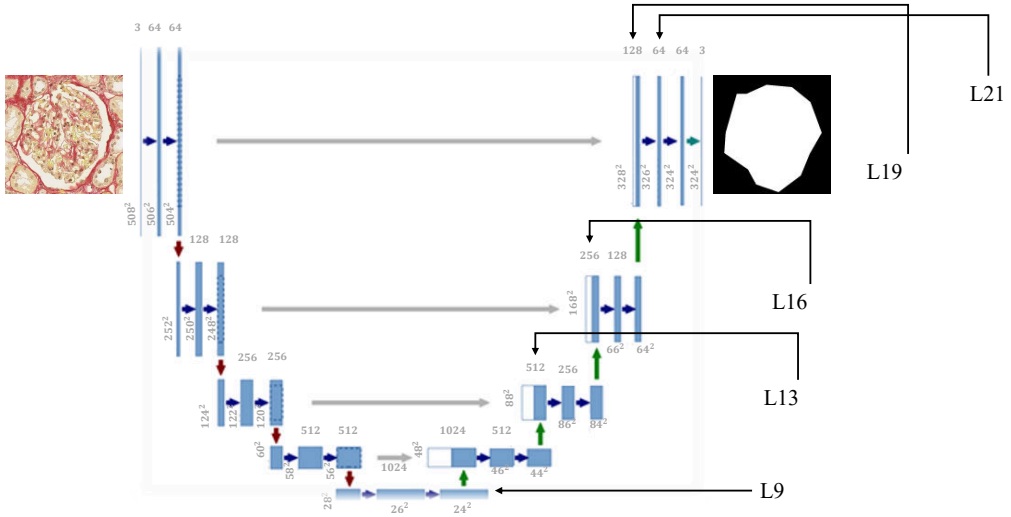


Figure 4.6: UNet architecture (taken from [13]) with a feature map size corresponding to the input patch of 508×508 pixels. Layers $L9$, $L13$, $L16$, $L19$, and $L21$ correspond to the last (transpose) convolutional layers in the marked blocks.

which mainly contains fully-connected layers. However, for segmentation tasks in histopathology, it is rarely applied [118]. One potential reason is the complexity of segmentation models, which are usually encoder-decoder architectures with skip-connections, such as the UNet [13]. Thus, it becomes more complex to determine which part of the network should be considered as a feature-extractor and which will be task-related (segmentation). In Subsection 4.2.1.1 the original DANN will be applied for the task of glomeruli segmentation in multiple stains. The limitations of such an approach will be discussed. Subsection 4.2.1.2 proposes a combination of stain transfer and DANN, which significantly enhance the adaptation process.

4.2.1.1 Vanilla DANN

As in all previous experiments, the PAS staining is considered to be annotated. Therefore, a pre-trained PAS segmentation model is adapted using unannotated data from each of the target domains {Jones H&E, Sirius Red, CD68, CD34}. The UNet architecture used in the segmentation model is given in Figure 4.6. Several layers are investigated to determine at which layer the feature extractor and segmentation model should be separated. Following the conclusions by Brion et al. [191] the adaptation is evaluated on layers 9, 13, 16, 19, 21. In preliminary experiments, it has been found that starting from a pre-trained UNet (source, annotated domain) stabilised training. Starting from a randomly initialised segmentation network that is simultaneously trained and adapted, in most cases (runs and stains), diverged. Thus, the UNet is pre-trained for 100 epochs using PAS data, training details given in Appendix C.2.1.

Table 4.5 shows the obtained results for each of the tested layers. The baseline results, which represent the direct application of the PAS pre-trained model to the target data, prior adaptation, are reported in column vPAS. For each of the tested layers, three separate adaptation runs are performed, and the overall averages with

Table 4.5 The segmentation performance of the PAS model adapted at different layers to a target stain.

Target stain	Measure	vPAS	Adaptation layer				
			Layer 9	Layer 13	Layer 16	Layer 19	Layer 21
Jones H&E	F_1	0.000	0.795 (0.011)	0.815 (0.016)	0.789 (0.021)	0.778 (0.016)	**0.828 (0.007)**
	Precision	0.083	0.788 (0.054)	**0.812 (0.037)**	0.784 (0.022)	0.785 (0.008)	0.801 (0.017)
	Recall	0.000	0.807 (0.048)	0.818 (0.019)	0.796 (0.053)	0.771 (0.025)	**0.856 (0.008)**
Sirius Red	F_1	0.000	0.065 (0.071)	0.653 (0.196)	0.006 (0.002)	0.287 (0.482)	**0.848 (0.012)**
	Precision	0.003	0.121 (0.137)	**0.926 (0.037)**	0.016 (0.006)	0.678 (0.284)	0.875 (0.011)
	Recall	0.000	0.055 (0.048)	0.535 (0.231)	0.004 (0.002)	0.272 (0.463)	**0.823 (0.033)**
CD68	F_1	0.000	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	Precision	0.000	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	Recall	0.000	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
CD34	F_1	0.000	0.151 (0.261)	**0.748 (0.015)**	0.636 (0.063)	0.497 (0.202)	0.230 (0.374)
	Precision	0.000	0.139 (0.241)	**0.845 (0.045)**	0.771 (0.000)	0.766 (0.025)	0.393 (0.396)
	Recall	0.000	0.164 (0.285)	**0.674 (0.051)**	0.546 (0.093)	0.404 (0.262)	0.194 (0.323)

corresponding standard deviations are reported.

It is not surprising that the direct application of the PAS-based model cannot segment in target stains due to domain shift caused by intra-stain variation, which is confirmed by vPAS results in Table 4.5. Therefore, the adaptation in different UNet layers has the potential to reduce such domain shift. Adapting at the UNet's bottleneck (Layer 9) is not able to reduce the domain shift in any of the test stains except Jones H&E. Generally, adapting between PAS and Jones H&E is the least dependent on the position of the GRL. Such findings align with all results presented in this thesis since several presented approaches successfully remove the domain shift between these two stains (see Chapter 3, Section 4.1 and Chapter 5). As already discussed, one potential explanation is that these two stainings are biologically closer than others, simplifying the adaptation process. When stains are far from each other, which is the case for the other stainings in these experiments, adapting at the bottleneck is not always effective. For example, adapting to stain CD34 in 1 out of 3 runs gives an overall F_1-score of 0.452; however, in the other two runs, the adaptation was unsuccessful, giving an overall F_1-score close to 0. Regarding the other layers, the adaptation at Layer 21 quantitatively gives the best results in two out of four tested stainings. However, in stain CD34, the adaptation at this layer

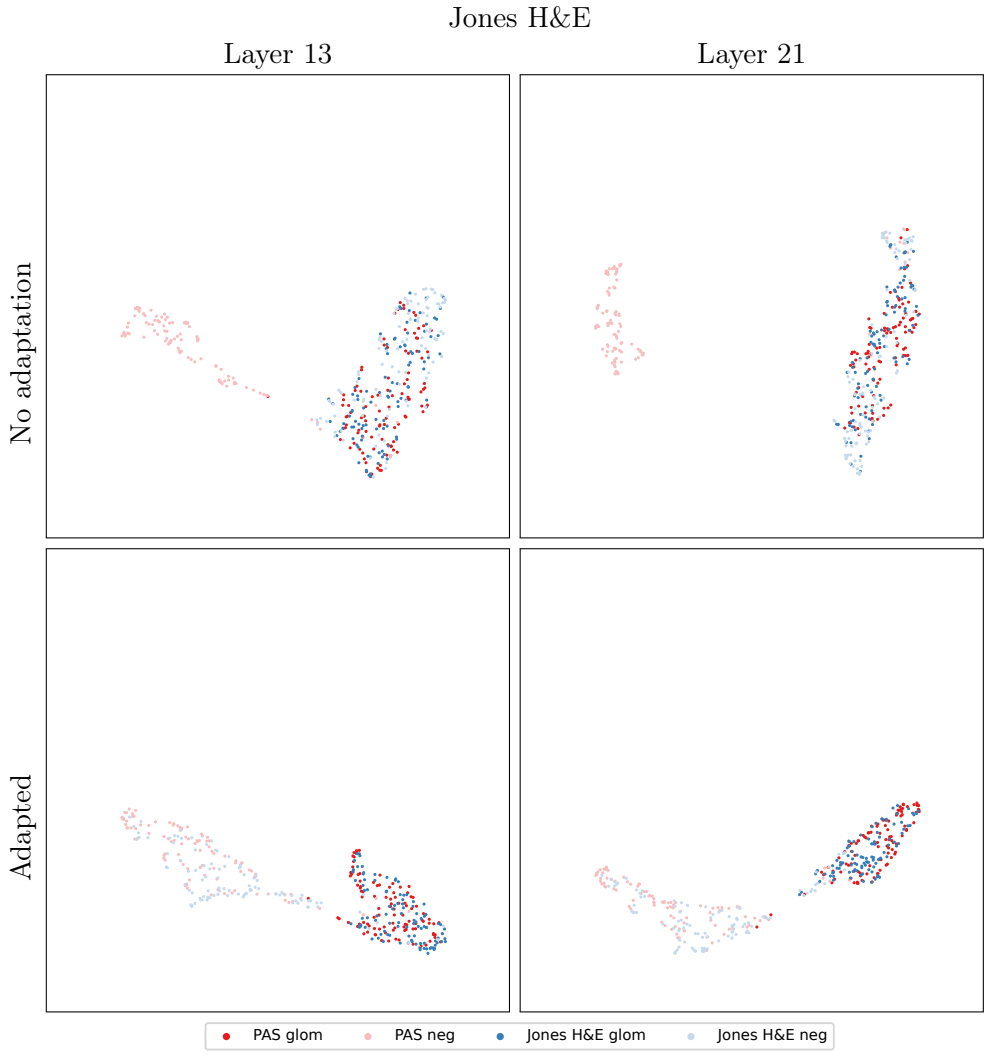


Figure 4.7: Two-dimensional UMAP embeddings of the Jones H&E representations in the segmentation model (pre-trained on PAS data), sampled from Layer 13 and Layer 21, before and after the adaptation, using 100 patches per stain per class. Each point represents a patch from the respective class and stain (glomeruli patches are centred on glomeruli).

is unstable — similarly to adaptation at the bottleneck for stain CD34 in 1 out of 3 runs, it gives the best overall F_1-score of 0.66 while in the other two runs, the overall F_1-score is close to 0. Overall, adapting at Layer 13 gives the most stable results across all the tested stainings, and quantitatively, they are close to the best results obtained for each stain generally.

In order to visualise the adaptation process, UMAP embeddings of 200 random patches from the source and target domains are plotted (100 glomeruli patches and 100 tissue patches), see Figure 4.7 and Figure 4.8. The embeddings are plotted for adaptation Layer 13 and Layer 21 before and after the adaptation. Figure 4.7 presents UMAP embeddings for Jones H&E (the best performing stain) and Figure

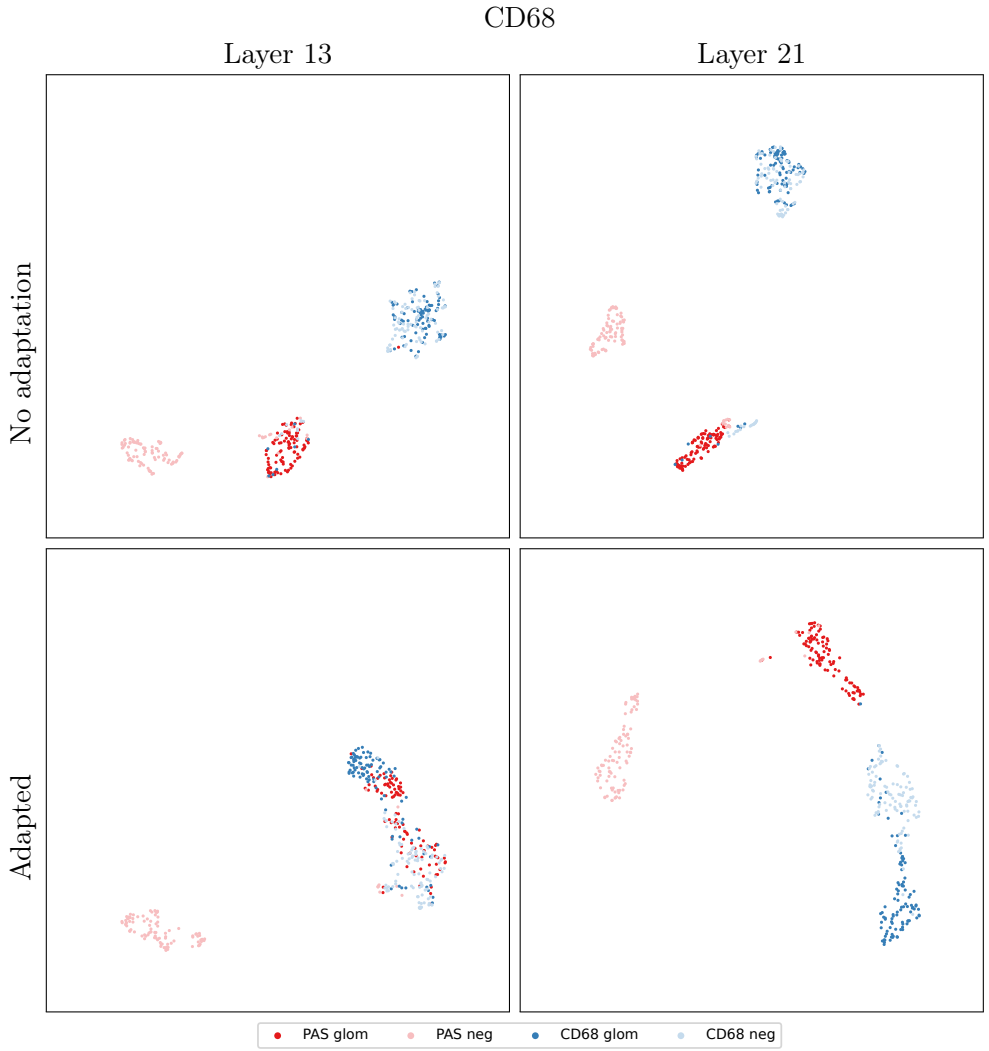


Figure 4.8: Two-dimensional UMAP embeddings of the CD68 representations in the segmentation model (pre-trained on PAS data), sampled from Layer 13 and Layer 21, before and after the adaptation, using 100 patches per stain per class. Each point represents a patch from the respective class and stain (glomeruli patches are centred on glomeruli).

4.8 presents CD68 (the worst-performing stain). It can be observed that for Jones H&E, adaptation leads to proper alignment in the adapted layer, which is preserved in later layers (Layer 21) and results in good segmentation performance. However, for stain CD68, adaptation is successful at the adaptation layer, but the alignment is not preserved in the latter layers, which results in poor segmentation results.

The obtained adapted models are invariant with respect to two stainings — PAS and target stain to which it has been adapted. The performances of the adapted model on the source stain (PAS) are presented in Table 4.6. The results of the baseline model (a pre-trained model from which the adaptation process starts) are given in column PAS. The other columns contain the results of the adapted model

Table 4.6 Segmentation models adapted at Layer 13 for the corresponding target staining applied to PAS data. N.B. models are able to segment in both source and target domains (excluding CD68 where adaptation is not successful for the target staining and the resulting model works only in the PAS domain).

Score		Target Staining			
	PAS	Jones H&E	CD68	Sirius Red	CD34
F_1	0.910	0.902 (0.006)	0.911 (0.001)	0.899 (0.003)	0.891 (0.006)
Precision	0.882	0.884 (0.010)	0.888 (0.002)	0.885 (0.005)	0.884 (0.010)
Recall	0.940	0.920 (0.011)	0.937 (0.001)	0.912 (0.010)	0.897 (0.004)

(at Layer 13) when applied to PAS data. These results confirm that the adaptation process does not affect the source domain's performance and that the model can segment across two stainings simultaneously. The exception is CD68 staining, where adaption is not successful for the target stain. Nevertheless, the adaptation did not ruin the performance of the source domain.

The instability of the Vanilla DANN approach can be attributed to the differences between stains. One way to improve its performance is to force distribution alignment at several layers. Alternatively, the adaptation task can be facilitated by first reducing domain shift in the pixel-space domain. However, as has been shown, stain transfer can produce an imperceptible domain shift. In the following, DANN will be used in combination with stain transfer as a mechanism to remove the remaining domain shift between real and virtually stained images.

4.2.1.2 Stain Transfer for DANN

A CycleGAN model is trained to translate between PAS and each target stain. The model architecture and training strategy are the same as in Chapter 3. Separate segmentation models are trained for each target staining, using the annotated stain (PAS) translated to a target by pre-trained CycleGAN models. To be consistent with the previous section, this approach is referred to as MDS2 [20]. Furthermore, for each target stain three MDS2 segmentation models are trained, whose average performances in the target domain are given in Table 4.7, MDS2 row. In order to reduce the remaining domain shift, the DANN approach is adopted. According to the results presented in Table 4.5, Layer 13 is the most promising for attaching a discriminator. A separate adaptation is performed for each target segmentation model, starting from a pre-trained MDS2 model, using unannotated target data and annotated PAS-stain data translated to the target. Training details and discriminator architecture are the same as in the Vanilla DANN approach.

Table 4.7 presents the results after adaptation in row MDS2-adapted, which a) show that the DANN approach is able to improve results in all stainings; b) confirm the existence of a feature-space discrepancy between real and virtually stained images. The individual results of the improvements obtained for each MDS2 model in each target stain are in Appendix C.2.2.

In three out of four target stainings, this strategy obtained close-to baseline results (baseline results for each stain are given in Table 4.2). In stain CD68, this strategy achieves 0.708 F_1-score, which is an improvement of 10% compared to the

Table 4.7 MDS2 DANN adaptation results.

Training Strategy	Score	Test Staining			
		Jones H&E	CD68	Sirius Red	CD34
MDS2[20]	F_1	0.876 (0.003)	0.603 (0.072)	0.814 (0.034)	0.863 (0.018)
	Precision	0.855 (0.005)	0.499 (0.100)	0.730 (0.057)	0.825 (0.038)
	Recall	0.899 (0.002)	0.790 (0.032)	0.924 (0.006)	0.906 (0.007)
MDS2[20] adapted	F_1	**0.880** (0.002)	**0.708** (0.023)	**0.849** (0.011)	**0.871** (0.007)
	Precision	0.851 (0.007)	0.691 (0.041)	0.781 (0.024)	0.845 (0.016)
	Recall	0.912 (0.004)	0.735 (0.077)	0.932 (0.007)	0.899 (0.006)

Table 4.8 Silhouette scores measuring the separation between each stain's glomeruli class and the PAS (translated to given staining) glomeruli class (averaged over three models). Calculated using the features of 100 random patches per stain, extracted from the adaptation layer (Layer 13). Values closer to zero mean better alignment.

Training Strategy	Test Staining			
	Jones H&E	CD68	Sirius Red	CD34
MDS2[20]	0.007 (0.004)	0.017 (0.010)	0.023 (0.006)	0.008 (0.003)
MDS2[20] adapted	**0.005** (0.004)	**0.005** (0.002)	**0.010** (0.006)	**0.007** (0.002)

non-adapted stain-transfer-based model. In the other three stains, DANN-based adaptation boosted performance to close to baseline (fully supervised training on annotated target data, see Table 4.2).

In order to visualise the adaptation process, UMAP embeddings from 200 random patches from the source and target domains are plotted (100 glomeruli patches and 100 tissue patches), see Figure 4.9 and Figure 4.10. It can be seen that since the process starts from an MDS2 model (trained on PAS → target images), which is already able to segment glomeruli in the target domain, the domains appear to be aligned even without adaptation. The adaptation further refines this alignment, which is confirmed by the silhouette scores given in Table 4.8, where lower values are obtained in the adapted compared to the MDS2 model. Such alignment is reflected in the overall increase in precision (less false positives), which is observed in all tested stainings.

The same strategy can be applied to MDS1, which assumes that the translated images (target to source) are segmented using pre-trained source models (in this case, PAS). A domain shift exists between the source data and the target translations to the source. As in the previous experiment, the DANN approach is used to reduce the domain shift. The obtained results are presented in Table 4.9. Surprisingly, the adaptation in the direction target→PAS is not as successful as the direction PAS→target. An improvement is observed in stainings CD68 and CD34, which are, as previously discussed, biologically far from PAS and thus it is expected that stain transfer injects more noise into translations (i.e. the domain shift is bigger than in other stainings). However, for histochemical stainings Jones H&E and Sirius Red (biologically closer to PAS), MDS1 already achieves baseline results (see Table 4.2) and adaptation negatively affects the performances. The drop in performance mainly comes from the decrease in precision (more false positives) that is observed in all tested stainings. Such behaviour can be explained by classes mixing in the

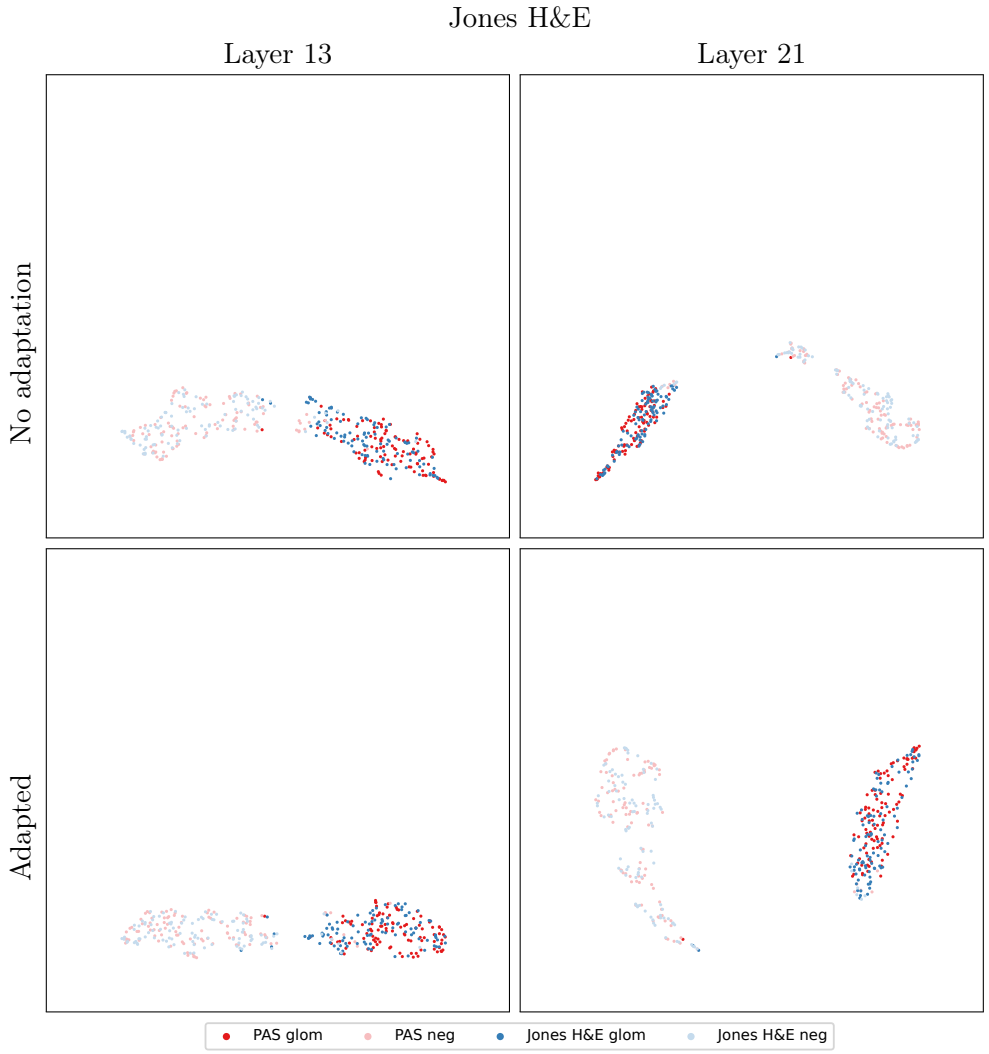


Figure 4.9: Two-dimensional UMAP embeddings of the Jones H&E representations in the MDS2 segmentation model (pre-trained on PAS translated to Jones H&E), sampled from the Layer 13 and Layer 21, before and after the adaptation, using 100 patches per stain per class. Each point represents a patch from the respective class and stain (glomeruli patches are centred on glomeruli).

feature space, and thus more advanced domain adaptation techniques such as FADA [192] can be beneficial. The potential of such methods for both MDS1 and MDS2 remains to be explored in the future.

4.3 Conclusions

To summarise, this chapter introduced UDA-GAN, a state-of-the-art model for stain invariant segmentation that outperforms all other existing pixel-space alignment approaches in five different stainings. The model is domain invariant (including

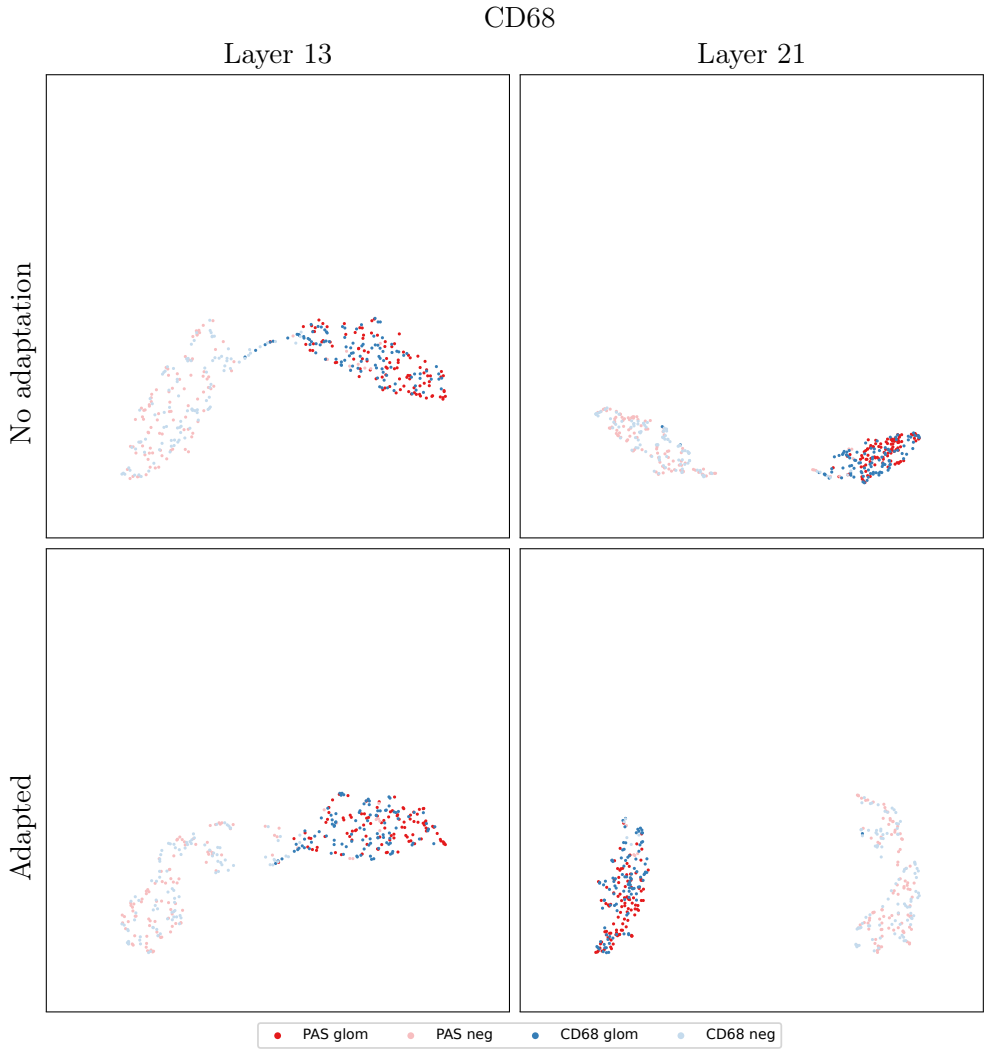


Figure 4.10: Two-dimensional UMAP embeddings of the CD68 representations in the MDS2 segmentation model (pre-trained on PAS translated to CD68), sampled from Layer 13 and Layer 21, before and after the adaptation, using 100 patches per stain per class. Each point represents a patch from the respective class and stain (glomeruli patches are centred on glomeruli).

unseen stains), can be easily extended to new stainings, and the training procedure is general so that it can be used in different segmentation and classification tasks.

UDA-GAN's stain invariance has been shown through quantitative results and by analysing the resulting feature distribution and attention, demonstrating that patches from all stainings are better aligned than competing approaches. The approach uses pixel space alignment, which aids visual interpretation and verification. The results have been discussed and related to those found in the literature. Namely, choosing translation model epochs by visual inspection is not the best approach (although no formal method for selecting the epochs to be used nor evaluating quality exists), and the direction of translation of the data cannot be simply prescribed a

Table 4.9 MDS1 DANN adaptation results.

Training Strategy	Score	Test Staining			
		Jones H&E	CD68	Sirius Red	CD34
MDS1[20]	F_1	**0.871**(0.023)	0.641 (0.033)	**0.884** (0.018)	0.764 (0.026)
	Precision	0.847 (0.040)	0.867 (0.018)	0.854 (0.036)	0.891 (0.009)
	Recall	0.897 (0.003)	0.510 (0.045)	0.918 (0.006)	0.669 (0.043)
MDS1[20] adapted	F_1	0.839 (0.037)	**0.675** (0.031)	0.861 (0.005)	**0.816** (0.019)
	Precision	0.791 (0.064)	0.700 (0.091)	0.813 (0.006)	0.803 (0.047)
	Recall	0.893 (0.003)	0.667 (0.099)	0.915 (0.004)	0.830 (0.010)

priori. Finally, UDA-GANs limitations were presented; it is still unclear exactly why the model's performance degrades with certain stainings despite visually accurate translations. One hypothesis is that some stainings are far from each other in terms of biological structures highlighted; in particular immunohistochemical staining methods highlighting migratory immune cells can be far from conventional histology staining methods making the translation task harder. Therefore, the translation model is forced to introduce hidden information. Consequently, distributional differences between features extracted from real and virtually stained images could still exist, as previously demonstrated in Chapter 3.

This chapter also presented an approach to reducing inter-stain domain shift by adversarial training for feature-space alignment. A widely used mechanism for domain adaptation based on Gradient Reversal Layer (GRL) is used for this purpose. It is found that GRL position can affect the adaptation process and that the final results are prone to high standard deviations for a specific position. Nevertheless, the obtained models are invariant with respect to source and target staining, and it is demonstrated that the model is able to segment across both domains. However, it is shown that this method is not able to reduce intra-stain variation when applied to difficult stain combinations (e.g. PAS and CD68). To overcome such limitations, a combined solution of a stain transfer and feature-space adaptation is proposed. Such a method greatly improves upon domain shift reduction solely based on stain transfer, and the benefits are demonstrated for all stainings, particularly for difficult stain combinations (general and immunohistochemical staining). The method achieves almost baseline results in three out of four tested stainings. These experiments suggest that the best way to address inter-strain variation is a combination of feature and pixel-space alignment strategies, which will be exploited in the next chapter.

Chapter 5

HistoStarGAN — Integrated Stain Transfer and Stain Invariance

Taking into account the differences in tissue components visible under different stainings and the variation that occurs inside one staining protocol, virtual staining can be considered an ill-posed problem since a single image can be virtually stained in multiple valid ways. However, the previously discussed methods for virtual staining (presented in Chapter 3) reduce virtual staining to a deterministic mapping due to their technical characteristics, producing a single and fixed output for a given input. Although training repetitions of the same model, or different epochs during training, can result in different translations (as demonstrated in Chapter 3), the common characteristic of all previously discussed methods is a deterministic translation and therefore limited output diversity upon model training. For tasks that use virtual staining intending to reduce model input variability, such as normalisation methods, the diversity of virtual staining is non-desirable, and thus, methods with deterministic behaviour are widespread in the literature (see Chapter 2). Nevertheless, as previously demonstrated in Chapter 4 for inter-stain variation and in the literature for intra-stain variation [111], invariant solutions can greatly benefit from diverse translations of a single image. Additionally, if a single model can obtain diverse translations between multiple stainings simultaneously, i.e. a diverse multi-domain model, that would significantly reduce the number of needed virtual staining models, providing a way to obtain stain invariant solutions more efficiently.

Inspired by the current state-of-the-art methods for diverse multi-domain style transfer, such as StarGANv2[15] and TUINT [16], the same methods could be used for diverse virtual staining. However, contrarily to the domain of natural images where style transfer can alter source-specific image characteristics as long as the final output has a large fidelity (i.e. the size of ears is modified when translating a cat to a dog; in female to male transfer, the amount of hair and hairstyle is altered, a beard can be added/deleted, etc.), in the medical domain such extensive alterations are not allowed. For example, in digital histopathology applications, such models could result in removing/inventing a specific cell population (like cancerous) or structures (such as glomeruli), affecting the usefulness of the translations. Thus, the direct application of these models is not straightforward.

This chapter proposes an extension of the StarGANv2 model, named HistoStarGAN, for applications in renal histopathology. The presented model performs plausible and diverse stain transfer between multiple stainings, preserving the structure of interest during the translation process. Additionally, using the annotations provided in single staining, the model obtains stain invariant segmentation of the selected structure across multiple stainings. The proposed solution, explained in more detail in Section 5.1, results in a single model that is, for the first time, able to perform simultaneous stain normalisation, stain transfer and stain invariant segmentation. The HistoStarGAN can, for the first time, generalise stain transfer to unseen stainings, which is demonstrated in numerous examples in this chapter. Moreover, the new state-of-the-art performances for stain invariant glomeruli segmentation is achieved, outperforming the previous (proposed in Chapter 4) by a large margin. Visual and quantitative results are given in Section 5.2. The HistoStarGAN's model architecture and training parameters are analysed in more detail in ablation studies presented in Section 5.3. Upon training, the HistoStarGAN model is able to archive multiple translations for a single image. This property is exploited to generate the first artificially created dataset, KidneyArtPathology, presented in more details in Section 5.4. The KidneyArtPathology dataset can be accessed online [9].

5.1 HistoStarGAN

5.1.1 Model Description

HistoStarGAN architecture is presented in Figure 5.1. The model is composed of five modules: generator (G), discriminator (D), mapping network (F), style encoder (E) and segmentation network (S). The models G, D, F and E are elements of StarGANv2's architecture.

The mapping network F generates a stain-specific style by transforming the random latent code z into the target stain's style. The style is injected into the generator G during translation, which enables diverse generations as different latent codes result in different stain-specific styles. In order to ensure that the generator uses the injected style information, the model is constrained with a style reconstruction loss, i.e. the style encoder E extracts the style from the generated image, and the difference between that style and the style provided to the generator during translation is minimised (blue arrows in the Figure 5.1). In order to explicitly allow style diversification, the model is trained to produce different outputs for different styles in the given target domain by the style diversification loss (orange arrows in the Figure 5.1). Moreover, the model is constrained using a cycle-consistency loss (green brackets in Figure 5.1), e.g. the difference between the original and reconstructed images is minimised. Reconstruction is performed using the same generator, but the style information is extracted from the mapping network by taking the original, real image as the input.

The generator G is an encoder-decoder network, with an instance normalisation layer in the encoder and an adaptive instance normalisation layer in the decoder. In this way, the encoder removes stain-specific characteristics from the image while the decoder injects target-stain characteristics during the generation process. Thus,

[9] https://main.d33ezaxrmu3m4a.amplifyapp.com/

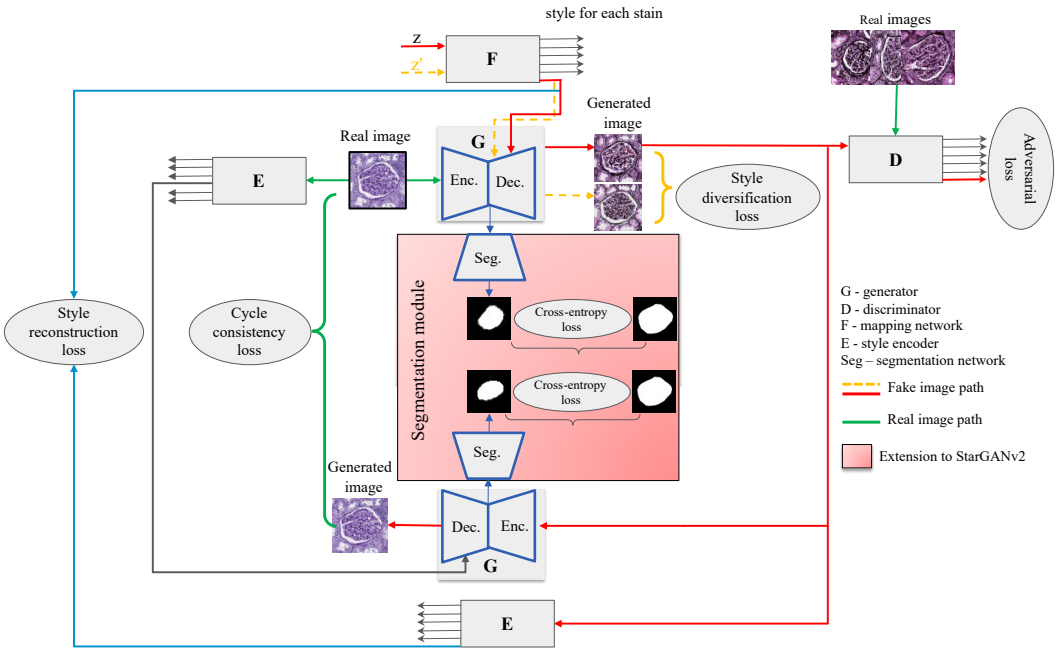


Figure 5.1: HistoStarGAN — an end-to-end trainable model for simultaneous stain transfer and stain invariant segmentation. The red block denotes the difference compared to the StarGANv2 model [15].

the features extracted by the generator's encoder should be stain invariant. Under the assumption that a structure of interest is visible in all stainings, i.e. the stain invariant solution is feasible, the representation extracted in the bottleneck should be sufficient to perform the considered object-related task. Therefore, a segmentation module is attached to the bottleneck. Trained end-to-end with the other modules, this extension forces the preservation of structures of interest during the translation process.

More formally, let $\mathcal{X}$ be the set of histopathological images, $\mathcal{S}$ the set of available segmentation masks for the structure of interest, and $\mathcal{Y}$ the set of stainings found in $\mathcal{X}$. Given an image $x \in \mathcal{X}$, its original staining $y \in \mathcal{Y}$ and corresponding segmentation mask m_{seg}, the model is trained using the following objectives:

Adversarial objective: The latent code $z \in \mathcal{Z}$, source domain $y \in \mathcal{Y}$ and target domain $\tilde{y} \in \mathcal{Y}$ are randomly sampled. The target style code $\tilde{s} = F_{\tilde{y}}(z)$ is obtained via the mapping network F. Thus the generator G, with input image x from domain y and style $\tilde{s}$, generates and output image $G(x, \tilde{s})$. This image is evaluated by the discriminator's output, which corresponds to a domain $\tilde{y}$, $D_{\tilde{y}}$. These are trained using the adversarial objective, such that

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{x,y}[\log D_y(x)] + \mathbb{E}_{x,\tilde{y},z}[\log(1 - D_{\tilde{y}}(G(x, \tilde{s})))]. \tag{5.1}$$

Style reconstruction: In order to ensure that generator G uses the provided style code $\tilde{s}$ when producing output $G(x, \tilde{s})$, the mapping netowrk E is employed to extract the provided style from the generated image, $E_{\tilde{y}}(G(x, \tilde{s}))$, and the following

style reconstruction objective is formed:

$$\mathcal{L}_{\text{sty}} = \mathbb{E}_{x,\tilde{y},z}[\|\tilde{s} - E_{\tilde{y}}(G(x,\tilde{s}))\|_1]. \tag{5.2}$$

Style diversification: The generator is forced to produce different outputs for different styles produced by the mapping network F. Given different latent codes z_1 and z_2, style codes $\tilde{s_1}$ and $\tilde{s_2}$ are generated. Then, the following diversity cost is maximised:

$$\mathcal{L}_{\text{ds}} = \mathbb{E}_{x,\tilde{y},z_1,z_2}[\|G(x,\tilde{s_1}) - G(x,\tilde{s_2})\|_1]. \tag{5.3}$$

Cycle-consistency: The following cycle-consistency loss forces the generator to reconstruct the original image given the source style code $\hat{s} = E_y(x)$:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x,y,\tilde{y},z}[\|x - G(G(x,\tilde{s}),\hat{s})\|_1]. \tag{5.4}$$

Segmentation objective: The cross-entropy loss is used to train the segmentation branch on both real data and their translation to a random stain, such that:

$$\mathcal{L}_{\text{seg}} = \mathbb{E}_{x,m_{seg}}[m_{seg} \log Seg(x)] + \mathbb{E}_{x,m_{seg},z}[m_{seg} \log Seg(G(x,\tilde{s}))]. \tag{5.5}$$

Full objective: The importance of each of these losses is controlled by hyperparameters and combined in the following full objective:

$$\min_{G,F,E} \max_{D} \mathcal{L}_{\text{adv}} + \lambda_{\text{sty}}\mathcal{L}_{\text{sty}} - \lambda_{\text{ds}}\mathcal{L}_{\text{ds}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}. \tag{5.6}$$

5.1.2 Training Setup

5.1.2.1 Dataset

Training the HistoStarGAN model's segmentation branch is supervised, and requires the segmentation masks for all images in the dataset. However, as discussed in previous chapters, this assumption is unrealistic. It is more reasonable to assume that annotations exist for limited examples, e.g. as is common in the field, only for one staining [19, 64]. Similarly as in the previous chapters, the PAS staining is considered to be annotated (the source stain), while the other stainings (Jones H&E, Sirius Red, CD34, CD68) are considered to be unannotated.

To overcome the lack of annotations in target stainings, the CycleGAN's deterministic nature and limited capacity to perform geometrical changes are used to artificially generate a fully-annotated dataset. As previously, the CycleGAN model is trained in an unsupervised manner, using randomly extracted patches from a given pair of stainings. Separate CycleGAN models are trained for each pair of PAS-target stains. When trained, the CycleGAN models are applied to the source (annotated) dataset in order to generate annotated samples in the target stainings.

5.1.2.2 Training Details

As in the previous chapters, the task of glomeruli segmentations is considered. CycleGAN models are trained using the same training setting as described in Chapter 3, for 50 epochs, but with patches of size 512×512 pixels extracted from the training patients in each staining. Upon training, a balanced dataset is formed by translating 500 glomeruli and 500 random negative patches from PAS to all target stains. When training the HistoStarGAN model, extensive data augmentation is performed, following the conclusions by Karras et al. [193] that data augmentation is a crucial factor when training GANs with limited data. The CycleGAN-based dataset is limited in several aspects: a) the size — the annotated dataset is not very big compared to datasets typically used to train GAN models; b) the variability of the glomeruli's size/shape is limited to the variability present in the annotated dataset; c) the capacity of CycleGAN models to mimic a real target distribution affects the appearance of stain-specific tissue components. Thus, adding unsupervised augmentation that leads to realistically-looking images during training is important to increase dataset variability and establish more diverse translations. The augmentations used are horizontal/vertical flipping, affine transformations and elastic deformation, in addition to image enhancement and additive Gaussian noise.

The following augmentations are applied 50% of time with an independent probability of 0.5 (batches are augmented 'on the fly') for each method; elastic deformation ($\sigma = 10$); affine transformations — random rotation in the range $[0°, 180°]$, random shift sampled from $[-5, 5]$ pixels, random magnification sampled from $[0.95, 1]$, and horizontal/vertical flip; brightness and contrast enhancements with factors sampled from $[0.0, 0.2]$ and $[0.8, 1.2]$ respectfully; additive Gaussian noise with $\sigma \in [0, 0.01]$.

The HistoStarGAN model is trained using the following loss weights: $\lambda_{\text{sty}} = 1$, $\lambda_{\text{ds}} = 1$, $\lambda_{\text{cyc}} = 1$, and $\lambda_{\text{seg}} = 5$. To stabilise training, the weight of style diversification loss, λ_{ds}, is linearly decreased to zero over 100 000 iterations [15]. Also, the weight of the segmentation loss, λ_{seg}, is 0 for the first 10 000 iterations until the model starts to generate recognisable images from each stain. Although their fidelity is not high enough, at this moment the segmentation model receives enough meaningful information to start learning. HistoStarGAN is trained for 100 000 iterations. The architectures for the generator, discriminator, style encoder, and mapping network are the same as in the original StarGANv2 architecture [15] as well as optimisers and learning rates. The segmentation branch's architecture is the same as the generator's decoder, without the adaptive instance normalisation layer. The segmentation branch is trained using the Adam optimiser with a learning rate of 10^{-5}. As for the other networks, exponential moving averages over parameters [10] is applied during training to obtain the final segmentation network, as experimentally, it gives better results than the best model saved based on validation performance.

The HistoStarGAN model's segmentation branch is trained using a balanced dataset, mainly containing artificial histological images produced by both CycleGAN and HistoStarGAN. However, tasks in digital histopathology are usually concerned with sparse structures, and using an imbalanced dataset to account for the tissue diversity is beneficial for learning [17, 64]. Moreover, CycleGAN-based translations can be noisy [63], which could affect the stain invariant properties of the segmen-

Table 5.1 Quantitative results for HistoStarGAN compared to UDA-GAN. Each model is trained on annotated PAS (source staining) and tested on different (target) stainings. Standard deviations are in parentheses, and the highest F_1-scores for each staining are in bold.

Model	Score	Test Staining					
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
UDA-GAN	F_1	**0.903** (0.003)	0.849 (0.031)	0.720 (0.016)	**0.875** (0.016)	0.800 (0.033)	0.829 (0.072)
	Precision	0.878 (0.018)	0.787 (0.060)	0.688 (0.110)	0.835 (0.035)	0.720 (0.064)	0.782 (0.079)
	Recall	0.930 (0.014)	0.923 (0.010)	0.777 (0.095)	0.921 (0.007)	0.903 (0.016)	0.891 (0.064)
HistoStar-GAN	F_1	0.871 (0.009)	**0.870** (0.007)	**0.755** (0.006)	0.859 (0.004)	**0.840** (0.004)	**0.839** (0.048)
	Precision	0.845 (0.029)	0.864 (0.019)	0.845 (0.039)	0.883 (0.018)	0.839 (0.032)	0.855 (0.018)
	Recall	0.899 (0.016)	0.877 (0.007)	0.684 (0.024)	0.836 (0.017)	0.842 (0.024)	0.828 (0.084)

tation module. Thus, after training HistoStarGAN as a whole, the segmentation module is fine-tuned for one epoch using real unbalanced PAS-stained images while the rest of the model is fixed (the choice of the number of fine-tuning epochs is discussed in Section 5.3). This dataset contains all PAS glomeruli (662 extracted from the training patients) and seven times more negative patches (4634) to account for tissue variability. The Adam optimiser is used with a batch size of 8 and a learning rate of 0.0001. The same augmentation as in Lampert et al. [17] is applied during fine-tuning.

5.2 Results

This Section will demonstrate that HistoStarGAN results in a single model able to perform diverse stain transfer, stain normalisation and stain invariant segmentation. Moreover, having a stain-invariant encoder, the HistoStarGAN model can, for the first time, generalise stain transfer to unseen stainings.

5.2.1 Diverse Multi-Domain Stain Transfer

A trained model is able to perform diverse stain transfer between any pair of stainings seen during training. The diverse transfer is obtained by sampling different random codes, which are transformed by the mapping network into target-stain specific styles. Some examples of PAS image translations, alongside corresponding segmentations, are provided in Figure 5.2. The obtained translations are plausible histopathological images, where the structures of interest, glomeruli in this case, are preserved during translation. The differences between translations are at the level of microscopic structures (e.g. the appearance of nuclei, the thickness of a membrane, etc.), see Figure 5.3. Since the HistoStarGAN is multi-domain, the same model is

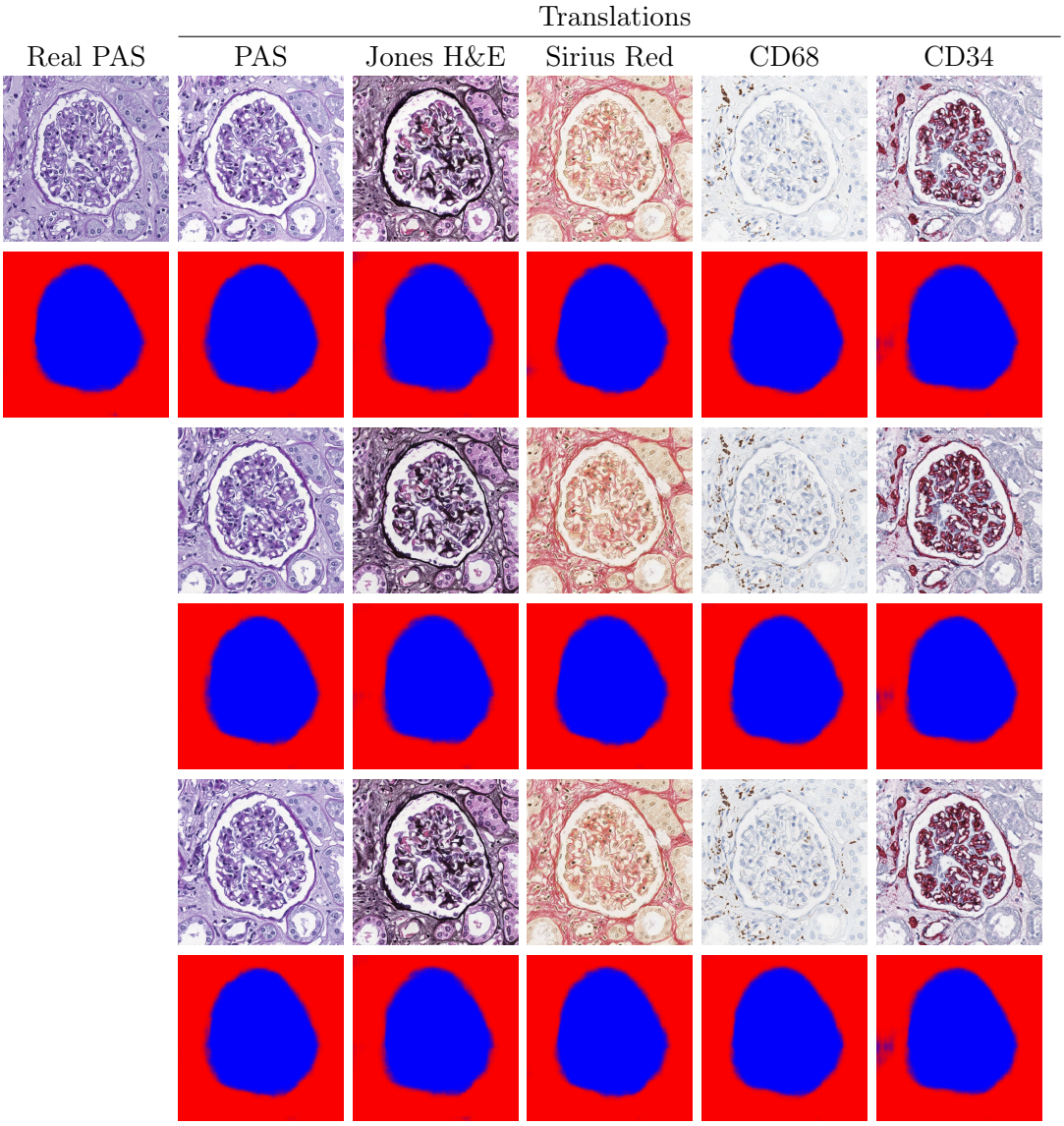


Figure 5.2: Diverse HistoStarGAN translations of a PAS glomeruli patch to target stains (including PAS) with corresponding segmentations. Fake images in each row are generated using the same random vector transformed into a stain-specific style by the Mapping network. The differences between translations are in microscopic structures (e.g. membrane weight or nucleus appearing), which is barely visible in these figures. Full resolution images, in which these differences are more visible, are available online.

also able to perform translations between other staining pairs, examples of which are provided in Figure 5.4.

Generalisation of stain transfer: Since HistoStarGAN is trained on a variety of stains, the generator's encoder is stain invariant, which for the first time enables virtual staining of unseen stains. Examples of which are presented in Figure 5.5,

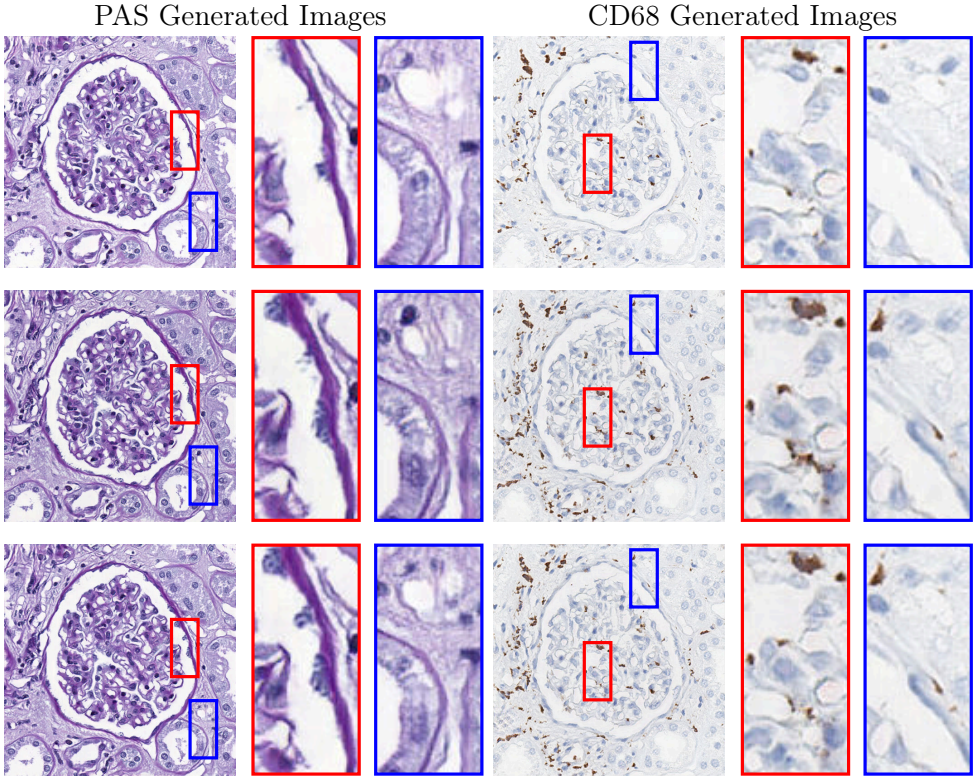


Figure 5.3: Closer look to the differences between the HistoStarGAN translations.

in which a new stain modality named H&E in addition to three double-stainings CD3-CD68, CD3-CD163 and CD3-CD206 are translated to stainings seen during training. Moreover, in Figure 5.6 the model is applied to the AIDPATH dataset [18] composed of images which are publicly available variations of the PAS stain. This demonstrates that HistoStarGAN can generalise and obtain stain normalisation (column PAS) and stain transfer (other columns) simultaneously, alongside stain-invariant segmentation. Videos representing the exploration of the latent space during translation, are provided online[10]. Additional results are provided in the Appendix D.

5.2.2 Stain Invariant Segmentation

A model composed of the generator's encoder and the segmentation branch can perform stain-invariant segmentation of WSIs across various stainings. Table 5.1 presents the segmentation results for test WSIs from all stainings (virtually) seen during training. The model's performance is compared to UDA-GAN, which uses the same CycleGAN models for data augmentation. Since the patch size is 512×512, each patch is cropped to 508×508 during UDA-GAN training. The presented results are the averages of three independent training repetitions with corresponding standard deviations.

[10] https://main.d33ezaxrmu3m4a.amplifyapp.com/

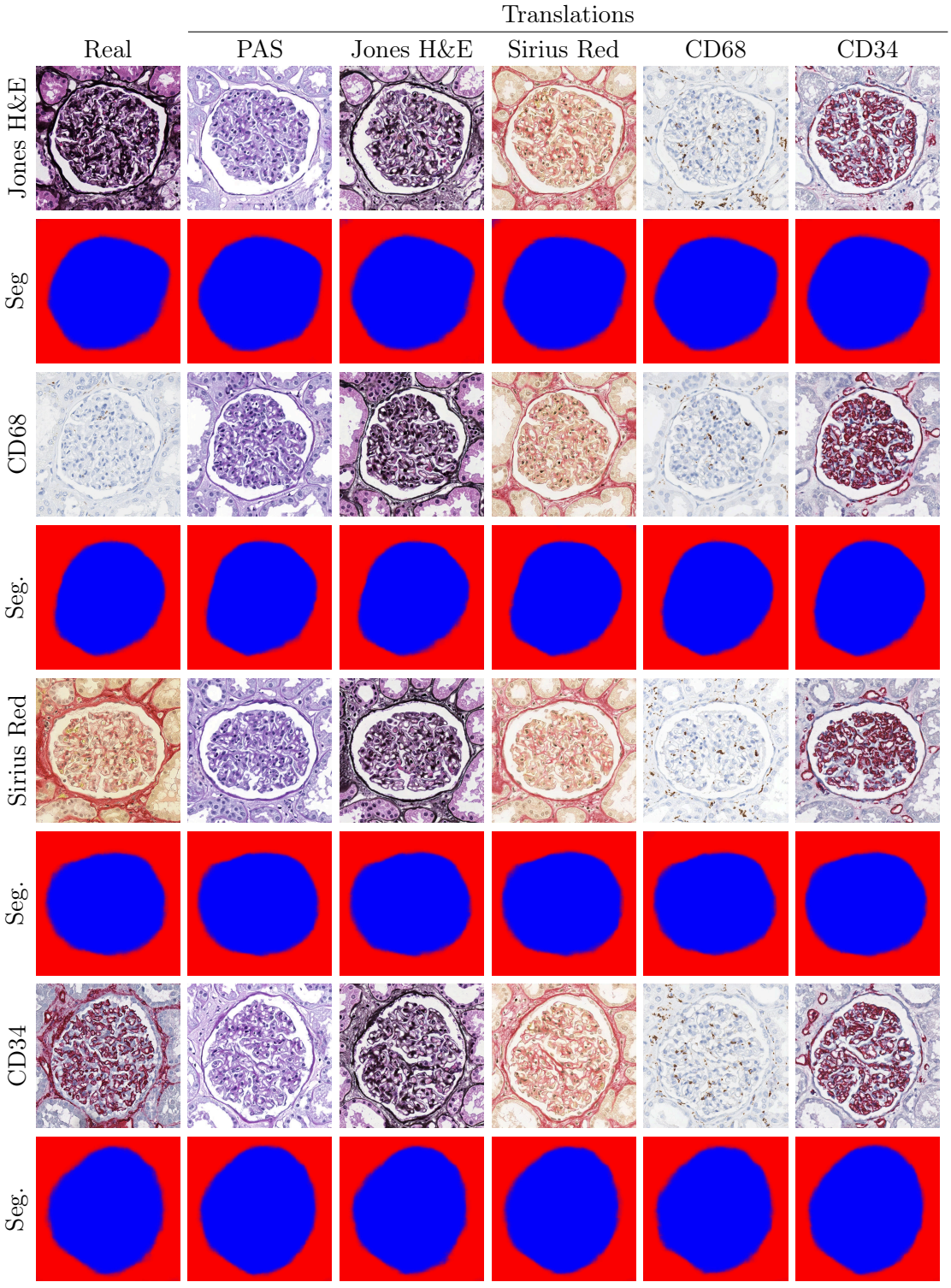


Figure 5.4: HistoStarGAN translations between different stains with corresponding segmentation.

Table 5.2 Quantitative results for HistoStarGAN compared to UDA-GAN on unseen stains. Each model is trained on annotated PAS (source staining). Standard deviations are in parentheses, and the highest F_1 scores for each staining are in bold.

Model	Score	Test Staining						
		H&E	CD3	CD3-CD68	CD3-CD163	CD3-CD206	CD3-MS4A4A	Overall
UDA-GAN	F_1	0.681 (0.031)	0.648 (0.111)	0.258 (0.062	0.260 (0.050)	0.330 (0.058)	0.330 (0.071)	0.418 (0.194)
	Precision	0.865 (0.079)	0.550 (0.161)	0.171 (0.047)	0.168 (0.039)	0.230 (0.054)	0.240 (0.070)	0.371 (0.281)
	Recall	0.563 (0.028)	0.824 (0.032)	0.538 (0.070)	0.586 (0.039)	0.598 (0.029)	0.542 (0.029)	0.608 (0.108)
HistoStar-GAN	F_1	**0.813** (0.022)	**0.741** (0.009)	**0.597** (0.011)	**0.611** (0.015)	**0.593** (0.014)	**0.570** (0.012)	**0.654** (0.099)
	Precision	0.855 (0.018)	0.850 (0.007)	0.835 (0.022)	0.891 (0.021)	0.881 (0.026)	0.882 (0.025)	0.866 (0.022)
	Recall	0.777 (0.056)	0.656 (0.011)	0.465 (0.017)	0.465 (0.021)	0.447 (0.020)	0.422 (0.016)	0.539 (0.144)

The HistoStarGAN can generalise across all virtually seen stainings during training, outperforming UDA-GAN trained using the same translation models. Apart from being an end-to-end model that simultaneously performs virtual staining and segmentation, HistoStarGAN also results in an increase in precision. This can be attributed to the fact that HistoStarGAN better recognises the negative tissue (less false positives). However, recall is lower, which indicates that more glomeruli (or parts of them) are missed compared to UDA-GAN. This could be the consequence of predicting a segmentation mask from the feature space directly, without skip-connections. However, extending the HistoStarGAN model with skip-connections between the encoder and segmentation branch experimentally showed to negatively affect training stability.

Generalisation of stain invariance: To test the generalisation of HistoStarGAN, the model is applied to new stainings not seen during training. Using the annotations of four whole-slide images from each of the new stainings (H&E, CD3, CD3-CD68, CD3-CD163 and CD3-CD206), the averages of three independent training repetitions with corresponding standard deviations are presented in Table 5.2. Overall, HistoStarGAN generalises better and is more robust compared to UDA-GAN. A potential cause of UDA-GAN failure in some stains can be its learning process where stain invariance is forced only on the pixel level. Although the HistoStarGAN uses the identical CyleGAN translations, the framework extracts stain-invariant features, which are better transferred to unseen stains. Moreover, the segmentation branch of HistoStarGAN is trained on a dataset with more variety since HistoStarGAN translations are also included (see Figure 5.1).

5.3 Ablation Studies

HistoStarGAN model builds upon StarGANv2 in several aspects — the first extension is attaching a segmentation module to the generator; the second is using

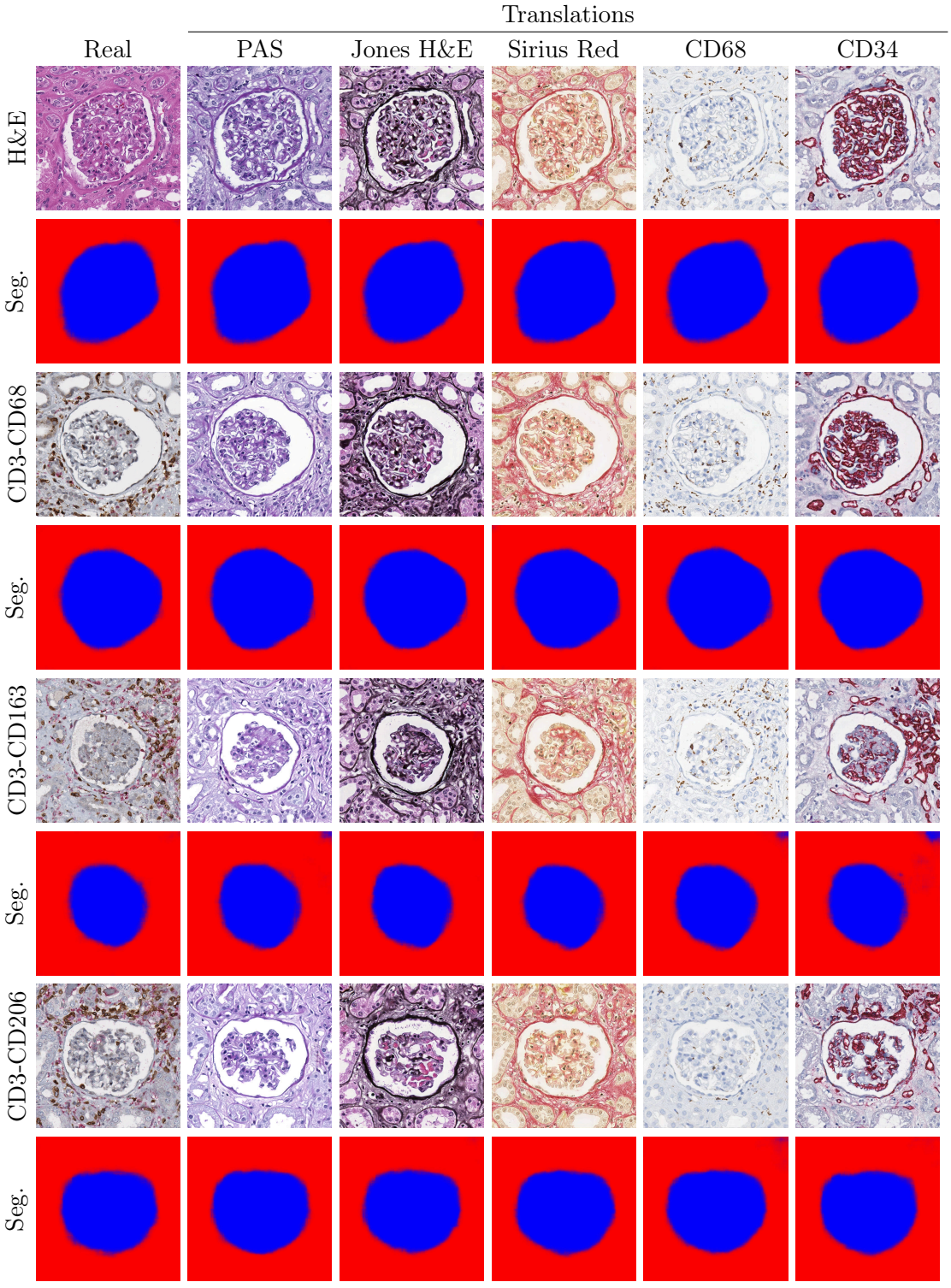


Figure 5.5: HistoStarGAN — a generalisation of stain transfer and segmentation to unseen stain modalities.

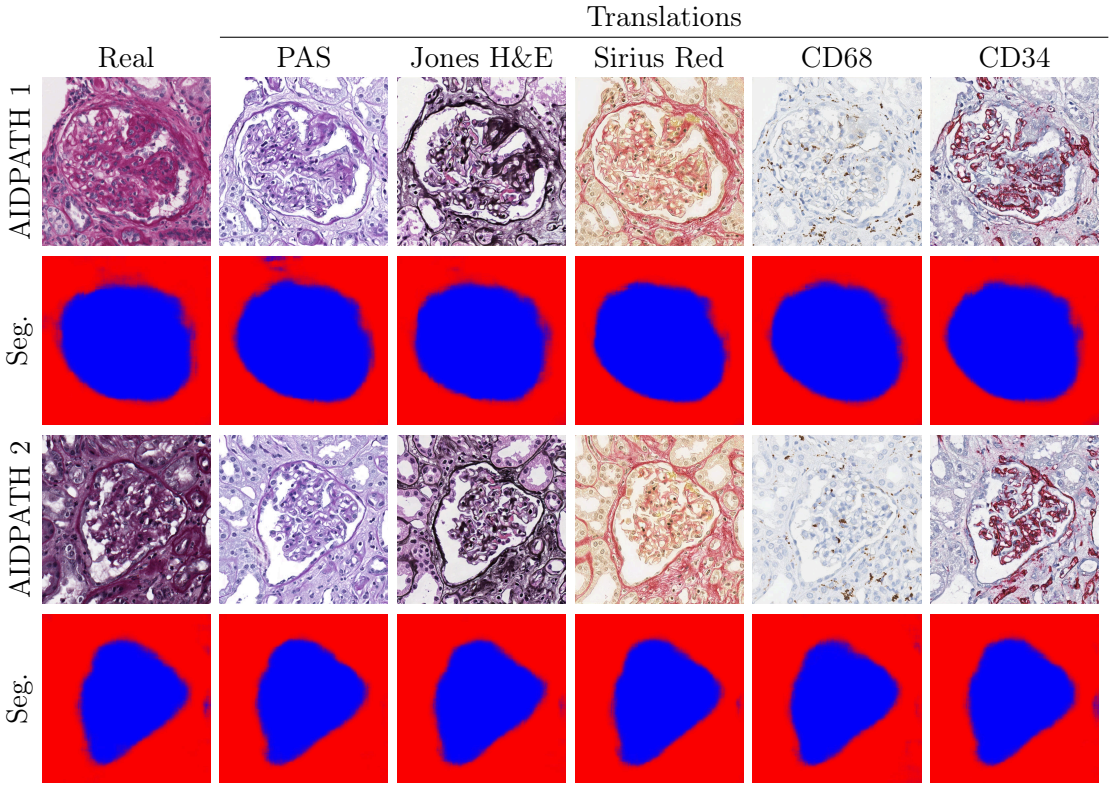


Figure 5.6: HistoStarGAN applied for stain normalisation, stain transfer and glomeruli segmentation of the publicly available AIDPATH (PAS-based) dataset.

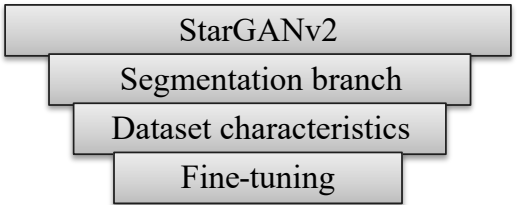


Figure 5.7: Illustration of the differences between the StarGANv2 and HistoStarGAN. Each of these differences is justified via an ablation study.

pre-trained CycleGANs to create the training dataset, and the third is fine-tuning the segmentation module using only source data. In the following, each of these aspects will be discussed via an ablation study, in a bottom-to-top direction as illustrated in Figure 5.7.

5.3.1 Fine-Tuning

Table 5.3 demonstrates that fine-tuning the segmentation branch using the fixed generator's encoder increases segmentation performance. In practice, fine-tuning for just a single epoch on an imbalanced source dataset increases overall performance on virtually seen stains (test patient WSIs) by around 6% in the overall F_1-score.

Table 5.3 Fine-tuning effects on the segmentation model's performance, in which the model is fine-tuned for one epoch.

Model	Score	Test Staining					
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
HistoStarGAN	F_1	0.820	0.816	0.705	0.792	0.777	0.782
	Precision	0.779	0.791	0.764	0.792	0.802	0.761
	Recall	0.865	0.843	0.655	0.782	0.795	0.779
HistoStarGAN (fine-tuned)	F_1	**0.876**	**0.875**	**0.762**	**0.855**	**0.842**	**0.842**
	Precision	**0.851**	**0.871**	**0.861**	**0.886**	**0.842**	**0.862**
	Recall	**0.902**	**0.879**	**0.683**	**0.825**	**0.843**	**0.826**

However, longer fine-tuning, although potentially beneficial for particular stainings, does not offer any benefits. Also, the performance on the source's validation set does not correlate with the obtained improvements, i.e. the fine-tuned model with the lowest validation loss does not bring the best overall results on unseen stains. Thus, fine-tuning for one epoch is chosen.

5.3.2 Dataset Characteristics

A balanced and (virtually) fully annotated dataset is used to train HistoStarGAN. A labelled dataset is required since the segmentation branch is trained in a supervised manner, and thus alternative data sampling strategies such as random data sampling [19, 64] are not suitable. However, the fully annotated dataset does not need to be balanced. Thus, using the findings of Lampert et al. [17], an imbalanced source (PAS) dataset is formed where all glomeruli in the training patient images are extracted (662) and seven times as many negative patches (4634). This dataset is translated using CycleGAN to all target stains (CycleGAN models are always trained in an unsupervised way, on random patches). Thus, a fully annotated and highly imbalanced dataset is created on top of which the HistoStarGAN model is trained.

The results obtained in this setting are presented in Figure 5.8, in which one glomeruli patch from the PAS stain is translated into multiple stainings using different latent codes. The obtained translations for one stain pair differ from each other in terms of glomeruli texture and surrounding structures. The model is able to change the size of glomeruli by changing the size of the Bowman's capsule (white space), in addition to varying the appearance of stain-specific markers (i.e. macrophages in CD68 stain). Nevertheless, the segmentation branch successfully recognises all glomeruli variations. Compared to training with a balanced dataset, this setting offers more translation variability related to the surrounding tissue and structure of the glomeruli themselves. However, although globally these translations look realistic, the internal structures inside the glomeruli are not. Since in HistoStarGAN these translations must also be successfully segmented, this leads to an increase in the false positive rate. For example, it is evident that the produced translations often contain 'artificial looking patterns' such as a tendency to group microscopic structures (e.g. nuclei) into diagonal, horizontal or vertical stripes

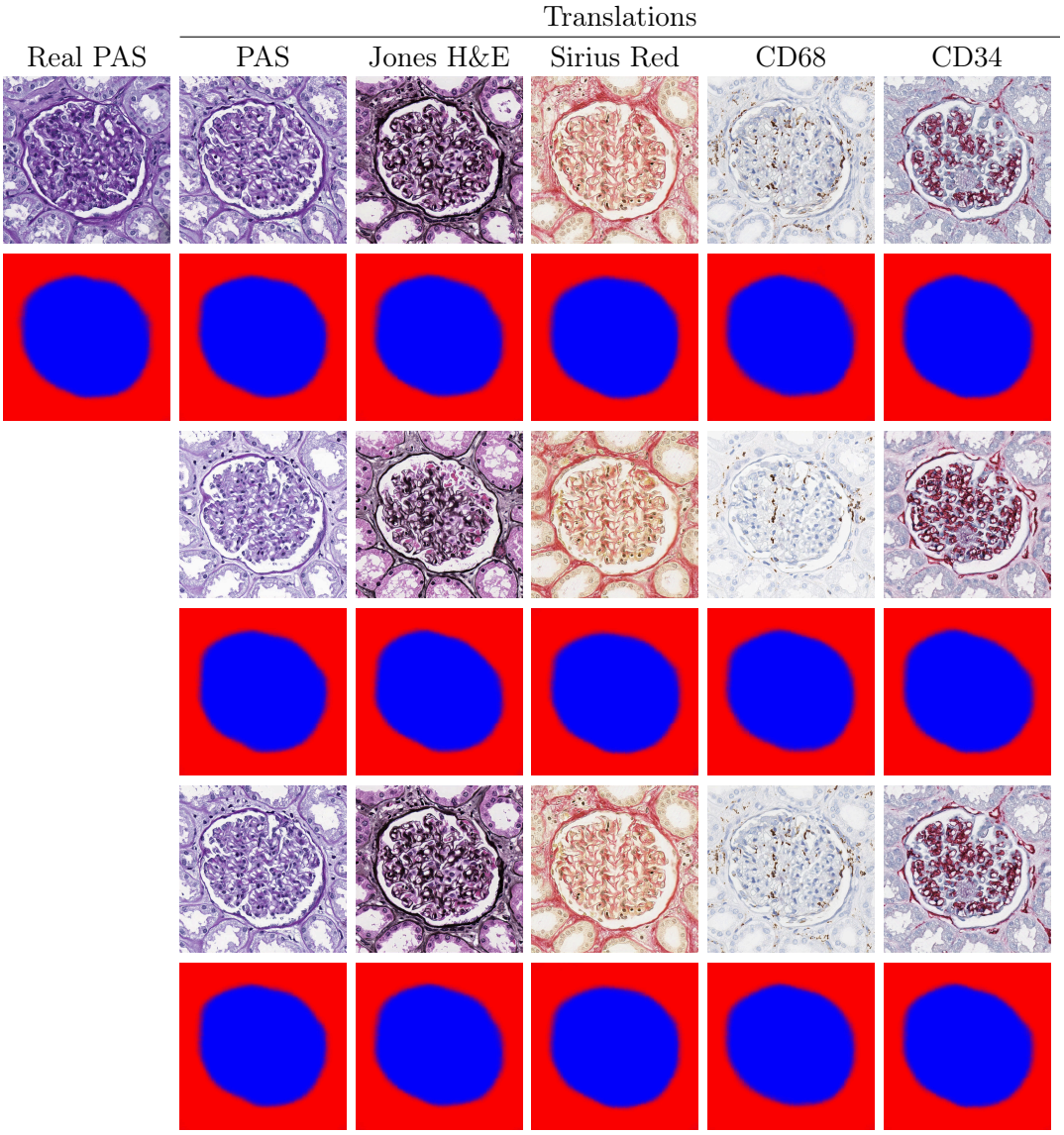


Figure 5.8: HistoStarGAN translations using a model trained using a CycleGAN-generated imbalanced annotated dataset. The images in each column are generated using the same latent codes.

(in Figure 5.8 this is mostly visible in translations to Sirius Red and Jones H&E). Therefore, the proposed model uses a balanced dataset since it greatly reduces such possibilities.

5.3.3 Segmentation Branch

Without the segmentation branch, HistoStarGAN is reduced to StarGANv2 (see Figure 5.7). Moreover, removing the segmentation module removes the requirement for the dataset to be fully annotated. Thus, such a model can be trained using a

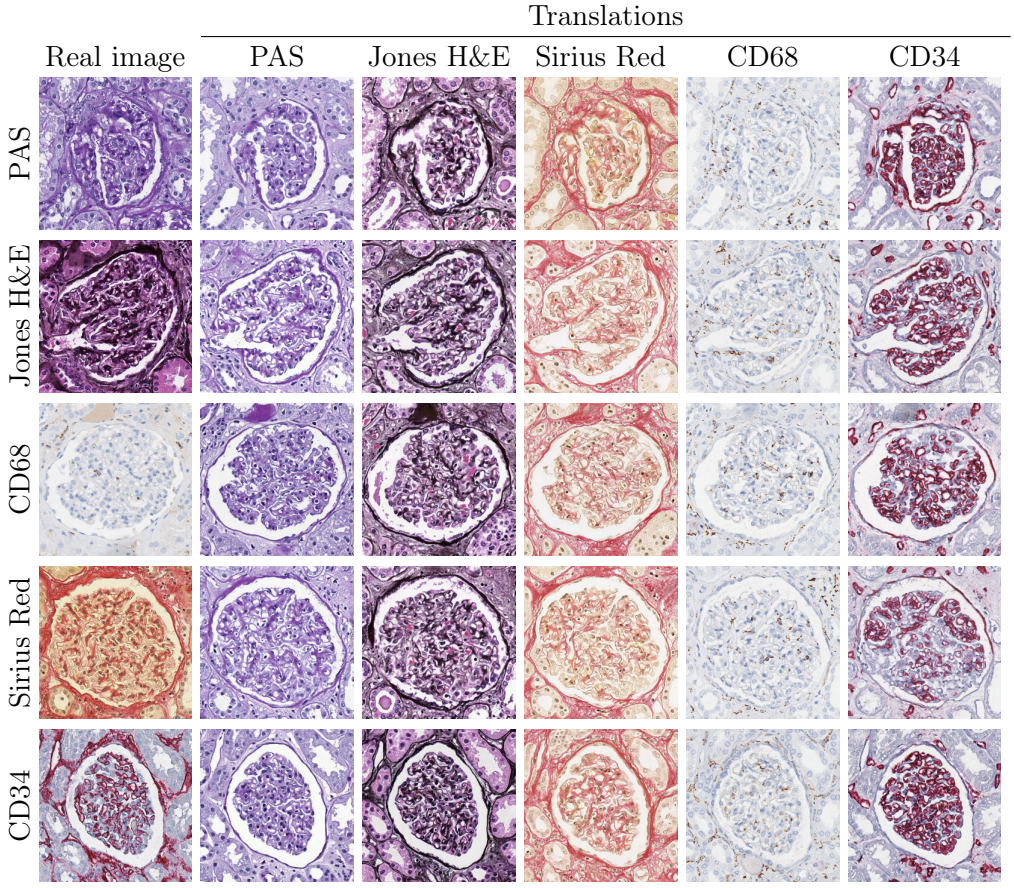


Figure 5.9: StarGANv2 trained on CycleGAN-generated translations of a balanced PAS dataset to other stains.

dataset composed of random patches (uniformly sampled in each stain), as well as created by CycleGAN translations of the annotated stain (PAS), both balanced and imbalanced. Each of these will be separately analysed.

CycleGAN-Based Balanced Dataset A StarGANv2 model trained on a balanced dataset produced by CycleGAN translations of the annotated stain, can also result in diverse multi-domain translations between multiple stainings, as illustrated in Figure 5.9. Usually, the glomeruli structures are visually preserved. However, when measuring the quality of obtained translation using the performances of pre-trained segmentation models for each staining (baseline models trained in a fully-supervised setting), the conclusion can be different. Figure 5.10 presents a visual comparison of HistoStarGAN and StarGANv2, where translations obtained by StarGANv2 are segmented using pre-trained models from each staining. Since both models can obtain diverse translations, Figure 5.10 represent the average translation and average segmentation over 50 random latent codes. It can be seen that HistoStarGAN gives more accurate segmentations, especially in difficult cases, e.g. sclerotic glomeruli, rows 5 and 7. Nevertheless, in the absence of a segmentation branch, the

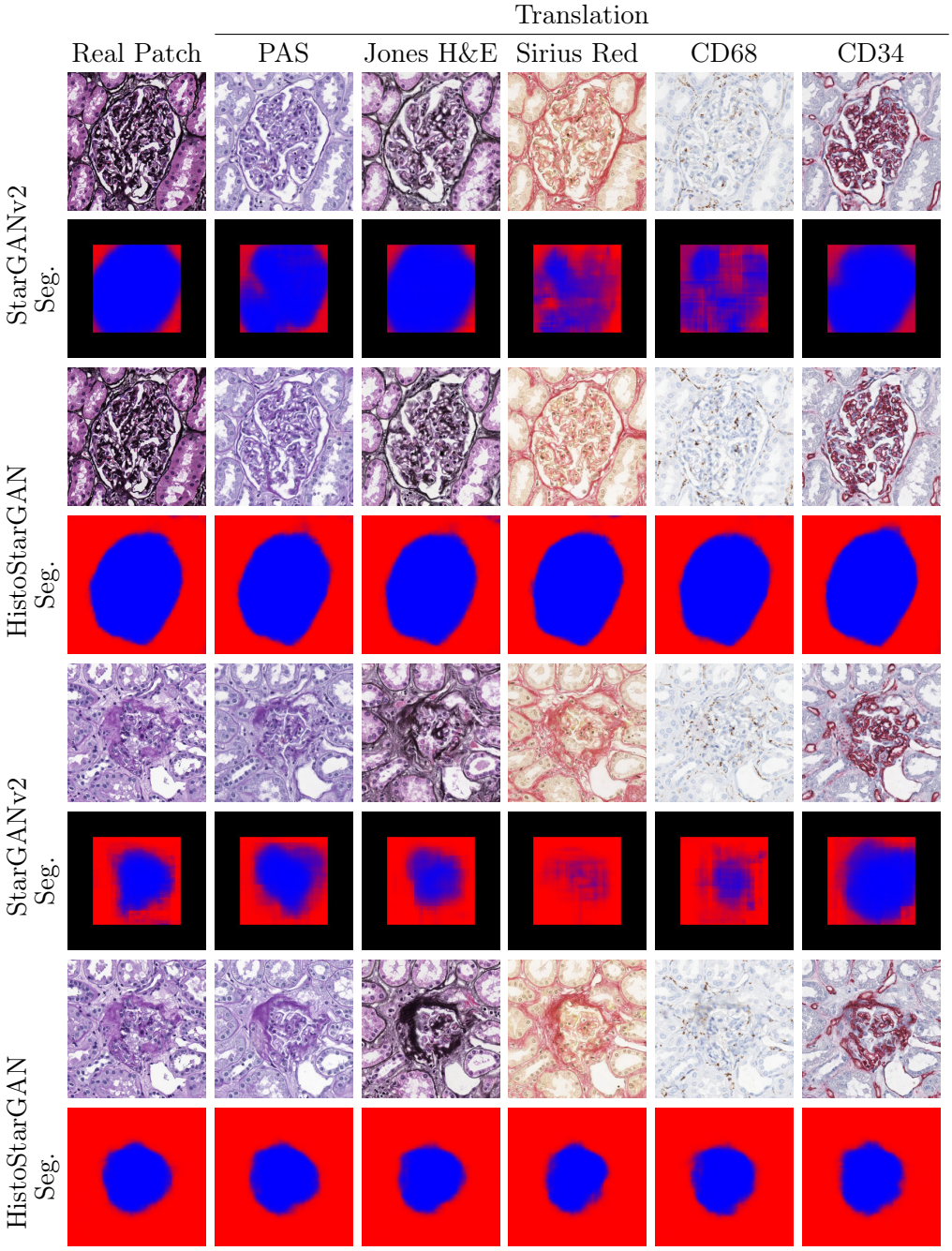


Figure 5.10: StarGANv2 trained on a balanced CycleGAN-generated dataset compared to HistoStarGAN. The segmentations for StarGANv2 are obtained by stain-specific, pre-trained baseline models. Each translation and segmentation is averaged over 50 random latent codes. N.B: HistoStarGAN yields more accurate segmentations. Pre-trained models result in 324×324 pixel images which are placed in the centre of 512×512 pixel black squares.

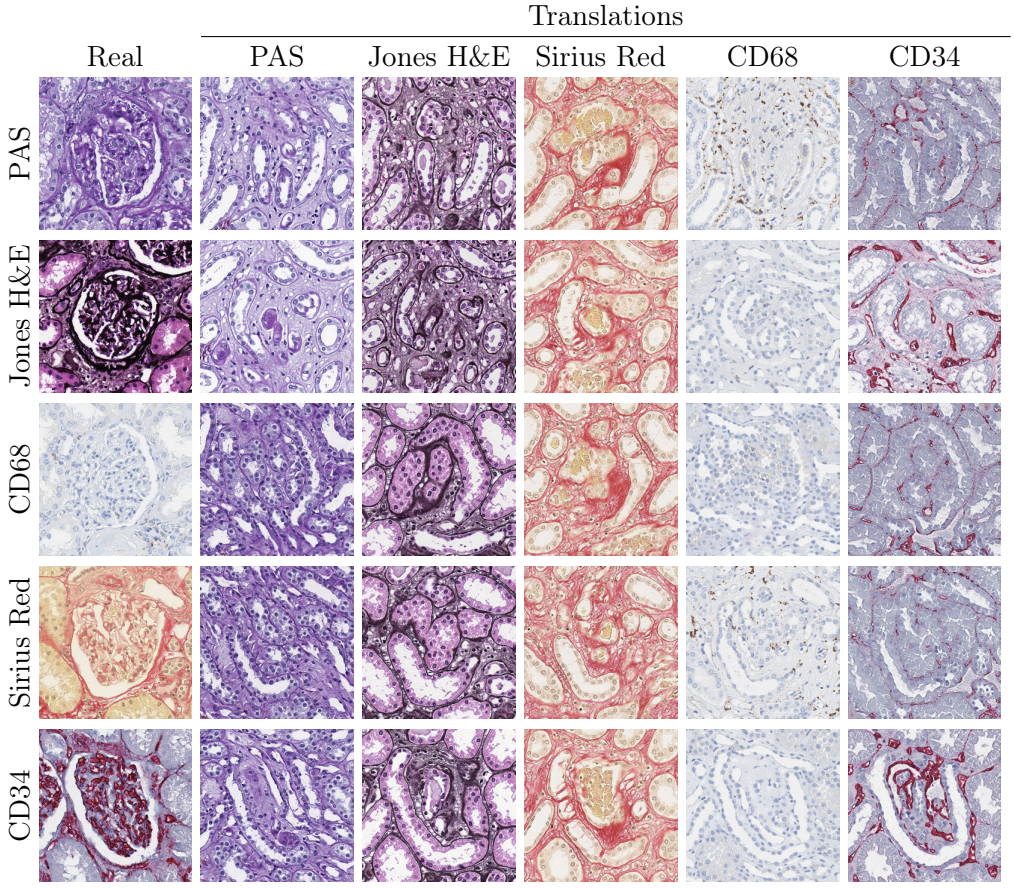


Figure 5.11: StarGANv2 model trained on CycleGAN-generated translations of an imbalanced PAS dataset to other stains.

dataset composition itself cannot be a strong guarantee that glomeruli structures will be preserved. Thus, an explicit requirement, such as attaching the proposed segmentation branch, is beneficial to ensure the correctness of the translation and robust segmentation.

CycleGAN-Based Imbalanced Dataset If the dataset used to train the StarGANv2 is imbalanced, the model can no longer preserve the structure of interest. Some examples of stain transfers obtained with this model are presented in Figure 5.11. Since the HistoStarGAN model trained using the same imbalanced dataset does not have such a problem, this demonstrates the significance of the proposed segmentation branch.

Random Dataset: StarGANv2 can also be trained on a dataset composed of random patches extracted from each stain (PAS, Jones H&E, Sirius Red, CD68 and CD34). Thus, 5000 random patches from the training patients are extracted uniformly, and the total dataset containing 25 000 images is used to train the StarGANv2 model. As in the CycleGAN-generated imbalanced data setting, this model

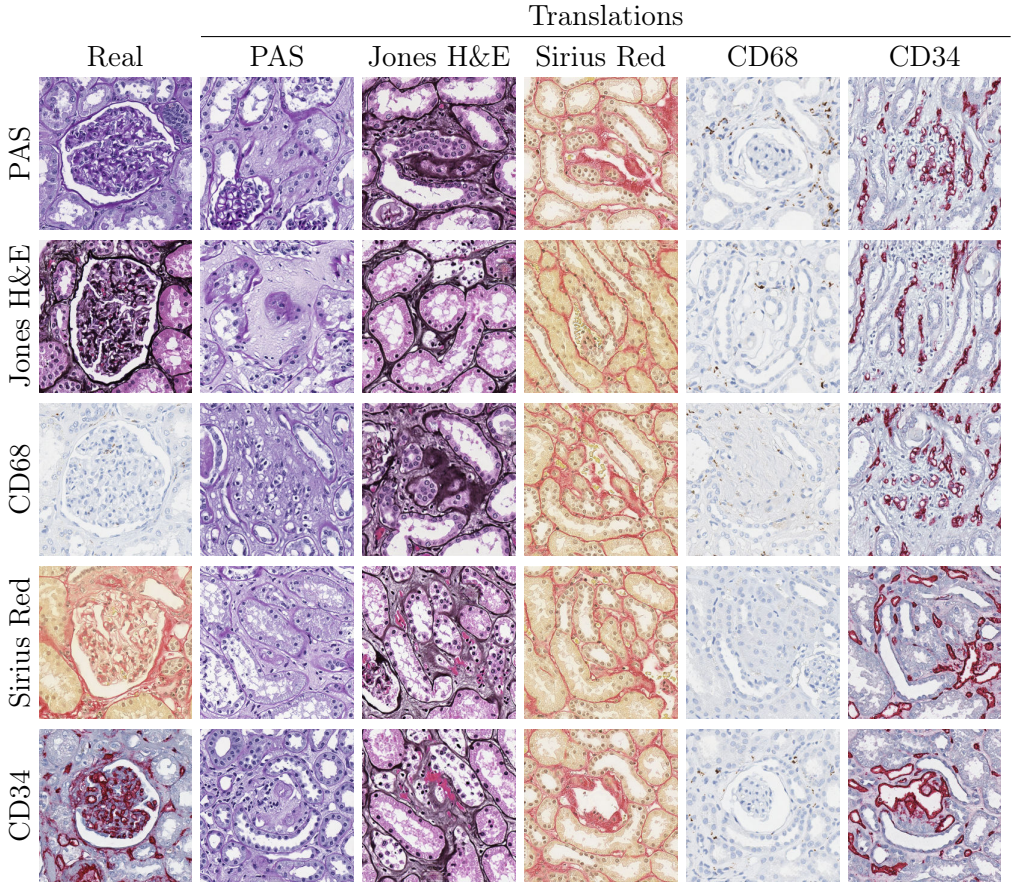


Figure 5.12: StarGANv2 model trained on random patches extracted from all stains, applied to translated glomeruli patches.

cannot preserve the structure of interest, as demonstrated in Figure 5.12. Moreover, it is prone to bigger alterations of microscopic structures, which limits the usefulness of such translations and confirms the benefits of the proposed HistoStarGAN model.

5.4 KidneyArtPathology Dataset

This Section describes the KidneyArtPathology dataset — artificially created, fully annotated dataset released to encourage the progress of deep learning-based solutions in the field of renal pathology[12]. The dataset contains 5000 images from five stainings, in a resolution of 512×512 pixels, fully annotated for the task of glomeruli segmentation. To achieve this, a HistoStarGAN model is trained using as representative as possible dataset — two immunohistochemical stainings (CD68

[11]Image credits in order of appearance: Strasbourg Cathedral: https://www.visitstrasbourg.fr/en/discover/must-see-attractions/the-cathedral/, Kragujevac Museum: https://www.topworldtraveling.com/articles/travel-guides/15-best-things-to-do-in-kragujevac-serbia.html, Dog: https://github.com/fastai/imagenette.

[12]The dataset will be released upon selection performed by pathologists.

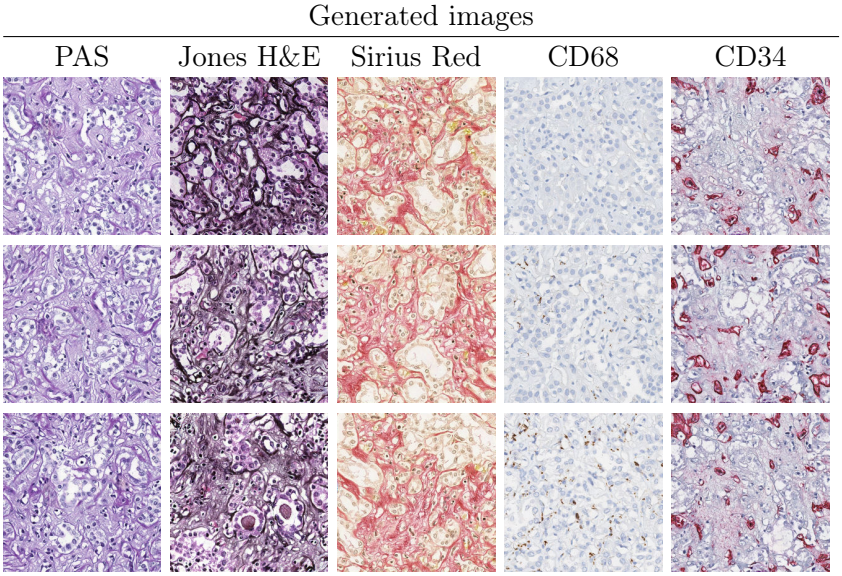


Figure 5.13: The HistoStarGAN model generates plausible histopathological images from random noise.

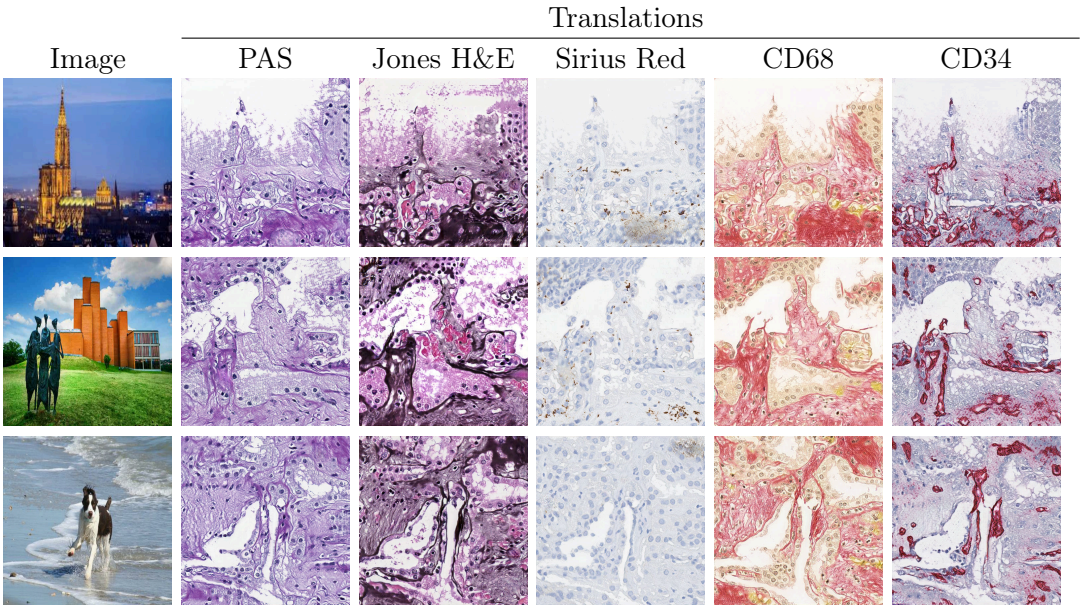


Figure 5.14: KidneyArtPathology — histopathological image generation from natural images[11].

and CD34) in addition to three histochemical stainings (PAS, Jones H&E and Sirius Red). Thus, the model can generate various translations in multiple stainings, which has allowed the creation of the KidneyArtPathology dataset [13]. The dataset

[13]KidneyArtPathology has been released online, https://main.d33ezaxrmu3m4a.amplifyapp.com/

is composed of HistoStarGAN translations of PAS stain images (the same as those used to generate the annotated dataset used for HistoStarGAN training) which are translated using ten random latent vectors into each staining, including PAS. Moreover, the associated pre-trained HistoStarGAN model (also publicly available) can be further used to augment private datasets with their annotated translations to stainings used during HistStarGAN's training.

There are several benefits of such a dataset, which fall under three categories:

- Pathological — Non-invasive Pathology Training, the diverse appearance of glomeruli can be helpful in the early stages of a pathologist's training.
- Benchmarks — the absence of publicly available real-world datasets poses huge challenges for rigorous comparison in the literature. Thus, such a large collection of annotated patches can serve as a benchmark.
- Domain adaptation — The data can be used in addition to a private dataset (which can contain limited data) to build more robust models, e.g. as an augmentation or domain adaptation strategy.

New histopathological images: Since the style of a stain is encoded by the mapping network, it is possible to generate new patches in each training staining by providing a random image, rather than a source histopathological patch, to be translated to a given stain. In addition to a fully-annotated glomeruli dataset, new histopathological images in different stains can be generated. Some generated examples are provided in Figure 5.13. Alternatively, by providing a non-histopathological image, the HistoStarGAN is able to convert it to a histopathological image, examples of which are provided in Figure 5.14, and additional are provided in Appendix D (although medical use of this is admittedly most likely limited to non-existent, it is an interesting side property of HistoStarGAN).

5.5 Limitations and Opportunities

HistoStarGAN is an end-to-end model that can segment and generate diverse plausible histopathological images alongside their segmentation masks. However, for some latent codes the model can produce specific artefacts in the translations. These are usually well-incorporated into the overall texture of the image, which makes them not obvious at first glance. A closer look at one such example is given in Figure 5.15. The primary hypothesis is that the discriminator does not have enough capacity to spot these artefacts, and thus a more sophisticated discriminator could be considered.

Furthermore, the performance of the HistoStarGAN model can be affected by choice of source and target stainings and the quality of CycleGAN translations, which remains to be explored in future work. HistoStarGAN is based on CycleGAN translations, which are proven to be noisy (Chapter 3), and therefore some of these invisible artefacts may be recreated by HistoStarGAN. Thus, it is unclear what information from the source domain is preserved in translations and whether any of the applied augmentations can perturb them. It is possible that incorporating additional real examples from the target stainings can improve translation quality.

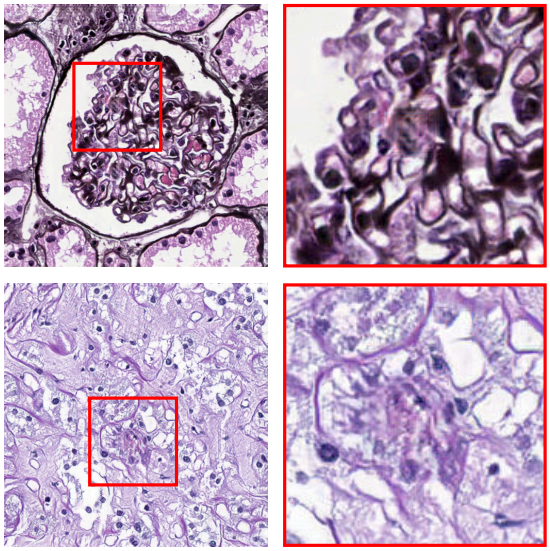


Figure 5.15: HistoStarGAN — common artefacts. N.B. the artefacts are well-incorporated into the image texture and not obviously noticeable at first glance.

Additionally, it has been confirmed experimentally that the choice of loss functions used for the diversity loss can significantly affect the HistoStarGAN translations. For example, the model will be forced to perform structural changes if diversity loss is set to maximise the structural similarity between images (this is illustrated on the linked project's website). Although from a style-transfer perspective, such a solution is interesting, changing internal structures in this way is not biologically justifiable. However, according to a recent study [141], medically-irrelevant augmentation can increase model robustness; thus, this represents another exciting direction for the work ahead.

Nevertheless, the model has great potential to be used for tackling the lack of annotations in histopathological datasets. Depending on the availability of labels in a given target domain, diverse HistoStarGAN translations with corresponding segmentation masks can help training from unsupervised to fully supervised settings, using one or more source (annotated) domains.

Finally, it is crucial to note that translations obtained by the HistoStarGAN model should not be used for diagnostic purposes. All images are artificially generated, and the model can perturb diagnostic markers. Thus, the dataset composed of HistoStarGAN translations should only be used for general-purpose analysis related to glomeruli (e.g. counting).

5.6 Conclusions

The HistoStarGAN model represents the first end-to-end trainable solution for simultaneous stain normalisation, stain transfer and stain invariant segmentation. For the first time, obtaining highly plausible stain transfer from unseen stainings is

possible without any additional change to the model (e.g. fine-tuning). Moreover, the model achieves new state-of-the-art results in stain invariant segmentation, successfully generalising to six unseen stainings. The proposed solution is general and extendible to new stainings or use cases.

Being able to generate diverse translations for a given input, the proposed solution has allowed the generation of the first artificially created, fully-annotated dataset — KidneyArtPathology. Furthermore, the model is trained on five widely used stainings, and pre-trained models are available, enabling offline data augmentation (e.g. on private datasets) by stain transfer to these stainings.

However, the obtained virtual staining can realistically change the appearance of microscopic structures, such as the number of visible nuclei, size of Bowman's capsule or membrane thickness. Thus, it is ill-advised to use these translations for applications that rely on the assumption that the translation process preserves all structures in the original image. Moreover, such virtually stained images must not be used for diagnostic purposes since diagnostic markers can be altered. As such, this study goes towards saving resources and time for annotating the high-quality datasets required for developing deep learning-based models, not to replace the process of physical staining.

Nevertheless, it remains an open question why the model's performance varies across different stainings. One hypothesis is that the model, being based on CycleGAN translations, replicates some of their limitations. However, the influence of these CycleGAN models and the choice of the stainings used in the dataset remains to be explored in future work.

Chapter 6

Conclusions and Perspectives

Digital histopathology is a promising research area where many daily tasks, such as segmentation or classification, have the potential to be facilitated by deep-learning-based solutions. However, many state-of-the-art approaches are data hungry, requiring huge collections of annotated data to perform well. Additionally, considering the variations that can occur due to the staining process and staining protocols, already collected and annotated datasets can only be reused with limited success. This thesis has investigated the potential of Generative Adversarial Networks (GANs) in two directions for addressing these problems — stain transfer to enable reusing already available data collections; and developing stain invariant solutions which would alleviate the need for additional data acquisition. The application focus was glomeruli segmentation in renal pathology with multiple stainings, all of which have been annotated by trained experts for evaluation purposes.

The work presented in this thesis was preceded by an extensive literature review (Chapter 2) of existing GAN-based approaches that tackle stain variation and stain invariance. That led to the identification of several shortfalls in current methods and motivated solutions to these issues. Particularly, the significant lack of approaches that tackle stain transfer between multiple stainings and the lack of stain invariant solutions was recognised, which is resolved by this thesis.

Chapter 3 proposed, in parallel with Gadermayr et al. [19], the use of CycleGAN-based approaches to obtain plausible stain transfer between multiple stainings. It has been demonstrated that such methods can significantly reduce domain shift caused by inter-stain variation, enabling models trained for one stain modality to be applied on another without additional annotations. Although CycleGAN-based approaches have been quickly accepted by the community and became the standard for stain transfer (and more commonly for stain normalisation), the limitations of these methods are rarely addressed. Chapter 3 of this thesis provided a critical look at these methods, raising the importance of such limitations through extensive experiments. A significant conclusion from this analysis is that the visual inspection as an evaluation criterion, widely adopted in practice, is ill-advised, primarily if the evaluation is performed by non-experts (e.g. non-pathologists). Trained experts can spot some failure modes; therefore, experts-in-loop are very important. However, inspection solely based on visual evaluation is not recommended as conclusions can be misleading due to human-imperceptible artefacts that the stain transfer model

can inject. The presence of imperceptible artefacts has been related to stain-specific characteristics. This finding was additionally exploited to propose an unsupervised augmentation method that increases deep learning model robustness in a supervised setting. Thus, imperfections of the stain transfer model can be used to build better deep-learning-based solutions. Nevertheless, the artefacts present in the translations can affect the pre-trained model's performance unpredictably, making such quantitative evaluations unreliable. These findings raise the importance of developing objective evaluation criteria for artificially generated medical data.

Chapter 4 introduced Unsupervised Domain Augmentation using Generative Adversarial Networks — UDA-GAN — the first solution that encourages empirical stain invariance for the task of glomeruli segmentation. That was achieved by taking advantage of the plausibility of the translations obtained by the proposed stain transfer approach. It has been shown that the obtained model can segment across multiple stains and generalises to some unseen stains. Such results confirm that obtaining a stain invariant solution for glomeruli segmentation is feasible and can be achieved using a limited set of annotations. However, the model's performance can vary depending on the staining characteristics. Moreover, UDA-GAN is more stable and achieves better quantitative results than competitive, stain-specific solutions, indicating that stain transfer as a domain shift reduction strategy could have greater potential as a part of the solution than the solution itself. This was also confirmed by feature-space adaptation experiments, where a combination of stain transfer and feature-space adaptation outperforms both strategies individually.

In Chapter 5, the findings of the previous chapters are brought together to propose HistoStarGAN, the first end-to-end trainable model that simultaneously performs stain transfer, stain normalisation and stain invariant segmentation. Such a model better generalises to unseen stainings than the previously proposed UDA-GAN approach (Chapter 4). Moreover, for the first time, it is possible to translate unseen stains to the set of chosen staining modalities in addition to stain invariant glomeruli segmentation. Such ability has been used to generate KidneyArtPathology, the first artificially created, fully annotated kidney pathology dataset, released to encourage the progress of deep learning-based solutions in the field of renal pathology. Such a data collection, in addition to a pre-trained model, offers several benefits to the community, such as non-invasive pathology training, benchmarking and the development of better models on small and private datasets.

Nevertheless, based on the findings presented in this thesis, the use of artificially created datasets in the medical domain should be carefully regulated. Their clinical use should be limited given the current state-of-the-art, as these methods can perturb diagnostically relevant information. From the deep learning perspective, it is evident that such approaches can enhance model learning and increase a model's robustness. However, short-cut learning is still likely to happen, making such models sensitive to unexpected and sometimes imperceptible data variations. Thus, solutions that can learn from the limited available data while generalising to unannotated unseen domains, such as solutions proposed in Chapter 4 and Chapter 5 of this thesis, should gain more focus in the literature.

This thesis also confirms that experts in the loop are significant from the perspective of developing automated solutions. Their understanding and acceptance of automatic solutions are essential to clinical application. Finding a reasonably high

number of experts willing to participate in studies is still difficult. Resistance to automatic solutions is not neglectable from both the patient and expert viewpoint. In order to reduce the gap, it is essential to include experts from the earliest stages of development. Understanding the potential and, more importantly, the limitations and sources of errors of automated solutions are the first steps toward making clinically appropriate and widely-acceptable solutions.

6.1 Perspectives

This section discusses possible future research directions that arise from the work presented in this thesis. Some of the directions are related to the methods proposed in this thesis, whilst others are related to general problems in the field of digital histopathology.

6.1.1 Synthetic Medical Datasets

The computer vision community has a noticeable trend to replace real datasets with synthetic [194–196]. That way, multiple issues, such as data privacy and the laborious annotation processes, can be alleviated to a great extent. There are similar attempts in the medical field [197–199] where this thesis has been contributed by the KidneyArtPathology dataset. However, the potential and use of artificially generated datasets in the medical domain remain to be explored.

The medical data generators, such as herein proposed HistoStarGAN, require some portion of real data collected from hospitals. These data typically fall under strict legal regulations. Thus, it is of crucial importance to carefully examine whether artificially created medical data contains any sensitive patient-related information [200]. Moreover, considering that a neural network can unintentionally memorise part of the training dataset [201] and the possibility of model inversion [202], the potential to deduce sensitive information from pre-trained generative models should be investigated.

Additionally, the representativeness of the synthetic data should be considered. It is possible that artificially created data contain imperceptible artefacts, which can affect the generalisation properties of models that learn from such data. Moreover, similar to the real datasets, synthetic data can be biased towards a specific patient group (e.g. gender), which could result in unintentionally biased models. Nevertheless, the possibility of generating a representative synthetic dataset using real (currently) biased datasets or algorithms [203, 204] remains to be explored in the work ahead.

6.1.2 Learning From Limited Data

Obtaining annotations in the medical domain requires expert knowledge, which is a significant bottleneck in developing automated solutions. Typically, it is possible to obtain annotations for some limited data (e.g. one staining); however, as has also been confirmed by this thesis, due to domain shift, the model trained on such data will fail when applied to data with different characteristics (e.g. different stain or stain variation). The solutions presented in this thesis (Chapter 4 and Chapter

5) demonstrate that it is possible to develop a stain-invariant solution for glomeruli segmentation. However, it remains to be explored if it is possible to develop a model that is able to detect, irrespective of the staining protocol, other diagnostically relevant structures (e.g. tubules) using the annotations provided in a single stain. Another promising avenue is to develop a single model that is able to detect several diagnostically relevant structures in a stain invariant manner.

6.1.3 Federated Learning

Currently, the datasets used in the medical domain are very limited in size and representativeness [205]. Even in the cases where learning from a given medical dataset obtains near-to-human performances for the considered task, the size of the dataset is significantly smaller compared to those used in other artificial intelligence applications [205]. Moreover, the available data can be unintentionally biased towards specific genders or ethnicities resulting in unfair model predictions in under-represented patient groups [203, 204].

Given that vast amounts of data are generated daily in hospitals worldwide, theoretically, obtaining a representative dataset for a given task should be possible. However, apart from their heterogeneity due to different technical equipment and protocols, medical data falls under rigorous regulations for data protection that largely prevents them from being merged and forming a representative data collection.

The idea of Federated Learning (FL) [206], where a model is trained in a decentralised/distributed manner, is of great importance for the medical domain [207]. Instead of bringing all the data to one place, where the model is trained, the model itself is distributed to the sites (i.e. nodes) where the data are stored. Training iterations are performed locally on each node, followed by a global model update. Although such a training schema does not require explicit data sharing, the data still could be exposed by specific attacks or model inversion [202]. Another challenge is ensuring secure communication during training and defence against various model-related attacks, e.g. that the model does not get 'poisoned' during training [208]. All of this represents a promising and exciting research direction for the future of AI in the medical domain.

6.1.4 Advanced Approaches to Automatic Solutions

Current deep-learning approaches exploit available data to solve complex tasks performed by experts, e.g. pathologists. However, the expert's conclusion is not solely based on the data the model uses for training. In addition to experience, domain-specific knowledge is crucial in the decision-making process. Nevertheless, trying to approximate such a complex process by learning uniquely from one type of data, e.g. images, in some cases might be an oversimplification of the problem. In this context, steps toward an effective multi-domain solution seem promising. A system which can incorporate medical records in various formats (e.g. text, image, voice) could have a considerable potential to obtain general and robust knowledge. However, learning from such heterogeneous data is still not straightforward.

A promising direction for advanced approaches is the incorporation of the attention mechanism in the model's learning. Apart from state-of-the-art attention

approaches, an exciting direction can be to incorporate eye-tracking into the learning process [209, 210]. This can be used for training a model to focus on important aspects of the image [211], increasing its interpretability. Moreover, once the model is trained, it can serve as an advisory tool for the expert, pointing to the portions of the images that are not carefully investigated enough.

Appendix A

Medical Background

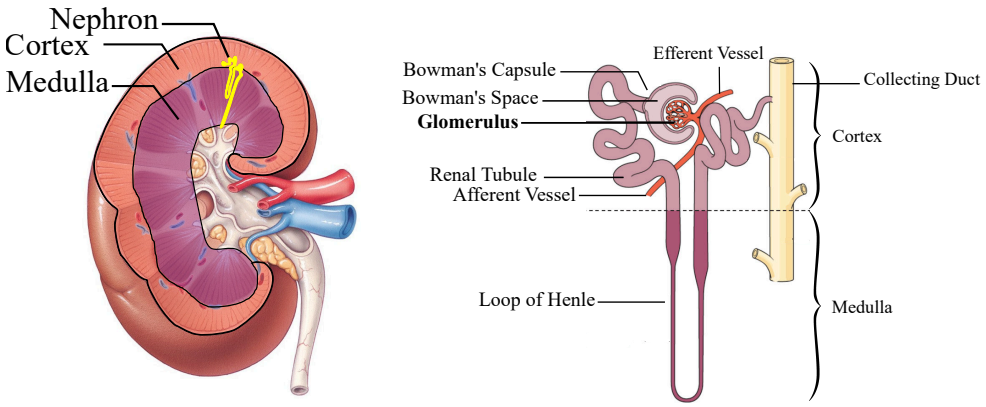
Staining

The tissue sections extracted and prepared to be microscopically examined are transparent, providing very little detail of tissue structure. In order to increase the contrast in the tissue and to make tissue components visible, the staining process is required. The staining always involves the visual labelling of some biological entity by attaching or depositing in its vicinity a marker of characteristic colour or form [7]. Since different tissue components are biochemically different, they will react with the provided staining differently. The basic principles on which the stain dye a tissue is building chemical compounds established between the dye and the tissue. For example, the acidophilic tissue components such as erythrocyte cytoplasm and collagen fibres [212] will bind to acid dyes such as hematoxylin; the basophilic tissue components such as nuclei [212] will bind to basic dyes such as eosin. Staining with haematoxylin and eosin (H&E) is routinely performed in histopathology. However, many other stains have been developed to highlight particular tissue components. Nevertheless, it is important to note that the colours in which tissue components appear to depend on the staining; they are not related to the tissue structures' colour.

Moreover, immunohistochemical stainings are employed to enable more specific information, such as the expression of a particular protein (antigen). The basic working principle of these stainings is antigen-antibody binding — e.g. a solution containing a special antibody is laid over the tissue, therefore, only cells having targeted antigen will attach the antibody, which further can be visualised [31]. There is an important difference between immunohistochemical (IHC) staining methods highlighting only one specific protein with a chromogenic label, as opposed to histochemical (HC) staining methods that are less specific. More general (histochemical) stains use chemicals that can interact with several tissue components, making them visible from different perspectives. In contrast, specific stains, such as immunohistochemical stains, are designed to highlight only specific proteins (e.g. a group of cells).

In the following, a brief description of stainings used in this thesis is given:

Hematoxylin and Eosin (H&E) staining: is classical histochemical staining



(a) Illustration of the kidney structure.

(b) Illustration of the nephron structure.

Figure A.1: Illustration of the kidney and nephron structures.

with a long history of clinical usage [213]. It highlights very general tissue components and enables the analysis of most organs and diseases. It contains two components, hematoxylin, which binds acidic structures such as cell nuclei, and eosin which binds basic components such as cytoplasmic proteins. Hematoxylin reaction results in the appearance of blue colour on the image while eosin produces pink. Thus, clear nuclear contrast can be achieved to reveal the distribution of cells [214]. This staining is the gold standard in diagnosing several types of cancer, and it is usually routinely performed in histological examinations.

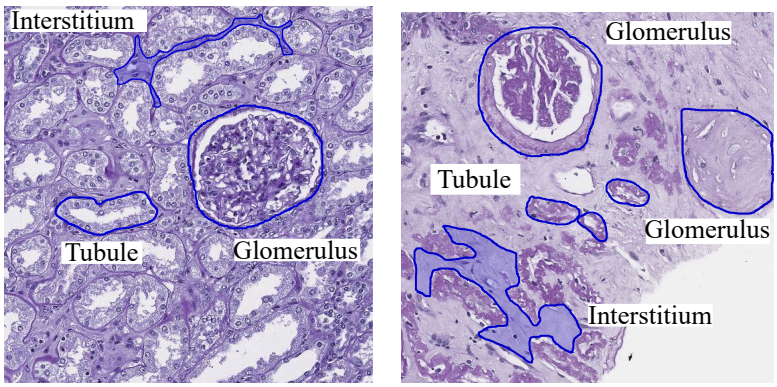
Jones Hematoxylin and Eosin (H&E) staining: is histochemical silver staining widely adapted in renal pathology to mark the basement membrane in black, nuclei in blue and the background in pink. Usually, it is used to demonstrate abnormalities of the glomerular basement membrane.

Periodic Acid-Schiff Reaction (PAS): is histochemical staining used to identify carbohydrates by exposing a tissue section to periodic acid oxidation and then staining them with Schiff's reagent [212]. It marks carbohydrate-containing cell components in magenta (shades of purplish pink) [33]. PAS is most commonly used to demonstrate cells filled with glycogen deposits, or the glycocalyx [33]. In renal pathology, according to Banff classification scheme [215], its usage is recommended for the identification of glomerulitis, tubulitis and any destruction of the tubular basement membrane.

Sirius Red: is a histochemical stain which marks connective tissue (collagen) red and cytoplasm in lighter violet or pink [33].

CD68: is an immunohistochemical stain which reflects the expression of a specific protein during macrophage differentiation and activation.

CD34: is an immunohistochemical stain which highlights blood vessels, specifically the inner layer (endothelium).



(a) Histology of the healthy kidney.

(b) Histology of the chronically rejected kidney.

Figure A.2: Comaprison of the histological differences (PAS stain) between healthy and rejected kidney sample.

Kidney

Kidneys are bean-shaped organs, part of a human's urinary system, that filter the blood and produce urine. Adult kidneys average $150g$ and are approximately $11cm$ long, $6cm$ wide and $3cm$ thick [216]. The kidney is divided into two main parts — the cortex and medulla, see Figure A.1a. The functional unit of the kidney is the nephron, and the average human kidney contains approximately 1 million nephrons [31]. The nephrons are located in both cortex and medulla, with some parts lying in the cortex (e.g. glomeruli) and the others in the medulla (e.g. loop of Henle). The main components of the nephron are illustrated in Figure A.1b. The blood comes through the afferent vessel to the glomerulus — a specialised network of capillaries which perform the filtration and goes out of the glomerulus via the efferent vessel. The glomeruli filtrated products are collected into Bowman's space (i.e. primary urine) and further processed in the renal tubule to finally form the urine in the collecting duct. The renal tube is essential in re-absorbing some of the glomeruli filtrated substances, e.g. glucose, water and amino acids, and returns them to the blood. For example, the loop of Henle additionally absorbs different salts and enables the concentration of the urine. Therefore, the tubular apparatus concentrates the filtrate from up to $200l$ of primary urine to $1.5l$ of the final urine [6].

The body's total blood volume is circulated through the kidneys about 300 times each day [31]. The kidneys have essential functions in the human body, such as filtration, pressure regulations, maintenance of acid-base balance etc. However, the prevalence of chronic kidney disease is very high, around 13.4.% [217], which indicates that kidney diseases may be more common than diabetes [217]. Kidney transplantation is the preferred treatment for end-stage renal disease patients as it offers better life quality than dialysis and has a higher survival rate [218]. Every year, thousands of patients undergo kidney transplantation. Nevertheless, the transplantation process can have short-term success. Some transplants have been rejected in the first months after the transplantation (acute rejection), while others

can be rejected after several years (chronic rejection).

Chronic rejection can be attributed to the immune response [219], and it is manifested by the progressive decrease in the glomeruli filtration rate. On the histological scan, chronic rejection affects all parts of the kidney — arteries, interstitium, glomeruli, and tubules [219], as demonstrated in Figure A.2. Therefore, detecting these structures can be an important indicator of kidney health status. According to the Banff classification scheme [215], the adequate kidney biopsy specimen should contain ten or more glomeruli and at least two arteries, while a minimal number of glomeruli is seven.

Appendix B

Stain Transfer

B.1 Dataset

All stain transfer models are trained on data extracted from the training patients in each staining in an unsupervised way. The patches are randomly extracted using a uniform sampling strategy, and the final dataset is composed of 5000 random patches per staining, in a size of 508×508 pixels, scaled to the range $[-1, 1]$, except for experiments in Chapter 5, where the patches have the size of 512×512 pixels.

In the case of intra-stain variation, where AIDPATH dataset [18] is used, patches were randomly extracted from patients 1 and 7 using a uniform sampling strategy. During test time, pre-trained models were applied to patients 1, 3 and 7 (patient 3 is kept as an out-of-training distribution sample since it contains sufficient glomeruli - 49).

B.2 CycleGAN Models

The model's architecture of 9 ResNet blocks is used. All CycleGAN models are trained for 50 epochs, with a learning rate of 0.0002, using the Adam optimiser and a batch size of 1. From the 25^{th} epoch, the learning rate linearly decayed to 0, and the cycle-consistency and identity weights halved. In all experiments, the translation model from the last (50^{th}) epoch is used. Moreover, Shrivastava et al.'s strategy [220] of updating the discriminator using the 50 previously generated samples is adopted to reduce model oscillation. The training parameters for CycleGAN models are taken from the original paper ($w_{\text{cyc}} = 10$, $w_{\text{id}} = 5$) [12]. Preliminary experiments showed that the visual translation quality is not highly dependent on the weight values of Eq. (3.4) (although some combinations required more training to obtain realistic translations). As has been discussed in more detail in Chapter 3 (Section 3.5), visual criteria is not a good proxy for assessing the quality of stain translation, and so the weighting values proposed by the original authors [12] are used since they already achieve visually good results.

B.3 StarGAN Training

The generator's and discriminator's architecture and training settings from the CycleGAN model (without instance normalisation in the discriminator as suggested in the original paper [14]) are employed. The loss weights are taken from the original StarGAN article ($w_{\text{cyc}} = 10$, $w_{\text{cls}} = 1$, $\lambda_{gp} = 10$) [14]. The model is trained for 50 epochs, and, similarly to the CycleGAN training strategy, the model from the 50^{th} epoch is used. Preliminary experiments were conducted with various values for the weight parameters in Eq. (3.5). Some combinations lead to unstable training or require more epochs to produce realistic translations in all stain combinations. Thus, the parameter values from the original paper were used [14].

B.4 Segmentation Models

The U-Net architecture [13] is used for segmentation. The U-Net training set comprised all the glomeruli from the 4 training patients - 662 for PAS, 624 for Jones H&E, 529 for CD68, 654 for Sirius Red, and 568 for CD34. Seven times more negative patches are extracted from each stain to account for the variance of non-glomeruli tissue. The validation sets (2 patients) were composed of 588 (PAS), 593 (Jones H&E), 524 (CD68), 579 (Sirius Red) and 598 (CD34) glomeruli patches.

The same training parameters are used for all experiments: the batch size of 8, learning rate of 0.0001, 250 epochs, and the network with the lowest validation loss is kept. The slide background (non-tissue) is removed by thresholding each image by its mean value, then removing small objects and closing holes. All patches are standardised to $[0, 1]$ and normalised by the mean and standard deviation of the (labelled) training set. The following augmentations are applied with an independent probability of 0.5 (batches are augmented 'on the fly'), in order to further force the network to learn general features: elastic deformation ($\sigma = 10$, $\alpha = 100$); random rotation in the range $[0°, 180°]$, random shift sampled from $[-205, 205]$ pixels, random magnification sampled from $[0.8, 1.2]$, and horizontal/vertical flip; additive Gaussian noise with $\sigma \in [0, 2.55]$; Gaussian filtering with $\sigma \in [0, 1]$; brightness, colour, and contrast enhancements with factors sampled from $[0.9, 1.1]$; stain variation by colour deconvolution [167], α sampled from $[-0.25, 0.25]$ and β from $[-0.05, 0.05]$. Due to the specificity of the U-Net architecture with valid convolutions, the central part of each is used (resulting in a segmentation patch size of 508×508). The predicted segmentation has a size of 324×324 pixels.

The best model is saved based on performance on the validation set, which is composed of patches extracted from validation patients. The performances of the best models are calculated over test patients in each of the experiments.

B.5 CycleGAN/StarGAN Noise Sensitivity - Additional Results

Figure B.1 and Figure B.2 represent reconstructions of target-stain images when intermediate translation to PAS stain obtained by CycleGAN/StarGAN model is corrupted with Gaussian zero-mean noise respectfully.

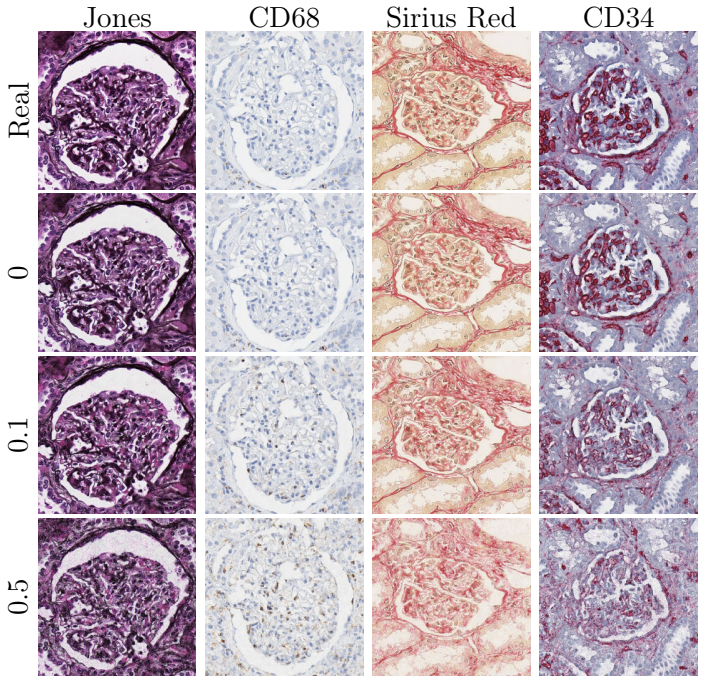


Figure B.1: CycleGAN — the effects of additive zero-mean Gaussian noise with different standard deviations to image reconstructions when added to the intermediate translation to PAS.

B.6 PixelCNN++ Model

The PixelCNN++ [181] architecture is used to model the underlying distribution of each stain: PAS, Jones H&E, CD68, Sirius Red, and CD34. The model employs 3 Resnet [221] blocks consisting of 5 residual layers in the encoding phase, with 2×2 downsampling between the ResNet blocks. The same architecture is employed in the decoding phase but with upsampling layers instead of downsampling. All residual layers utilise 160 filter maps in their convolutional layers and have a dropout of 0.5. The overall training for one PixelCNN++ model took approximately 15 days on an HPC with 4 V100 GPUs (in parallel).

Since each pixel value is conditioned on the product of all previously generated pixels, the models were trained and evaluated on patches of size 32×32 due to GPU memory limitations. For each stain, 1280000 train, validation, and test patches from the corresponding patents are extracted. The model is trained for 60 epochs with a learning rate of 0.001 and a decay rate of 0.999. The best model is saved based on the validation set's lowest bits-per-dimension score [222]. The validation set of 128000 patches randomly extracted from the validation patients is used. The original publicly available implementation [14] is employed.

[14] `https://github.com/openai/pixel-cnn`

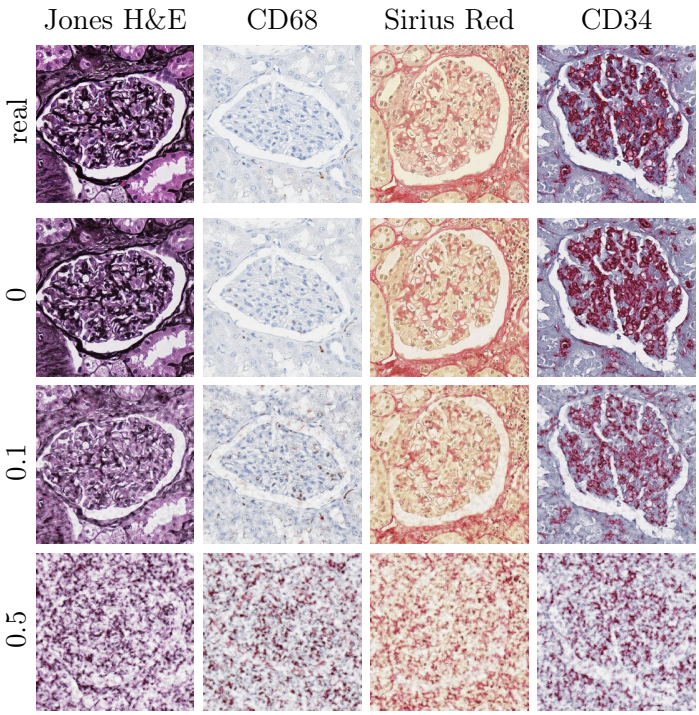


Figure B.2: StarGAN — the effects of additive zero-mean Gaussian noise added to intermediate representations of given stains on their reconstructions when added to the intermediate translation to PAS.

Appendix C

Stain Invariance

C.1 UDA-GAN

CycleGAN and StarGAN models are trained using the same dataset composition, architecture and training setting as provided in Appendix B. Similarly, as in all previous experiments, the segmentation model is UNet, having the same architecture as in B.4. The MDS1 approach uses baseline PAS models, which are trained according to the Appendix B.4. The MDS2 approach uses the same setting for the training, except that the image (PAS stain) is before feeding to the UNet model (during the training phase), translated using CycleGAN/StarGAN model into the given target staining. The proposed UDA-GAN augmentation is performed before other augmentation strategies. The segmentation model architecture and training setting are the same for all segmentation models (UDA-SD, UDA-GAN, MDS1, MDS2 and Multi UDA-CGAN).

C.1.1 UDA-GAN Robustness

Due to implementation differences, in addition to differences resulting from training settings and the CycleGAN epoch chosen (as demonstrated in Chapter 3), there are differences in the reported MDS1 results presented in Chapter 3 and the results presented in Section 4.1. The experiments in Chapter 3 are implemented in Tensorflow 2, while those from Section 4.1. are implemented in the Keras framework (now deprecated since Keras has been integrated into Tensorflow 2). However, these reveal that the proposed UDA-GAN approach is less sensitive to such variation than other approaches. Table C.1 presents the results for MDS1, MDS2 and UDA-GAN using TensorFlow implementation of a CycleGAN (same models as used in Chapter 3). Although absolute numbers are different, the UDA-GAN approach is far less sensitive to non-visible changes in translation when compared to MDS1 and MDS2. The overall gain in the hardest target, CD68, remains in the same range, 13%, compared to the next best-performing method (MDS1), confirming the importance of the proposed method in terms of its ability to generalise. In other stainings, the UDA-GAN model performs close to their baseline results (see Table 4.2).

Table C.1 Quantitative results for each strategy trained on PAS (source staining) and tested on different (target) stainings. Standard deviations are in parentheses, and the highest F_1 scores for each staining are in bold.

Training Strategy	Score	Test Staining					
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
MDS1[20]	F_1	-	**0.866** (0.017)	0.637 (0.034)	**0.880** (0.015)	0.754 (0.033)	0.784 (0.113)
	Precision	-	0.842 (0.035)	0.846 (0.050)	0.846 (0.031)	0.879 (0.027)	0.853 (0.018)
	Recall	-	0.894 (0.020)	0.516 (0.058)	0.918 (0.008)	0.662 (0.059)	0.747 (0.193)
MDS2[20]	F_1	-	0.852 (0.033)	0.628 (0.079)	0.820 (0.032)	0.848 (0.026)	0.787 (0.107)
	Precision	-	0.811 (0.061)	0.527 (0.115)	0.737 (0.054)	0.799 (0.049)	0.718 (0.132)
	Recall	-	0.901 (0.005)	0.803 (0.030)	0.928 (0.012)	0.905 (0.007)	0.884 (0.055)
UDA-CGAN	F_1	**0.900** (0.015)	0.853 (0.013)	**0.760** (0.035)	0.848 (0.045)	**0.849** (0.023)	**0.842** (0.051)
	Precision	0.864 (0.032)	0.794 (0.028)	0.735 (0.072)	0.785 (0.079)	0.805 (0.046)	0.797 (0.046)
	Recall	0.940 (0.007)	0.922 (0.007)	0.791 (0.021)	0.927 (0.007)	0.899 (0.008)	0.896 (0.060)

C.2 DANN

C.2.1 Training Details

The discriminator is a Fully Connected Neural Network (FCNN) whose first layer is the Gradient Reversal Layer, followed by GlobalAveragePooling. The network contains four fully-connected layers (the number of nodes is $512 - 256 - 64 - 1$) followed by Batch norm and using the LeakyRelu activation function, except for the last layer in which the sigmoid activation function is used.

In the original DANN manuscript [146], the authors proposed the usage of λ_{hp} in order to reduce the influence of a signal from the untrained discriminator at the early stages of training. However, according to preliminary experiments, when starting from a pre-trained model, as is in these experiments, better alignment is obtained when λ_{hp} is set to 1 starting from the first epoch.

The overall model containing the UNet and the Discriminator is trained using Stochastic Gradient Descent (SGD) with $lr = 0.0001$. Half of the training batch contains source (annotated) data, and the other half target (unannotated) data. The UNet is trained using categorical cross entropy and source data, while the Discriminator and the part of the UNet to which the Discriminator is attached are trained using binary cross entropy with label smoothing 0.1 and batches containing both source and target data.

UNet pre-training The UNet is pre-trained on PAS training data using a batch size of 8, a learning rate of 0.0001 and 100 epochs, and the network with the low-

est validation loss is kept. All patches are standardised to $[0, 1]$. The following augmentations are applied with an independent probability of 0.5 (batches are augmented 'on the fly'): elastic deformation ($\sigma = 10$, $\alpha = 100$); random rotation in the range $[0°, 180°]$, random shift sampled from $[-205, 205]$ pixels, random magnification sampled from $[0.8, 1.2]$, and horizontal/vertical flip; additive Gaussian noise with $\sigma \in [0, 2.55]$; Gaussian filtering with $\sigma \in [0, 1]$; brightness, colour, and contrast enhancements with factors sampled from $[0.9, 1.1]$; stain variation by colour deconvolution [167], α sampled from $[-0.25, 0.25]$ and β from $[-0.05, 0.05]$.

C.2.2 MDS2 Individual Models and Adaptation Results:

Table C.2 Jones H&E - MDS2 approach and its DANN adaptation.

Training strategy	Score	Model1	Model2	Model3	Average	Std
MDS2 baseline	F-score	**0.878**	0.872	0.879	**0.876**	0.003
	Precision	0.859	0.848	0.858	**0.855**	0.005
	Recall	0.898	0.898	0.901	0.899	0.002
MDS2 adapted	F-score	0.877	**0.883**	**0.881**	**0.880**	0.002
	Precision	0.845	0.860	0.848	0.851	0.007
	Recall	0.913	0.907	0.916	**0.912**	0.004

Table C.3 CD68 - MDS2 approach and its DANN adaptation.

Training strategy	Score	Model1	Model2	Model3	Average	Std
MDS2 baseline	F-score	0.668	0.503	0.638	0.603	0.072
	Precision	0.588	0.360	0.549	0.499	0.100
	Recall	0.773	0.836	0.762	**0.790**	0.032
MDS2 adapted	F-score	**0.678**	**0.712**	**0.733**	**0.708**	0.023
	Precision	0.738	0.638	0.699	**0.691**	0.041
	Recall	0.627	0.806	0.771	0.735	0.077

Table C.4 Sirius Red - MDS2 approach and its DANN adaptation.

Training strategy	Score	Model1	Model2	Model3	Average	Std
MDS2 baseline	F-score	0.841	0.766	0.835	0.814	0.034
	Precision	0.777	0.650	0.763	0.730	0.057
	Recall	0.918	0.932	0.923	0.924	0.006
MDS2 adapted	F-score	**0.855**	**0.833**	**0.859**	**0.849**	0.011
	Precision	0.795	0.747	0.799	**0.781**	0.024
	Recall	0.924	0.942	0.930	**0.932**	0.007

Table C.5 CD34 - MDS2 approach and its DANN adaptation.

Training strategy	Score	Model1	Model2	Model3	Average	Std
MDS2 baseline	F-score	0.839	0.880	**0.870**	0.863	0.018
	Precision	0.774	0.863	0.838	0.825	0.038
	Recall	0.915	0.898	0.905	**0.906**	0.007
MDS2 baseline	F-score	**0.869**	**0.880**	0.863	**0.871**	0.007
	Precision	0.835	0.867	0.831	**0.845**	0.016
	Recall	0.907	0.892	0.897	0.899	0.006

Appendix D

HistoStarGAN - Additional Results

D.1 Histopathological Images

The HistoStarGAN model is able to accurately segment and translate unseen renal histological images (unseen stainings) taken from the internet, some of the examples given in Figure D.1.

D.2 Semantic Generation Potential

The HistoStarGAN is also able to create histopathological images of glomeruli starting from the ellipse-like image, see Figure D.2. This opens a possibility for HistoStarGAN to be extended into a semantic histopathological image generator, a SPADE-like tool [223] to enable semantic generation of histopathological images based on a class-related drawing — e.g. glomeruli, tubule, nuclei etc. Based on preliminary experiments, it seems that HistoStarGAN is able to generate kidney structures based on specific patterns provided in the input, e.g. glomeruli-like structures based on circle-like textures or tubules based on empty circles. However, this does not necessarily imply that every circle-like structure will be represented as a glomeruli/tubule. The potential and limitations of such a use case remain to be explored in future work.

[15] Image credits in order of appearance: https://static.cambridge.org/binary/version/id/urn:cambridge.org:id:binary:20181009125204364-0075:9781107281981:61398fig4_6.png?pub-status=live, https://www.kidneypathology.com/English_version/Membranoproliferative_GN.html, https://www.kidneypathology.com/English_version/Histologic_patterns.html, https://commons.wikimedia.org/wiki/File:Membranous_nephropathy_-_alt_-_mpas_-_very_high_mag.jpg.

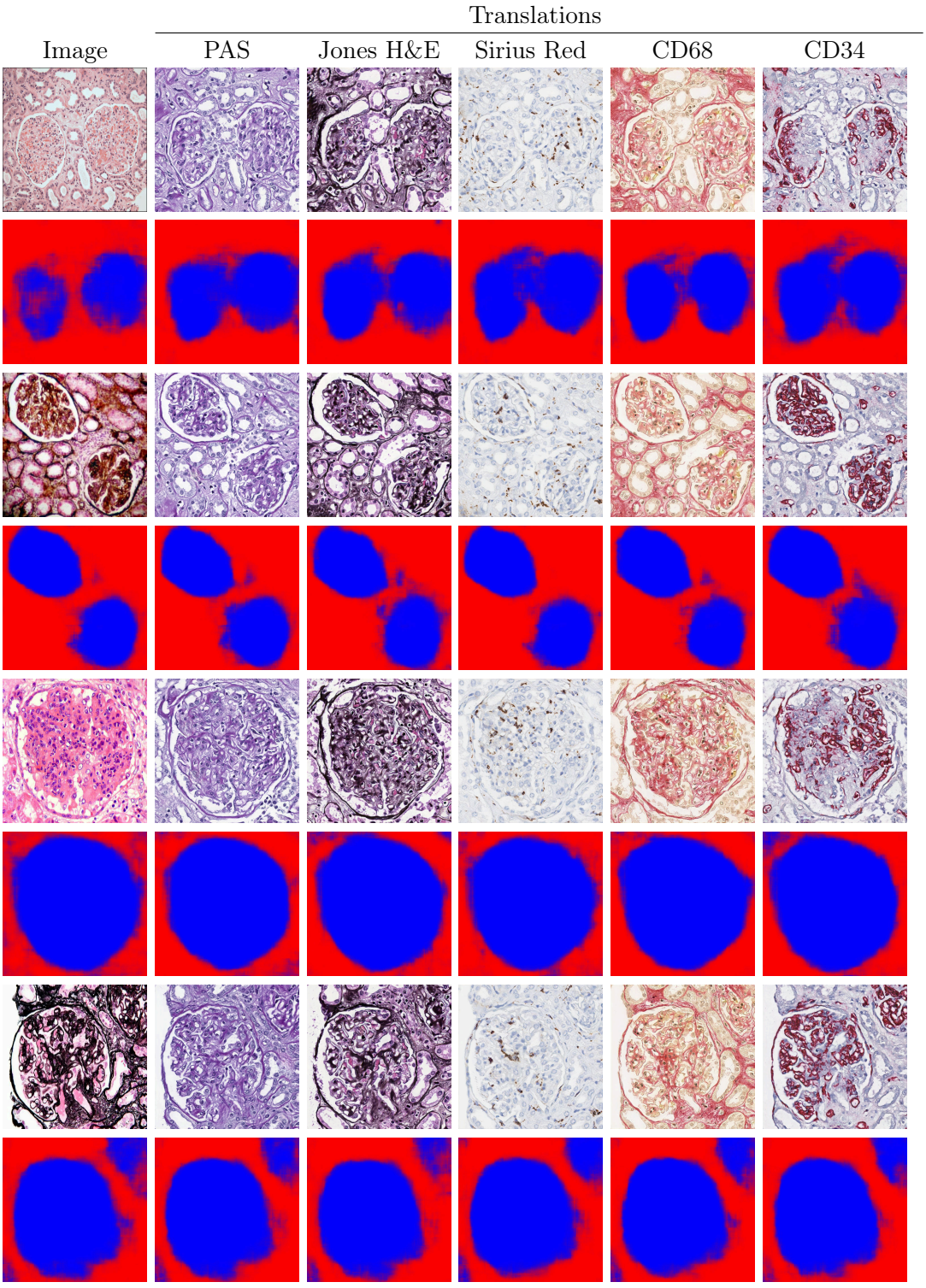


Figure D.1: HistoStarGAN - Images generated from histological images taken from the internet [15].

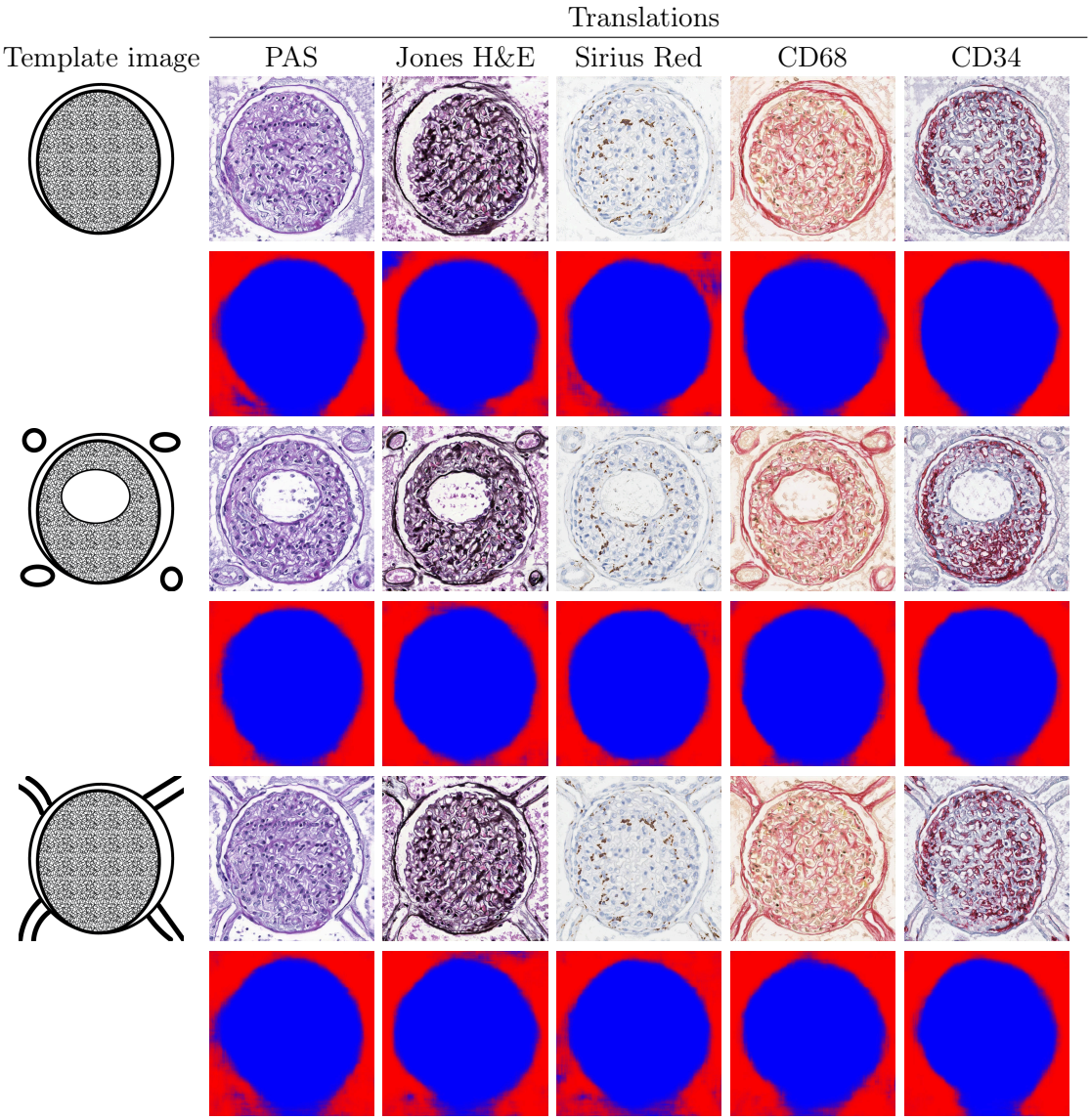


Figure D.2: HistoStarGAN - Semantic image generation based on the provided template image. The template image used in the first row is modified by adding circles and lines and provided as an input to the HistoStarGAN model in the third and fifth rows. All template images are created in Adobe Illustrator, using predefined patterns.

List of Figures

List of Tables

List of References

[1] Terrence J. Sejnowski. *The Deep Learning Revolution.* The MIT Press, 2018. ISBN 9780262038034.

[2] Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021.

[3] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: Methodological failures and recommendations for the future. *NPJ Digital Medicine*, 5(1):1–8, 2022.

[4] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 2672–2680, 2014.

[6] Maja Temerinac-Ott, Germain Forestier, Jessica Schmitz, Meyke Hermsen, JH Bräsen, Friedrich Feuerhake, and Cédric Wemmert. Detection of glomeruli in renal pathology by mutual comparison of multiple staining modalities. In *Proceedings of the International Symposium on Image and Signal Processing and Analysis*, pages 19–24, 2017.

[7] John D Bancroft and Marilyn Gamble. *Theory and Practice of Histological Techniques (sixth Edition).* Elsevier Health Sciences, sixth edition edition, 2008.

[8] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva de Souza, Alexi Baidoshvili, Geert Litjens, Bram van Ginneken, Iris Nagtegaal, and Jeroen van der Laak. The importance of stain normalization in colorectal

tissue classification with convolutional networks. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 160–163, 2017.

[9] Gabriela Csurka. *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer International Publishing, 2017.

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018.

[15] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGANv2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8188–8197, 2020.

[16] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14154–14163, 2021.

[17] Thomas Lampert, Odyssée Merveille, Jessica Schmitz, Germain Forestier, Friedrich Feuerhake, and Cédric Wemmert. Strategies for training stain invariant CNNs. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 905–909, 2019.

[18] Gloria Bueno, Lucia Gonzalez-Lopez, Marcial Garcia-Rojo, Arvydas Laurinavicius, and Oscar Deniz. Data for glomeruli characterization in histopathological images. *Data in Brief*, 29:105314, 2020.

[19] Michael Gadermayr, Vitus Appel, Mara Barbara Klinkhammer, Peter Boor, and Dorit Merhof. Which way round? A study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In *Proceedings*

of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 11071, pages 165–173, 2018.

[20] Michael Gadermayr, Laxmi Gupta, Vitus Appel, Peter Boor, Barbara M. Klinkhammer, and Dorit Merhof. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: A study on kidney histology. *IEEE Transactions on Medical Imaging*, 38:2293–2302, 2019.

[21] Creative Destruction Lab. Geoff hinton: On radiology, November 2016. URL `https://www.youtube.com/watch?v=2HMPRXstSvQ`.

[22] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67, 2021.

[23] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27(5):775–784, 2021.

[24] Stephen Chan and Eliot L Siegel. Will machine learning end the viability of radiology as a thriving medical specialty? *The British Journal of Radiology*, 92(1094):20180416, 2019.

[25] Juan Manuel Durán and Karin Rolanda Jongsma. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, 47(5):329–335, 2021.

[26] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.

[27] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[28] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[29] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics Guidelines for Trustworthy Ai*. Publications Office, 2019.

[30] Adrian P. Brady. Artificial intelligence in radiology: An exciting future, but ethically complex. *EMJ Radiology*, pages 54–57, 2021.

[31] Barbara Young, Phillip Woodford, and Geraldine O'Dowd. *Wheater's Functional Histology E-book: A Text and Colour Atlas*. Churchill Livingston/Elsevier Philadelphia, PA, sixth edition / barbara young, geraldine o'dowd, phillip woodford. edition, 2014.

[32] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E Zuckerman, Thomas Chong, Anthony E Sisk, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3(6):466–477, 2019.

[33] Anthony Mescher. *Junqueira's Basic Histology: Text & Atlas*. McGraw-Hill Education, fifteenth edition edition, 2018. ISBN 978-1-26-002618-4.

[34] Hani A Alturkistani, Faris M Tashkandi, and Zuhair M Mohammedsaleh. Histological stains: A literature review and case study. *Global Journal of Health Science*, 8(3):72, 2016.

[35] Kim S. Suvarna, Christopher Layton, and John D. Bancroft. *Bancroft's Theory and Practice of Histological Techniques*. Elsevier, eighth edition edition, 2019. ISBN 978-0702068645.

[36] Syed Ahmed Taqi, Syed Abdus Sami, Lateef Begum Sami, and Syed Ahmed Zaki. A review of artifacts in histopathology. *Journal of Oral and Maxillofacial Pathology : JOMFP*, 22(2):279, 2018.

[37] Birgid Schömig-Markiefka, Alexey Pryalukhin, Wolfgang Hulla, Andrey Bychkov, Junya Fukuoka, Anant Madabhushi, Viktor Achter, Lech Nieroda, Reinhard Büttner, Alexander Quaas, et al. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Modern Pathology*, 34(12):2098–2108, 2021.

[38] Varun Rastogi, Naveen Puri, Swati Arora, Geetpriya Kaur, Lalita Yadav, and Rachna Sharma. Artefacts: A diagnostic dilemma - a review. *Journal of clinical and diagnostic research: JCDR*, 7(10):2408–2413, 2013.

[39] Cédric Wemmert, Jonathan Weber, Friedrich Feuerhake, and Germain Forestier. Deep learning for histopathological image analysis. In *Deep Learning for Biomedical Data Analysis*, pages 153–169. Springer International Publishing, 2021.

[40] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 23885–23899, 2021.

[41] Laxmi Gupta, Barbara M. Klinkhammer, Peter Boor, Dorit Merhof, and Michael Gadermayr. GAN-based image enrichment in digital pathology boosts segmentation accuracy. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11764, pages 631–639, 2019.

[42] Anne Grote, Nadine S Schaadt, Germain Forestier, Cédric Wemmert, and Friedrich Feuerhake. Crowdsourcing of histological image labeling and object delineation by medical students. *IEEE Transactions on Medical Imaging*, 38: 1284–1294, 2018.

[43] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The camelyon dataset. *GigaScience*, 7(6): giy065, 2018.

[44] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 27, 2014.

[45] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with Pixelcnn decoders. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2016.

[46] Asir Saeed, Suzana Ilic, and Eva Zangerle. Creative GANs for generating poems, lyrics, and metaphors. *arXiv preprint arXiv:1909.09534*, 2019.

[47] Yi Yu, Abhishek Srivastava, and Simon Canales. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17:1–20, 2021.

[48] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video generative adversarial networks: A review. *ACM Computing Surveys*, 55(2), 2022.

[49] California State. Assembly bill no. 730, elections: Deceptive audio or visual media., 2020.

[50] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[51] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do gans learn the distribution? some theory and empirics. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[52] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys*, 54 (2):1–38, 2021.

[53] Pourya Shamsolmoali, Masoumeh Zareapoor, Eric Granger, Huiyu Zhou, Ruili Wang, M Emre Celebi, and Jie Yang. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*, 72: 126–146, 2021.

[54] Yongjun Hong, Uiwon Hwang, Jaeyoon Yoo, and Sungroh Yoon. How generative adversarial networks and their variants work an overview. *ACM Computing Surveys*, 52(1):1–43, 2020.

[55] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.

[56] David Morrison and David Harris-Birtill. Anonymising pathology data using generative adversarial networks. In *Medical Imaging 2022: Digital and Computational Pathology*, volume 12039, pages 267–271, 2022.

[57] Marlen Runz, Daniel Rusche, Stefan Schmidt, Martin R Weihrauch, Jürgen Hesser, and Cleo-Aron Weis. Normalization of HE-stained histological images using cycle consistent generative adversarial networks. *Diagnostic Pathology*, 16(1):1–10, 2021.

[58] Kevin de Haan, Yijie Zhang, Jonathan E Zuckerman, Tairan Liu, Anthony E Sisk, Miguel FP Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, et al. Deep learning-based transformation of H&E stained tissues into special stains. *Nature Communications*, 12(1):1–13, 2021.

[59] Neslihan Bayramoglu, Mika Kaakinen, Lauri Eklund, and Janne Heikkilä. Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 64–71, 2017.

[60] Raphaël Marée, Loïc Rollus, Benjamin Stévens, Renaud Hoyoux, Gilles Louppe, Rémy Vandaele, Jean-Michel Begon, Philipp Kainz, Pierre Geurts, and Louis Wehenkel. Collaborative analysis of multi-gigapixel imaging data using cytomine. *Bioinformatics*, 32(9):1395–1401, 2016.

[61] Gloria Bueno, M Milagro Fernandez-Carrobles, Lucia Gonzalez-Lopez, and Oscar Deniz. Glomerulosclerosis identification in whole slide images using semantic segmentation. *Computer Methods and Programs in Biomedicine*, 184:105273, 2020.

[62] Jelica Vasiljević, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Self adversarial attack as an augmentation method for immunohistochemical stainings. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1939–1943, 2021.

[63] Jelica Vasiljević, Zeeshan Nisar, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. CycleGAN for virtual stain transfer:is seeing really believing? *Artificial Intelligence in Medicine (under review)*, page 30, 2021.

[64] Jelica Vasiljević, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks. *Neurocomputing*, 460: 277–291, 2021.

[65] Jelica Vasiljević, Zeeshan Nisar, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. HistoStarGAN: A unified approach to stain normalisation, stain transfer and stain invariant segmentation in renal histopathology. *(under review)*, 2022.

[66] Joaquin Qui nonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2008.

[67] Nicolas Brieu, Armin Meier, Ansh Kapil, Ralf Schoenmeyer, Christos G Gavriel, Peter D Caie, and Günter Schmidt. Domain adaptation-based augmentation for weakly supervised nuclei detection. In *MICCAI 2019 Workshop COMPAY*, 2019.

[68] Laxmi Gupta, Barbara Mara Klinkhammer, Peter Boor, Dorit Merhof, and Michael Gadermayr. Stain independent segmentation of whole slide images: A case study in renal histology. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1360–1364, 2018.

[69] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58: 101544, 2019.

[70] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63, 2020.

[71] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[72] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[73] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[74] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70, pages 214–223, 2017.

[75] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a large-scale study. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[76] Hongxia Gao, Zhanhong Chen, Binyang Huang, Jiahe Chen, and Zhifu Li. Image super resolution based on conditional generative adversarial network. *IET Image Processing*, 14(13):3006–3013, 2020.

[77] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, 51(2):2007–2028, 2020.

[78] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018.

[79] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9015–9024, 2019.

[80] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5914–5922, 2019.

[81] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3673–3682, 2019.

[82] Fabienne Anglade, Danny A Milner Jr, and Jane E Brock. Can pathology diagnostic services for cancer be stratified and serve global health? *Cancer*, 126:2431–2438, 2020.

[83] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):801–814, 2019.

[84] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2208–2217, 2017.

[85] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5): 1–46, 2020.

[86] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 7523–7532, 2019.

[87] Yijie Zhang, Kevin de Haan, Yair Rivenson, Jingxi Li, Apostolos Delis, and Aydogan Ozcan. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light: Science & Applications*, 9(1):1–13, 2020.

[88] Ye Liu, Sophia J Wagner, and Tingying Peng. Multi-modality microscopy image style augmentation for nuclei segmentation. *Journal of Imaging*, 8, 2022.

[89] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.

[90] Panagiotis Dimitrakopoulos, Giorgos Sfikas, and Christophoros Nikou. ISING-GAN: Annotated data augmentation with a spatially constrained generative adversarial network. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1600–1603, 2020.

[91] Aman Rana, Alarice Lowe, Marie Lithgow, Katharine Horback, Tyler Janovitz, Annacarolina Da Silva, Harrison Tsai, Vignesh Shanmugam, Akram Bayat, and Pratik Shah. Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis. *JAMA Network Open*, 3(5):e205111–e205111, 2020.

[92] Dan Li, Hui Hui, Yingqian Zhang, Wei Tong, Feng Tian, Xin Yang, Jie Liu, Yundai Chen, and Jie Tian. Deep learning for virtual histological staining of bright-field microscopic images of unlabeled carotid artery tissue. *Molecular Imaging and Biology*, 22(5):1301–1309, 2020.

[93] Aman Rana, Gregory Yauney, Alarice Lowe, and Pratik Shah. Computational histological staining and destaining of prostate core biopsy RGB images with generative adversarial neural networks. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 828–834, 2018.

[94] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 953–956, 2019.

[95] Shaojin Cai, Yuyang Xue, Qinquan Gao, Min Du, Gang Chen, Hejun Zhang, and Tong Tong. Stain style transfer using transitive adversarial networks. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 163–172, 2019.

[96] Thomas de Bel, Meyke Hermsen, Jesper Kers, Jeroen van der Laak, and Geert J. S. Litjens. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, volume 102, pages 151–163, 2019.

[97] Thomas de Bel, John Melle Bokhorst, Jeroen van der Laak, and Geert Litjens. Residual CycleGAN for robust domain transformation of histopathological tissue slides. *Medical Image Analysis*, 70, 2021.

[98] Hongtao Kang, Die Luo, Weihua Feng, Shaoqun Zeng, Tingwei Quan, Junbo Hu, and Xiuli Liu. StainNet: A fast and robust stain normalization network. *Frontiers in Medicine*, 8, 2021.

[99] Aman Shrivastava, Will Adorno, Yash Sharma, Lubaina Ehsan, S. Asad Ali, Sean R. Moore, Beatrice C. Amadi, Paul Kelly, Sana Syed, and Donald E. Brown. Self-attentive adversarial stain normalization. In *Proceedings of the International Conference on Pattern Recognition*, pages 120–140, 2021.

[100] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Ling Shao. Structure preserving stain normalization of histopathology images using self-supervised semantic guidance. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 309–319, 2021.

[101] Jing Ke, Yiqing Shen, Xiaoyao Liang, and Dinggang Shen. Contrastive learning based stain normalization across multiple tumor in histopathology. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 571–580, 2021.

[102] Farhad Ghazvinian Zanjani, Svitlana Zinger, Babak Ehteshami Bejnordi, Jeroen AWM van der Laak, and Peter HN de With. Stain normalization of histopathology images using generative adversarial networks. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 573–577, 2018.

[103] Farhad G Zanjani, Svitlana Zinger, Babak E Bejnordi, Jeroen AWM van der Laak, et al. Histopathology stain-color normalization using deep generative models. In *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, pages 1–11, 2018.

[104] Pegah Salehi and Abdolah Chalechale. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In *Proceedings of the International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–7, 2020.

[105] Harshal Nishar, Nikhil Chavanke, and Nitin Singhal. Histopathological stain transfer using style transfer network with adversarial loss. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 330–340, 2020.

[106] Aïcha BenTaieb and Ghassan Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE Transactions on Medical Imaging*, 37 (3):792–802, 2017.

[107] Hanwen Liang, Konstantinos Plataniotis, and Xingyu Li. Stain style transfer of histopathology images via structure-preserved generative learning. In *International Workshop on Machine Learning for Medical Image Reconstruction*, 2020.

[108] Atefeh Ziaei Moghadam, Hamed Azarnoush, Seyyed Ali Seyyedsalehi, and Mohammad Havaei. Stain transfer using generative adversarial networks and disentangled features. *Computers in Biology and Medicine*, 142:105219, 2022.

[109] Tasleem Kausar, Adeeba Kausar, Muhammad Adnan Ashraf, Muhammad Farhan Siddique, Mingjiang Wang, Muhammad Sajid, Muhammad Zeeshan Siddique, Anwar Ul Haq, and Imran Riaz. SA-GAN: Stain acclimation generative adversarial network for histopathology image analysis. *Applied Sciences*, 12(1):288, 2021.

[110] Haseeb Nazki, Ognjen Arandjelović, InHwa Um, and David Harrison. MultiPathGAN: Structure preserving stain normalization using unsupervised multi-domain adversarial network with perception loss. *arXiv preprint arXiv:2204.09782*, 2022.

[111] Sophia Wagner, Nadieh Khalili, Raghav Sharma, Melanie Boxberg, Carsten Marr, Walter de Back, and Tingying Peng. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 257–266, 2021.

[112] Yuan Xue, Jiarong Ye, Qianying Zhou, L. Rodney Long, Sameer Antani, Zhiyun Xue, Carl Cornwell, Richard Zaino, Keith C. Cheng, and Xiaolei Huang. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. *Medical Image Analysis*, 67:101816, 2021.

[113] Wenyuan Li, Jiayun Li, Jennifer Polson, Zichen Wang, William Speier, and Corey Arnold. High resolution histopathology image generation and segmentation through adversarial training. *Medical Image Analysis*, 75:102251, 2022.

[114] Jeremias Krause, Heike I Grabsch, Matthias Kloor, Michael Jendrusch, Amelie Echle, Roman David Buelow, Peter Boor, Tom Luedde, Titus J Brinker, Christian Trautwein, et al. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. *The Journal of Pathology*, 254(1): 70–79, 2021.

[115] Hejun Wu, Rong Gao, Yeong Poh Sheng, Bo Chen, and Shuo Li. SDAE-GAN: Enable high-dimensional pathological images in liver cancer survival prediction with a policy gradient based data augmentation method. *Medical Image Analysis*, 62:101640, 2020.

[116] Apostolia Tsirikoglou, Karin Stacke, Gabriel Eilertsen, and Jonas Unger. Primary tumor and inter-organ augmentations for supervised lymph node colon adenocarcinoma metastasis detection. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 624–633, 2021.

[117] Adrian B Levine, Jason Peng, David Farnell, Mitchell Nursey, Yiping Wang, Julia R Naso, Hezhen Ren, Hossein Farahani, Colin Chen, Derek Chiu, et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *The Journal of Pathology*, 252(2):178–188, 2020.

[118] Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, and Mitko Veta. Learning domain-invariant representations of histological images. *Frontiers in Medicine*, 6:162, 2019.

[119] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3851–3860, 2020.

[120] Mara Graziani, Sebastian Otálora, Henning Muller, and Vincent Andrearczyk. Guiding CNNs towards relevant concepts by multi-task and adversarial learning. *arXiv preprint arXiv:2008.01478*, 2020.

[121] Niccolo Marini, Manfredo Atzori, Sebastian Otalora, Stephane Marchand-Maillet, and Henning Muller. H&E-adversarial network: A convolutional neural network to learn stain-invariant features through hematoxylin & eosin regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision Workshops (ICCVW)*, pages 601–610, 2021.

[122] Frauke Wilm, Katharina Breininger, and Marc Aubreville. Domain adversarial retinanet as a reference algorithm for the mitosis domain generalization (midog) challenge. *arXiv preprint arXiv:2108.11269*, 2021.

[123] Jian Ren, Ilker Hacihaliloglu, Eric A. Singer, David J. Foran, and Xin Qi. Unsupervised domain adaptation for classification of histopathology whole-slide images. *Frontiers in Bioengineering and Biotechnology*, 7:102, 2019.

[124] Sezgin M Ismail, Angela B Colclough, John S Dinnen, Douglas Eakins, DM Evans, Ernest Gradwell, Jerry P O'Sullivan, Joan M Summerell, and Robert G Newcombe. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *British Medical Journal*, 298 (6675):707–710, 1989.

[125] Sergio G Veloso, Mario F Lima, Paulo G Salles, Cynthia K Berenstein, Joao D Scalon, and Eduardo A Bambirra. Interobserver agreement of gleason score and modified gleason score in needle biopsy and in surgical specimen of prostate cancer. *International Brazilian Journal of Urology*, 33(5):639–651, 2007.

[126] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1107–1110, 2009.

[127] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5): 34–41, 2001.

[128] Abhishek Vahadane, Tingying Peng, Shadi Albarqouni, Maximilian Baust, Katja Steiger, Anna Schlitter, Amit Sethi, Irene Esposito, and Nassir Navab.

Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016.

[129] Massimo Salvi, Nicola Michielli, and Filippo Molinari. Stain color adaptive normalization (scan) algorithm: Separation and standardization of histological stains in digital pathology. *Computer Methods and Programs in Biomedicine*, 193:105506, 2020.

[130] Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Dingyi Hu, Shujiao Sun, Jun Shi, and Chenghai Xue. Stain standardization capsule for application-driven histopathological image normalization. *IEEE Journal of Biomedical and Health Informatics*, 25(2):337–347, 2021.

[131] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.

[132] Andrew Janowczyk, Ajay Basavanhally, and Anant Madabhushi. Stain normalization using sparse autoencoders (stanosa): Application to digital pathology. *Computerized Medical Imaging and Graphics*, 57:50–61, 2017.

[133] Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using GAN for histopathological images. *arXiv preprint arXiv:1710.08543*, 2017.

[134] Cong Cong, Sidong Liu, Antonio Di Ieva, Maurice Pagnucco, Shlomo Berkovsky, and Yang Song. Semi-supervised adversarial learning for stain normalisation in histopathology images. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 581–591, 2021.

[135] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[136] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 29, 2016.

[137] David Tellez, Maschenka Balkenhol, Nico Karssemeijer, Geert J. S. Litjens, Jeroen van der Laak, and Francesco Ciompi. H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection. In *Medical Imaging 2018: Digital Pathology*, page 34, 2018.

[138] Khrystyna Faryna, Jeroen van der Laak, and Geert Litjens. Tailoring automated data augmentation to H&E-stained histopathology. In *Medical Imaging with Deep Learning (MIDL)*, pages 168–178, 2021.

[139] Jia-Ren Chang, Min-Sheng Wu, Wei-Hsiang Yu, Chi-Chung Chen, Cheng-Kung Yang, Yen-Yu Lin, and Chao-Yuan Yeh. Stain mix-up: Unsupervised domain generalization for histopathology images. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 117–126, 2021.

[140] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18613–18624, 2020.

[141] Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, and Daniel L Rubin. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40:3945–3954, 2021.

[142] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta, and Joel H Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8533–8542, 2019.

[143] Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, Pim Moeskops, and Mitko Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. In *MICCAI 2017 Workshop on Deep Learning in Medical Image Analysis*, pages 83–91, 2017.

[144] Jian Ren, Ilker Hacihaliloglu, Eric A. Singer, David J. Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 201–209, 2018.

[145] Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in Bioengineering and Biotechnology*, 7:198, 2019.

[146] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International conference on machine learning (ICML)*, volume 37, page 1180–1189, 2015.

[147] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176, 2017.

[148] Yiyang Lin, Bowei Zeng, Yifeng Wang, Yang Chen, Zijie Fang, Jian Zhang, Xiangyang Ji, Haoqian Wang, and Yongbing Zhang. Unpaired multi-domain stain transfer for kidney histopathological images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[149] 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 on Digital Pathology*, 11435:47–55, 2019.

[150] Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE Transactions on Medical Imaging*, 40:1977–1989, 2021.

[151] Zidui Xu, Xi Li, Xihan Zhu, Luyang Chen, Yonghong He, and Yupeng Chen. Effective immunohistochemistry pathology microscopy image generation using CycleGAN. *Frontiers in Molecular Biosciences*, 7:571180, 2020.

[152] Zhaoyang Xu, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. Gan-based virtual re-staining: A promising solution for whole slide image analysis. *arXiv preprint arXiv:1901.04059*, 2019.

[153] Joshua J Levy, Christopher R Jackson, Aravindhan Sriharan, Brock C Christensen, and Louis J Vaickus. Preliminary evaluation of the utility of deep generative histopathology image translation at a mid-sized nci cancer center. *bioRxiv*, 2020.

[154] Ryoichi Koga, Noriaki Hashimoto, Tatsuya Yokota, Masato Nakaguro, Kei Kohno, Shigeo Nakamura, Ichiro Takeuchi, and Hidekata Hontani. Stain transfer for automatic annotation of malignant lymphoma regions in H&E stained whole slide histopathology images. In *International Forum on Medical Imaging in Asia*, volume 11792, pages 151–156. International Society for Optics and Photonics, 2021.

[155] Ying-Chih Lo, I-Fang Chung, Shin-Ning Guo, Mei-Chin Wen, and Chia-Feng Juang. Cycle-consistent GAN-based stain translation of renal pathology images with glomerulus detection application. *Applied Soft Computing*, 98: 106822, 2021.

[156] Caner Mercan, GCAM Mooij, David Tellez, Johannes Lotz, Nick Weiss, Marcel van Gerven, and Francesco Ciompi. Virtual staining for mitosis detection in breast histopathology. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1770–1774, 2020.

[157] Ansh Kapil, Tobias Wiestler, Simon Lanzmich, Abraham Silva, Keith Steele, Marlon Rebelatto, Guenter Schmidt, and Nicolas Brieu. DASGAN - joint domain adaptation and segmentation for the analysis of epithelial regions in histopathology pd-l1 images. In *MICCAI 2019 Workshop COMPAY*, 2019.

[158] Nassim Bouteldja, Barbara Mara Klinkhammer, Tarek Schlaich, Peter Boor, and Dorit Merhof. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. *arXiv preprint arXiv:2112.08837*, 2021.

[159] Fuyong Xing, Tell Bennett, and Debashis Ghosh. Adversarial domain adaptation and pseudo-labeling for cross-modality microscopy image quantification.

In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 740–749, 2019.

[160] Wouter Bulten and Geert Litjens. Unsupervised prostate cancer detection on H&E using convolutional adversarial autoencoders. In *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[161] Ke Mei, Chuang Zhu, Lei Jiang, Jun Liu, and Yuanyuan Qiao. Cross-stained segmentation from renal biopsy images using multi-level adversarial learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1424–1428, 2020.

[162] Xianxu Hou, Jingxin Liu, Bolei Xu, Bozhi Liu, Xin Chen, Mohammad Ilyas, Ian Ellis, Jon Garibaldi, and Guoping Qiu. Dual adaptive pyramid network for cross-stain histopathology image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 101–109, 2019.

[163] Joshua J. Levy, Nasim Azizgolshani, Michael J. Andersen, Arief Suriawinata, Xiaoying Liu, Mikhail Lisovsky, Bing Ren, Carly A. Bobak, Brock C. Christensen, and Louis J. Vaickus. A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies. *Modern Pathology*, 34(4):808–822, 2021.

[164] Amal Lahiani, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. Perceptual embedding consistency for seamless reconstruction of tilewise style transfer. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 568–576, 2019.

[165] Bingzhe Wu, Xiaolu Zhang, Shiwan Zhao, Lingxi Xie, C. Zeng, Zhihong Liu, and Guangyu Sun. G2c: A generator-to-classifier framework integrating multi-stained visual cues for pathological glomerulus classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1214–1221, 2019.

[166] Odyssee Merveille, Thomas Lampert, Jessica Schmitz, Germain Forestier, Friedrich Feuerhake, and Cédric Wemmert. An automatic framework for fusing information from differently stained consecutive digital whole slide images: A case study in renal histology. *Computer Methods and Programs in Biomedicine*, 208:106157, 2021.

[167] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in H&E breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136, 2018.

[168] Yuxin Cui, Guiying Zhang, Zhonghao Liu, Zheng Xiong, and Jianjun Hu. A deep learning algorithm for one-step contour aware nuclei segmentation of

histopathology images. *Medical & Biological Engineering & Computing*, 57: 2027–2043, 2019.

[169] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 529–536, 2018.

[170] Zeeshan Nisar, Jelica Vasiljević, Pierre Gançarski, and Thomas Lampert. Towards measuring domain shift in histopathological stain translation in an unsupervised manner. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.

[171] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[172] Dina Bashkirova, Ben Usman, and Kate Saenko. Adversarial self-defense for cycle-consistent GANs. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 635–645, 2019.

[173] Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a master of steganography. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2017.

[174] Jelmer M Wolterink, Konstantinos Kamnitsas, Christian Ledig, and Ivana Išgum. Deep learning: Generative adversarial networks and adversarial methods. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, chapter 23, pages 547–574. Elsevier, 2020.

[175] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, page 448–456, 2015.

[176] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. In *Advances in NIPS 2016 Deep Learning Symposium*, 2016.

[177] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[178] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[179] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[180] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[181] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the Pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[182] Friedrich Feuerhake, Benedikt Volk, Christoph Ostertag, Freimut Jungling, Jan Kassubek, M Orszagh, and Martin Dichgans. Reversible coma with raised intracranial pressure: An unusual clinical manifestation of cadasil. *Acta Neuropathologica*, 103:188–92, 2002.

[183] Jacqueline Pettersen, Julia Keith, Fuqiang Gao, J. David Spence, and Sandra Black. Cadasil accelerated by acute hypotension: Arterial and venous contribution to leukoaraiosis. *Neurology*, 88(11):1077–1080, 2017.

[184] Miri Adler, Avi Mayo, Xu Zhou, Ruth A Franklin, Matthew L Meizlish, Ruslan Medzhitov, Stefan M Kallenberger, and Uri Alon. Principles of cell circuits for tissue repair and fibrosis. *iScience*, 23(2):100841, 2020.

[185] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29), 2018.

[186] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[187] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[188] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.

[189] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[190] Marc Aubreville, Nikolas Stathonikos, Christof A Bertram, Robert Klopleisch, Natalie ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A Donovan, Andreas Maier, et al. Mitosis domain generalization in histopathology images–the midog challenge. *arXiv preprint arXiv:2204.03742*, 2022.

[191] Eliott Brion, Jean Léger, Ana M Barragán-Montero, Nicolas Meert, John A Lee, and Benoit Macq. Domain adversarial networks and intensity-based data augmentation for male pelvic organ segmentation in cone beam ct. *Computers in Biology and Medicine*, 131:104269, 2021.

[192] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 642–659. Springer, 2020.

[193] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[194] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3681–3691, 2021.

[195] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10849–10859, 2021.

[196] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1905–1914, 2021.

[197] Amirsina Torfi, Edward A Fox, and Chandan K Reddy. Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences*, 586:485–500, 2022.

[198] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[199] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digital Medicine*, 3(1):1–13, 2020.

[200] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2018.

[201] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

[202] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[203] Mélanie Bernhardt, Charles Jones, and Ben Glocker. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, pages 1–2, 2022.

[204] Ziad Obermeyer and Sendhil Mullainathan. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 89–89, 2019.

[205] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[206] Jakub Konečnỳ, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS) (workshop)*, 2015.

[207] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1):1–7, 2020.

[208] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *Proceedings of the European Symposium on Research in Computer Security*, pages 480–501, 2020.

[209] Komal Mariam, Osama Mohammed Afzal, Wajahat Hussain, Muhammad Umar Javed, Amber Kiyani, Nasir Rajpoot, Syed Ali Khurram, and Hassan Aqeel Khan. On smart gaze based annotation of histopathology images for training of deep convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[210] Joseph N Stember, Haydar Celik, E Krupinski, Peter D Chang, Simukayi Mutasa, Bradford J Wood, A Lignelli, Gul Moonis, LH Schwartz, Sachin Jambawalikar, et al. Eye tracking for deep learning segmentation using convolutional neural networks. *Journal of Digital Imaging*, 32(4):597–604, 2019.

[211] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 2022.

[212] Kurt E. Johnson. *Histology and Cell Biology*. The National medical series for independent study. Harwal Pub. Co., 1991. ISBN 978-0683062106.

[213] Michael Titford. The long history of hematoxylin. *Biotechnic & Histochemistry*, 80(2):73–78, 2005.

[214] Jingxi Li, Jason Garfinkel, Xiaoran Zhang, Di Wu, Yijie Zhang, Kevin De Haan, Hongda Wang, Tairan Liu, Bijie Bai, Yair Rivenson, et al. Biopsy-free in vivo virtual histology of skin using deep learning. *Light: Science & Applications*, 10(1):1–22, 2021.

[215] Lorraine C Racusen, Kim Solez, Robert B Colvin, Stephen M Bonsib, Maria C Castro, Tito Cavallo, Byron P Croker, A Jake Demetris, Cynthia B Drachenberg, Agnes B Fogo, et al. The banff 97 working classification of renal allograft pathology. *Kidney International*, 55(2):713–723, 1999.

[216] Raphael Rubin, David S Strayer, Emanuel Rubin, et al. *Rubin's Pathology : Clinicopathologic Foundations of Medicine*. Seventh edition. edition, 2015. ISBN 1-4511-8390-9.

[217] Nathan R Hill, Samuel T Fatoba, Jason L Oke, Jennifer A Hirst, Christopher A O'Callaghan, Daniel S Lasserson, and FD Richard Hobbs. Global prevalence of chronic kidney disease–a systematic review and meta-analysis. *PloS one*, 11(7):e0158765, 2016.

[218] Michael Abecassis, Stephen T. Bartlett, Allan J. Collins, Connie L. Davis, Francis L. Delmonico, John J. Friedewald, et al. Kidney transplantation as primary therapy for end-stage renal disease: A national kidney foundation/kidney disease outcomes quality initiative (NKF/KDOQI) conference. *Clinical Journal of the American Society of Nephrology*, 3(2):471–480, 2008.

[219] Mohamed Hassanein and Joshua J Augustine. Chronic kidney transplant rejection. In *StatPearls [Internet]*. 2021.

[220] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2107–2116, 2017.

[221] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[222] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1747–1756, 2016.

[223] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019.

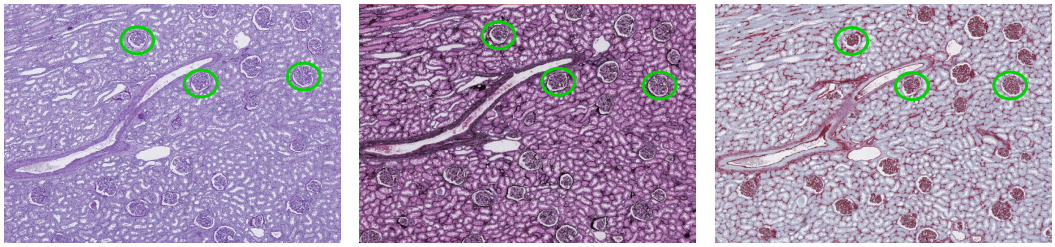
Résumé

Problématique

La révolution de l'apprentissage profond [1] ouvre la porte à des applications remarquables de l'intelligence artificielle dans le domaine médical. De nombreuses tâches cliniques quotidiennes ont un grand potentiel d'automatisation, ce qui a déclenché une grande quantité de travaux de recherche sur le sujet [2–4]. Dans un tel environnement, l'histopathologie numérique ne fait pas exception. Cependant, comme elle repose sur un processus de coloration des tissus étudiés, le résultat est sujet à de fortes variations en raison des différences dans la préparation des tissus et le protocole même de coloration. Chaque coloration met en évidence des structures spécifiques dans le tissu, et la variance introduite par les différents protocoles n'est pas seulement visuelle. Ces variations représentent une source de décalage de domaine et, en tant que telles, affectent considérablement les solutions basées sur l'apprentissage profond dans la pratique. Cela devient plus évident lorsqu'une tâche à accomplir s'attaque à des problèmes liés à des structures visibles sous plusieurs colorations. Récemment, les réseaux génératifs adversaires, *Generative Adversarial Networks* (GANs) [5], apportent de nouvelles opportunités pour l'apprentissage profond dans le domaine de l'histopathologie numérique. Cette thèse contribue à comprendre ce que les GANs peuvent et ne peuvent pas faire dans le domaine de l'histopathologie numérique, quand des résultats visuellement impressionnants sont dignes de confiance et quand ils sont trompeurs. La thèse étudie également les façons dont les GANs peuvent être utilisés pour construire des modèles d'apprentissage profond plus robustes.

Contexte

L'examen histopathologique commence par le prélèvement physique d'un échantillon de tissu dans le corps par biopsie ou chirurgie. Pour être examiné au microscope, l'échantillon subit plusieurs étapes de préparation, dont le processus important de coloration. La coloration introduit chimiquement un contraste dans les sections de tissu, rendant visibles des structures ou des cellules spécifiques, permettant ainsi leur analyse microscopique. Les colorants sont conçus pour être sélectifs ; ainsi, chacun met en évidence différentes structures tissulaires, ce qui permet des analyses différentes. Dans la figure 3 sont donnés des exemples de trois colorations pour le cas



(a) Periodic Acid Schiff (PAS) (b) Jones H&E (c) CD34

Figure 3: Un exemple de trois coupes consécutives d'un échantillon issu d'une néphrectomie, teintés avec trois colorants différents. Chaque coloration fournit des informations différentes sur le tissu, mais certaines structures communes, comme les glomérules, sont visibles dans toutes les colorations (cercles verts).

de la pathologie rénale. L'examen effectué par les pathologistes consiste en grande partie à reconnaître des structures spécifiques, par exemple les glomérules dans le rein (cercles verts de la figure 3) et à évaluer leur état de santé. En fonction de l'analyse, des colorations spécifiques peuvent être nécessaires car chaque coloration fournit des informations différentes. Ainsi, une pratique courante en histologie consiste à colorer différemment plusieurs coupes de la même biopsie et à les analyser ensemble. Par exemple, dans le cas de la pathologie rénale, les glomérules sont visibles sous plusieurs colorations et, comme l'illustre la figure 3, bien que différentes parties des glomérules soient mises en évidence dans chaque coloration, les experts peuvent les détecter quelle que soit la coloration. Par conséquent, les solutions automatiques de détection doivent être invariantes à ces différences de coloration.

Cependant, le processus de coloration est sujet à une grande variabilité [7] en raison des différences dans la préparation des tissus (temps d'exposition, fixation des tissus, épaisseur des sections, etc.) et dans le protocole de coloration. Par conséquent, lors du développement d'une solution invariante à la coloration, deux sources principales de variation doivent être prises en compte : la variation intra-coloration, qui est la variation de l'apparence d'une même coloration (par exemple, due aux procédures de différents laboratoires) et la variation inter-coloration, qui est la variation de l'apparence de différentes coloration, comme illustré dans la figure 4. Ces différences affectent considérablement les systèmes automatiques [8] car elles représentent une source de décalage de domaine [9]. Un pathologiste est capable de corriger ces variations grâce à son expérience contrairement aux algorithmes actuels basés sur l'apprentissage profond.

Capables de générer des échantillons à partir de distributions de données complexes, les GAN ont un grand potentiel pour surmonter certaines limitations de l'utilisation de l'apprentissage profond en histopathologie numérique. Un domaine d'application particulièrement prometteur est le transfert de couleurs, qui permet la re-coloration virtuelle d'une image histopathologique, c'est-à-dire la modification de son apparence pour donner l'impression qu'elle a été colorée avec une autre coloration ou une variation de coloration. Ainsi, le transfert de couleurs basé sur les GAN peut réduire le décalage de domaine introduit par la variation inter et intra-colorations dans l'espace des pixels. En outre, le transfert de couleurs peut être

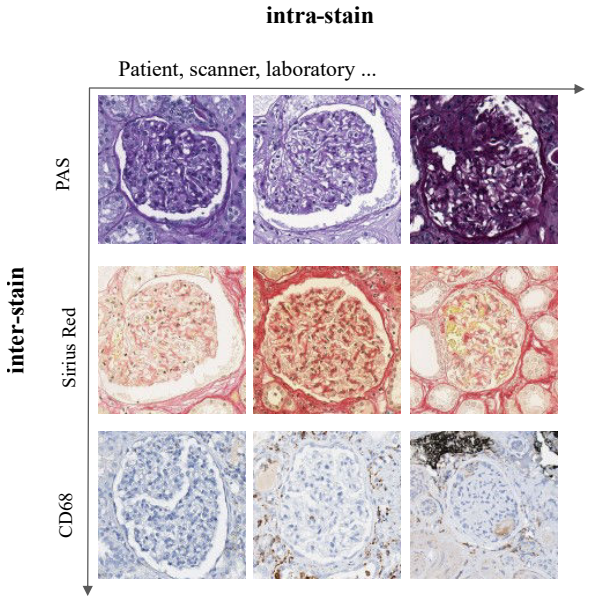


Figure 4: Variabilité des colorations : variations intra et inter-colorations sur des exemples de tissus rénaux. Chaque ligne contient des exemples colorés avec la même coloration.

utilisé pour créer des collections d'ensembles de données artificielles, ce qui pourrait résoudre les problèmes de confidentialité des données médicales.

Axes de recherche

Deux axes de recherche principaux ont été identifiés :

- Réseaux génératifs adversaires pour le transfert de couleurs : cet axe consiste à appliquer les GAN au transfert de couleurs, c'est-à-dire à la modification de l'apparence d'une image colorée avec la coloration A pour qu'elle ressemble à une image colorée avec la coloration B. Le transfert obtenu doit être réaliste : l'image microscopique générée, en l'absence d'informations spécifiques sur le patient telles que la maladie sous-jacente, doit être visuellement réaliste pour un pathologiste, à la fois au niveau de la coloration et de l'aspect morphologique des composants tissulaires. La thèse propose plusieurs façons d'obtenir des translations visuellement convaincantes et donne les limites de ces approches tant du point de vue du diagnostic que de leur application dans le domaine de la vision par ordinateur.

- Réseaux génératifs adversaires pour la construction de modèles robustes : cet axe de la thèse étudie comment les GANs peuvent être utilisés pour construire de meilleurs modèles d'apprentissage profond plus robustes et capables de résoudre le même problème dans des images provenant de plusieurs colorations (segmentation de structures à grande échelle, par exemple). L'apprentissage est effectué à l'aide d'un nombre limité d'annotations provenant d'une seule modalité de coloration, et l'objectif de la solution obtenue est de généraliser

ce modèle à de multiples colorations, même à celles qui n'ont jamais été vues durant l'apprentissage.

Méthodes

Réseaux génératifs adversaires - *Generative Adversarial Netowrks* (GANs)

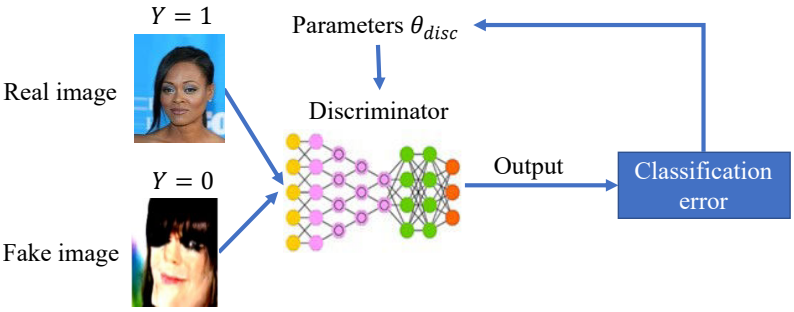
Les réseaux génératifs adversaires (GAN) [5] appartiennent à la classe des modèles génératifs implicites. Ils sont optimisés pour échantillonner la distribution des données d'un ensemble d'apprentissage par un jeu contradictoire à deux joueurs. Les joueurs sont généralement représentés par des réseaux neuronaux appelés générateur (G) et discriminateur (D). Le générateur effectue l'échantillonnage, c'est-à-dire qu'il est optimisé pour générer de nouvelles données, tandis que le discriminateur agit comme un classificateur, apprenant à distinguer les échantillons de données générés des échantillons réels. L'apprentissage de ces deux modèles est un jeu compétitif (contradictoire) puisque les objectifs des joueurs sont opposés. Le discriminateur vise à distinguer le mieux possible les données réelles des données générées, tandis que le générateur vise à créer des échantillons qui ne se distinguent pas des données réelles. Le résultat optimal d'un tel jeu est un équilibre de Nash, où le générateur produit des échantillons indiscernables (du point de vue du discriminateur) des données réelles. La figure 5 illustre graphiquement le jeu contradictoire et les joueurs pour la tâche de génération de visages.

Les réseaux génératifs adversaires ont suscité beaucoup d'attention depuis leur introduction en 2014, car ils permettent d'échantillonner à partir de distributions de données très complexes, ce qui a considérablement augmenté leurs domaines d'application. Un domaine d'application particulièrement intéressant du point de vue de l'histopathologie numérique est la traduction d'image à image. Cette tâche peut être définie comme la conversion d'une image x_A du domaine A en une image $\hat{x}_B$ du domaine B, en prenant le style du domaine B et en préservant le contenu de l'image x_A. Les réseaux adversaires conditionnels sont parfaitement adaptés à cela.

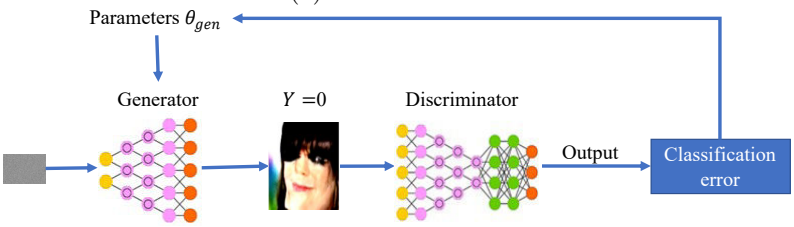
Transfert de couleurs

Dans le contexte de l'histopathologie numérique, des méthodes de traduction d'image à image peuvent être utilisées pour obtenir une coloration virtuelle. La coloration virtuelle consiste à modifier artificiellement l'apparence d'une image histopathologique après son acquisition, par exemple en faisant en sorte qu'une image initialement colorée par la coloration A ressemble à une image colorée par la coloration B de manière réaliste. *Le terme **'réaliste'** fait référence au fait qu'une image histologique seule, sans connaissance des sections adjacentes traitées avec d'autres modalités de coloration, et en l'absence d'informations spécifiques au patient telles que la maladie sous-jacente, semble visuellement correcte pour un expert à la fois concernant les caractéristiques de coloration et l'aspect morphologique des composants tissulaires.* Des architectures pour la traduction d'image à image, telles

[17] À titre d'illustration, la fausse image est générée à l'aide d'un modèle ProgressiveGAN pré-entraîné ; la vraie image est tirée du jeu de données Celeb-A [11].



(a) Discriminateur



(b) Générateur

Figure 5: Illustration de l'entraînement contradictoire a) le discriminateur est entraîné à distinguer les échantillons de données réels des échantillons générés ; b) le générateur est entraîné à l'aide des informations fournies par le discriminateur, afin de produire des échantillons impossibles à distinguer des données réelles[17].

que CycleGAN (voir figure 6), ont un grand potentiel pour être appliquées à la coloration virtuelle.

Le modèles est constitué de deux générateurs qui effectuent le tranfert de couleurs : $G_{AB} : A \to B$ pour le transfert de A vers B et $G_{BA} : B \to A$ pour le transfert de B vers A; et de deux discriminateurs D_A and D_B. L'objectif de D_A est de distinguer les images réelles du domaine A de celles translatées de B vers A; tandis que D_B a pour objectif de distinguer les images réelles du domaine B de celles translatées de A vers B. Une fois entraîné, le modèle est capable de vers des transferts $T_{A\to B}$ et $T_{B\to A}$ entre deux colorations A et B, en utilisant le générateur correspondant.

Invariance colorimétrique

Le modèle CycleGAN est capable d'obtenir des translations réalistes et plausibles entre différentes colorations, ce qui va être utile pour proposer une solution invariante à la coloration pour la segmentation des glomérules — Unsupervised Domain Augmentation using on Generative Adversarial Networks (UDA-GAN). UDA-GAN est une approche générale pour la formation de réseaux neuronaux convolutifs (CNN) invariants aux colorations pour une tâche spécifique. Après l'entraînement, le modèle est capable d'exécuter une tâche donnée dans différentes colorations, potentiellement inconnues pendant la période d'apprentissage. On suppose que des WSI annotées sont disponibles pour une coloration A alors que les WSI des autres colorations $B_1, B_2, \ldots, B_N$ ne sont pas annotées. L'objectif est d'augmenter la variabilité de

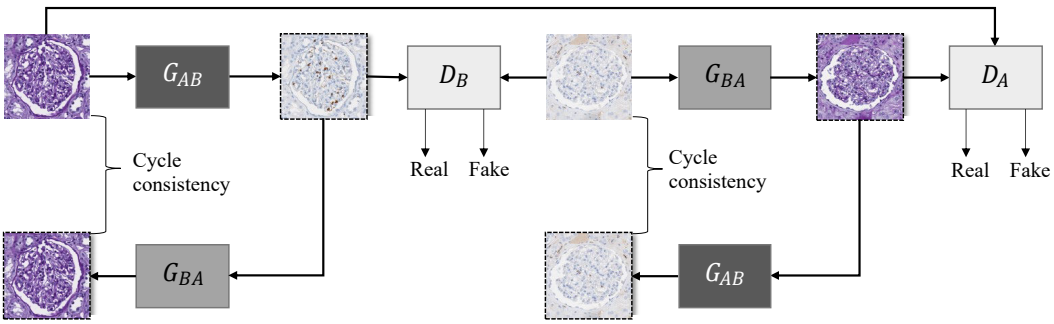


Figure 6: Architecture d'un CycleGAN pour la coloration virtuelle en histopathologie.

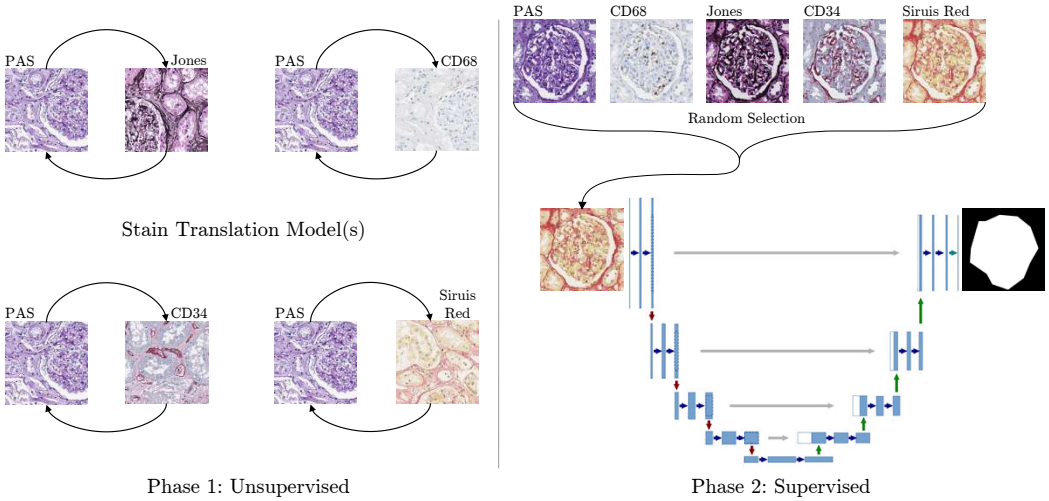


Figure 7: Schéma général de l'approche proposée. Phase 1, les modèles de translation sont entraînés pour traduire les images du domaine source vers les domaines cibles ; Phase 2, les patchs du domaine source sont transférés de manière aléatoire vers les domaines cibles pendant l'entraînement (image U-Net tirée de [13]).

l'ensemble d'apprentissage (annoté) en le traduisant de manière aléatoire dans les domaines non annotés (y compris le domaine original annoté). L'architecture globale de la méthode proposée est présentée dans la figure 7.

La méthode se décompose en deux phases :

a) (non supervisé) Modèle pour le transfert de couleurs – Afin d'obtenir des translations réalistes de la coloration annotée A vers des colorations non annotées $B_1, B_2, \ldots, B_N$, un modèle de translation d'image à image non supervisé basé sur un GAN est utilisé. Les modèles CycleGAN [12] et StarGAN [14] sont considérés pour cette tâche.

b) (supervisé) Modèle lié à la tâche (modèle de segmentation) – ce modèle est entraîné sur les données annotées après avoir été traduites en une coloration aléatoire non annotée. Comme la traduction ne modifie pas la structure globale de l'image, la vérité terrain du domaine d'origine reste valide. Ainsi,

au cours de l'apprentissage, divers échantillons annotés de toutes les colorations disponibles sont présentés au modèle, le forçant à apprendre des caractéristiques invariantes de la coloration. Une fois le modèle de segmentation entraîné, il peut être appliqué directement aux colorations non annotées, sans autre translation.

Transfert de coloration et invariance – une solution complète

Indépendamment du réalisme des translations obtenues, le modèle CycleGAN est bidirectionnel et ne permet donc qu'une coloration virtuelle entre deux colorations sélectionnées. De nombreuses méthodes de translation d'image à image non appariées multi-domaines existent [14–16]. Cependant, contrairement au modèle CycleGAN, elles ne sont pas directement applicables au transfert de couleurs. Ces architectures avancées comme StarGANv2 [15] incluent des modules additionnels qui permettent au générateur d'effectuer des changements importants sur une image tout en garantissant la réversibilité à travers une contrainte de cohérence cyclique. Cependant, lorsqu'il s'agit de transfert de couleurs, ces modifications pourraient conduire à la suppression/invention de la structure des tissus, empêchant l'application de ces translations à des tâches médicales ou de vision par ordinateur. Ainsi, l'obtention d'un modèle de transfert multi-colorations n'est pas simple. Dans cette thèse, une extension du modèle StarGANv2, appelée HistoStarGAN, est proposée. L'approche présentée est capable d'effectuer des colorations virtuelles réalistes et diverses tout en préservant la structure d'intérêt pendant le processus de traduction. De plus, le modèle fournit une segmentation invariante de la structure sélectionnée.

L'architecture du modèle HistoStarGAN est présentée dans la Figure 8. Le modèle contient cinq modules principaux : le générateur (G), le discriminateur (D), le réseau de mise en correspondance (F) et l'encodeur de style (E) (originellement de l'approche StarGANv2) en plus d'un réseau de segmentation (S) (en rouge dans la Figure 8) qui est relié au générateur. L'architecture de StarGANv2 (composée des modules G, D, E, F) est déjà suffisante pour effectuer la translation entre différentes colorations, cependant, les structures anatomiques importantes sont perturbés. Par exemple, les glomérules peuvent être supprimés. Le but du module de segmentation est d'éviter que les structures d'intérêt disparaissent pendant la translation.

Le générateur G est un réseau de type encodeur-décodeur dans lequel une couche de normalisation dans l'encodeur supprime les caractéristiques spécifiques à la coloration, tandis que la normalisation adaptative dans le décodeur injecte des informations spécifiques à la coloration. Ainsi, l'espace latent du générateur extrait une représentation des données invariante par rapport aux colorations. Un module de segmentation est relié à l'espace latent et vise à segmenter la structure d'intérêt en utilisant la représentation invariante de la coloration. Entraînée de bout en bout avec d'autres modules, la branche de segmentation force la préservation des structures d'intérêt dans l'espace latent pendant le processus de translation.

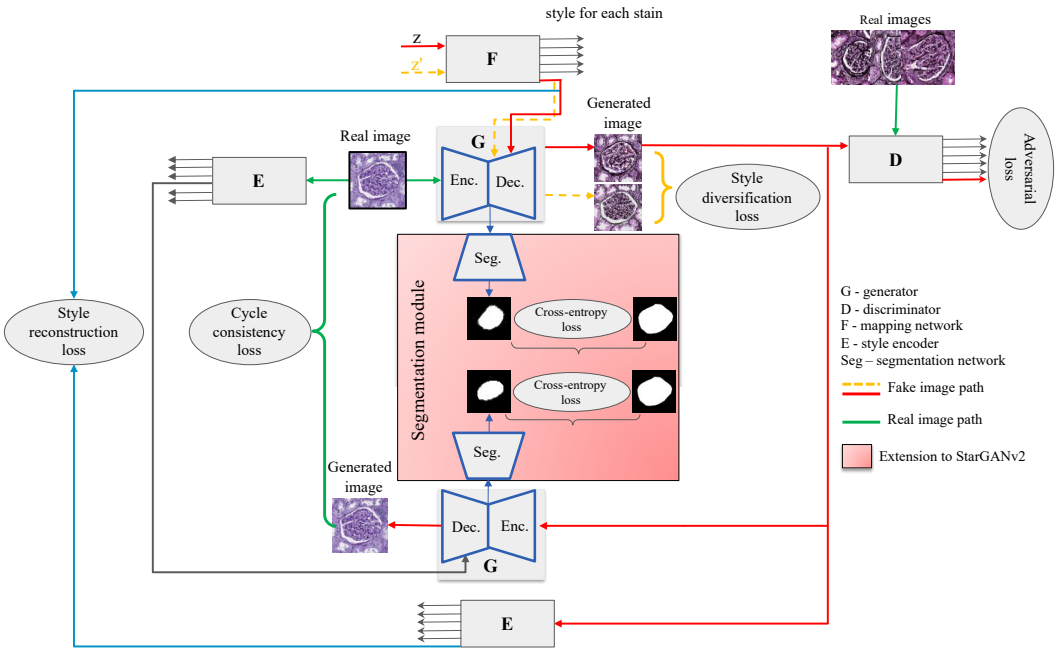


Figure 8: HistoStarGAN — un modèle entraînable de bout en bout pour le transfert simultané de couleurs et une segmentation invariante des colorations. Le bloc rouge indique la différence par rapport au modèle StarGANv2 [15].

Résultats

Transfert de coloration basé sur un CycleGAN

Les résultats des translations de CycleGAN $T_{PAS \to X}$ et $T_{X \to PAS}$, où $X \in \{\text{Jones H\&E}, \text{Sirius Red}, \text{CD68}, \text{CD34}\}$, sont donnés dans la figure 9. Toutes les traductions semblent plausibles, comme le confirment les pathologistes. Cependant, un tel transfert de couleurs n'est pas capable de réduire le décalage de domaine dans toutes les colorations de manière égale. On peut observer que certaines translations sont plus difficiles que d'autres, ce qui influence grandement la qualité des résultats.

Dans l'ensemble, la conception architecturale et la procédure d'apprentissage de cette méthode permettent le transfert de couleurs de telle sorte que les structures internes ne sont pas affectées, par exemple, les glomérules restent dans la même position (forme, orientation, etc.) avant et après la translation. La méthode est générale, elle n'est pas liée à une coloration spécifique, et peut donc être appliquée pour la translation entre n'importe quelle paire de colorations. Cependant, la conception architecturale et la procédure d'apprentissage qui conduisent à un modèle déterministe de transfert des couleurs (c'est-à-dire une correspondance une-à-une entre les colorations), entraînent des limitations supplémentaires liées à l'application pratique et à l'évaluation, telles que :

Qualité de la translation : les translations obtenues peuvent coder des informations supplémentaires afin d'effectuer une mise en correspondance déterministe. Cela peut affecter la capacité des translations à réduire le décalage de do-

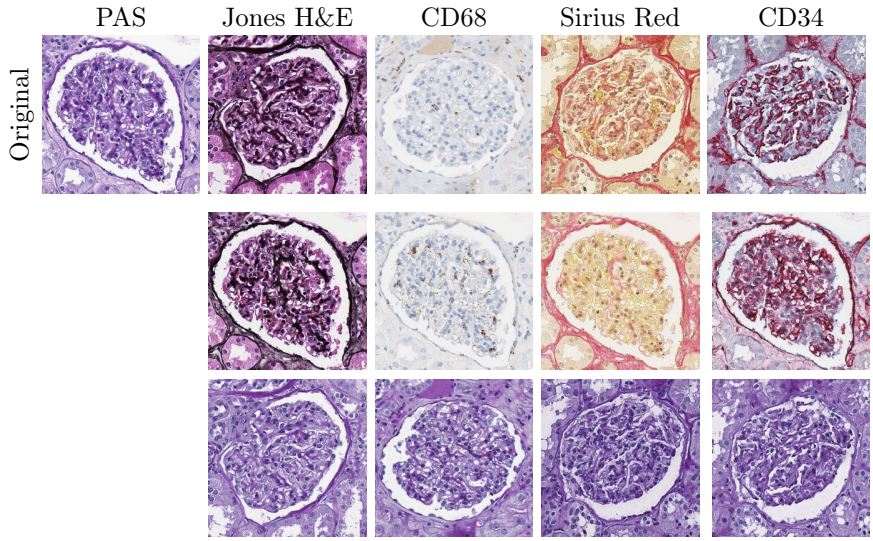


Figure 9: Transfert de couleurs obtenu avec les modèles CycleGAN. La première rangée contient les images réelles de chaque coloration. La deuxième rangée représente les translations $T_{PAS \to X}$ d'une image PAS vers la coloration cible. La dernière rangée représente les translations $T_{X \to PAS}$ d'images cibles réelles vers la coloration PAS.

maine entre les colorations.

Stabilité et reproductibilité de l'apprentissage : comme il n'y a pas de critère d'arrêt explicite, on peut arrêter l'apprentissage lorsque des traductions réalistes sont obtenues. Cependant, le mapping déterministe peut être différent à différentes époques ou répétitions de l'entraînement, ce qui pourrait conduire à des conclusions trompeuses sur la qualité des translations obtenues.

Généralisation : différentes combinaisons de couleurs pourraient coder des informations différentes, ce qui donnerait lieu à des conclusions ambiguës liées à l'application des modèles de transfert de couleurs.

Tous ces aspects sont étudiés plus en détail dans le chapitre 3 de la thèse. Il s'avère que le potentiel du transfert de couleurs basé sur le CycleGAN pour réduire le décalage de domaine introduit par la variation inter-coloration peut être grandement influencé par son architecture, c'est-à-dire que l'on peut utiliser beaucoup de couches de normalisation dans le modèle sans affecter le réalisme de la translation mais en affectant grandement sa capacité à réduire le décalage de domaine. Ainsi, l'inspection visuelle, qui est l'une des approches largement utilisées lors de l'évaluation du transfert de couleurs virtuelles, n'est pas suffisant. De plus, on constate que le bruit existe bel et bien dans les translations. Il peut être perturbé de manière à générer des échantillons divers, affectant la présence de marqueurs importants pour le diagnostic, par exemple les macrophages. Ces résultats sont ensuite utilisés pour proposer une méthode d'augmentation de données non supervisée, qui s'avère bénéfique pour l'apprentissage supervisé.

UDA-GAN

Le modèle UDA-GAN est capable d'atteindre une performance équivalente à la baseline dans trois des cinq colorations de test. Malgré le fait que le modèle ne voit que la coloration PAS (la seule coloration annotée dans ces expériences) 20% du temps pendant l'appretnissage, le modèle atteint une performance équivalente à la baseline dans ce domaine (source). Le modèle se rapproche également de la performance de base pour les colorations cibles Jones H&E et Sirius Red. Pour les colorations CD68 et CD34, le modèle atteint un score F_1 de $0,705$ et $0,799$, ce qui signifie qu'il apporte une amélioration de $11,9\%$ et 6% respectivement par rapport à la meilleure méthode suivante basée sur CycleGAN (MDS2). La performance moyenne sur les cinq colorations différentes montre que UDA-CGAN atteint un score F_1 moyen de $0,827$ ($0,808$ sans inclure la coloration PAS, afin d'être équitablement comparé aux approches MDS), tandis que MDS2, en tant que deuxième meilleure méthode, atteint un score F_1 de $0,748$. La plus grande différence relative est observée dans la coloration CD68, où l'amélioration globale est de $55,8\%$ par rapport à l'approche originale [17] et de $11,9\%$ par rapport à MDS2. En dehors de l'approche de base, UDA-GAN est la seule à obtenir des résultats acceptables dans cette coloration.

L'invariance du modèle est également démontrée par son application à deux nouvelles colorations — la coloration histologique H&E (une coloration générale non spécifique à une protéine particulière) et la coloration immunohistochimique CD3 (marqueur des cellules T). Ces colorations n'ont pas été observées lors de l'apprentissage du CycleGAN et de l'UDA-GAN. Les résultats obtenus confirment la capacité du réseau à réaliser une segmentation invariante aux colorations.

HistoStarGAN

L'entraînement d'HistoStarGAN aboutit à un modèle unique capable d'effectuer divers transferts de couleurs et une segmentation invariante aux colorations. De plus, en translatant une image dans son propre domaine, il est possible pour la première fois d'obtenir une normalisation des colorations. De plus, grâce à son codeur invariant, le modèle HistoStarGAN est capable d'effectuer la translation de couleurs et la segmentation de colorations non vues lors de l'apprentissage.

La capacité du modèle à effectuer la translation de couleurs multi-domaine (et les segmentations correspondantes) est illustrée dans la figure 11 ; la généralisation d'un transfert de couleurs à des colorations non vues dans la figure 12 ; et la normalisation de la coloration PAS dans la figure 13, dans laquelle l'ensemble de données AIDPATH [18] (disponibles publiquement) est traduit en plusieurs colorations.

Le modèle HistoStarGAN représente la première solution de bout en bout entraînable pour la normalisation simultanée des colorations, le transfert de couleurs et la segmentation invariante aux colorations. Pour la première fois, il est possible d'obtenir un transfert de couleurs hautement réaliste à partir de colorations non vues, sans aucune modification supplémentaire du modèle (par exemple, un fine tuning). De plus, le modèle définit de nouveaux résultats à l'état de l'art pour la segmentation invariante aux colorations, en se généralisant avec succès à six colorations inconnues. La solution proposée est générale et peut être étendue à de nouvelles colorations ou à de nouveaux cas d'utilisation.

En étant capable de générer diverses translations pour une entrée donnée, la

solution proposée offre un moyen de générer le premier ensemble de données créé artificiellement et entièrement annoté — KidneyArtPathology, qui est disponible publiquement pour encourager les recherches futures. En outre, le modèle est entraîné sur cinq colorations largement utilisées et des modèles pré-entraînés sont disponibles, ce qui permet d'augmenter les données hors ligne (par exemple, sur des ensembles de données privés) par le transfert des colorations utilisées pendant l'entraînement du modèle.

Contributions

Les contributions principales de cette thèse sont présentées sur la figure 10.

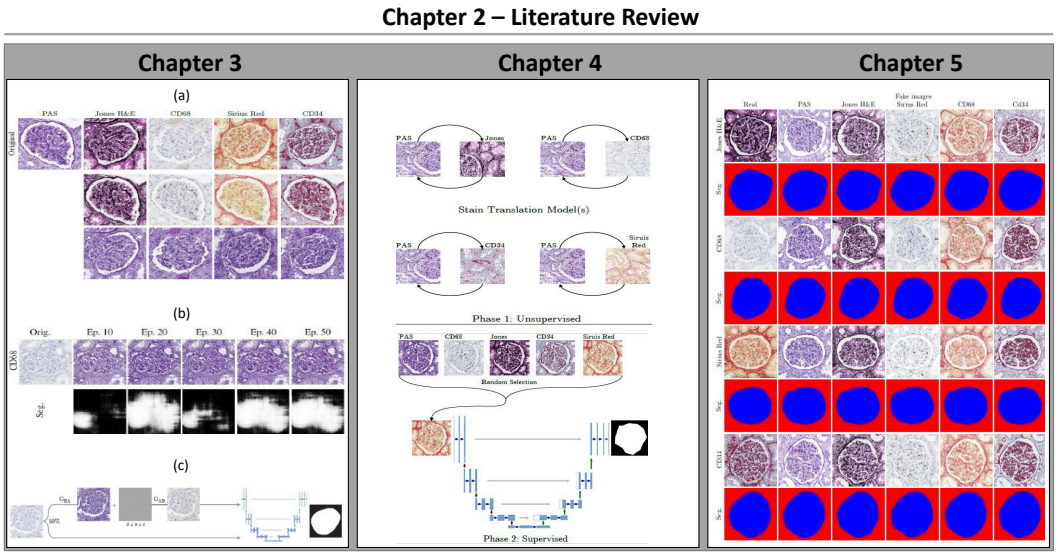


Figure 10: Un résumé visuel de la structure et des principales contributions présentées dans cette thèse.

Cette thèse, en parallèle avec d'autres auteurs [19, 20], a proposé pour la première fois l'utilisation d'une méthode de translation d'image à image basée sur un GAN pour le transfert de couleurs entre différentes modalités de coloration – la première contribution de la thèse, voir figure 10. Chapitre 3(a)[18]. Entretemps, l'approche basée sur CycleGAN a été établie comme une solution standard pour la coloration virtuelle en général et est largement adaptée au domaine de l'histopathologie numérique. Les travaux les plus importants sont classés et résumés dans le Chapitre 2 de cette thèse. Il est clair que la littérature s'est jusqu'à présent concentrée sur la normalisation des colorations, c'est-à-dire la standardisation de l'apparence des images histologiques dans une modalité de coloration, où les solutions basées sur CycleGAN sont les approches dominantes. Cependant, le transfert de couleurs entre différentes modalités de coloration, sur lequel cette thèse se concentre, est rarement abordé dans la littérature. La thèse révèle que le transfert entre

[18]L'ouvrage [19] a été publié en même temps que la préparation de la publication liée aux résultats basés sur CycleGAN présentés dans cette thèse. Après la publication de Gadermayr et al. [19], le travail a été étendu avec des analyses supplémentaires.

différentes colorations ouvre des questions plus complexes et présente des limitations spécifiques par rapport à la normalisation des colorations.

This thesis identifies that CycleGAN-based translations contain imperceptible information related to stain differences whose manipulation can modify the resulting translation in plausible way. As a consequence of this finding, the thesis proposes an unsupervised augmentation method that increases the robustness of deep-learning based solutions — the thesis' second contribution, see Figure 10 Chapter 3(c).

De plus, cette thèse à permis de découvrir et de démontrer la sensibilité des solutions basées sur CycleGAN à de petites modifications architecturales. Ces changements n'affectent pas nécessairement la qualité visuelle des translations obtenues mais influencent les conclusions générales relatives à l'utilité du transfert de couleurs tant du point de vue du diagnostic que de l'application — la troisième contribution de la thèse, voir figure 10 Chapitre 3(b)).

De plus, la thèse profite du réalisme des translations obtenues pour proposer la première solution qui permet d'avoir un modèle de segmentation robuste et invariant aux colorations pour la segmentation des glomérules dans des coupes histologiques rénales. Le modèle obtenu est capable de segmenter plusieurs colorations et est également capable de généraliser à certaines colorations non vues — la quatrième contribution de la thèse, voir figure 10 Chapitre 3(a) et Chapitre 4.

Les résultats et les conclusions des travaux menés durant cette thèse sont en outre utilisés pour proposer un modèle complet qui effectue simultanément le transfert de couleurs et la segmentation invariante aux colorations — la cinquième contribution de la thèse, voir la figure 10. Le modèle proposé est, pour la première fois, capable d'effectuer simultanément le transfert entre différentes colorations, la normalisation de couleurs dans une coloration et de généraliser le processus de translation à des colorations non vues. En outre, toutes les données générées (y compris les colorations originales et non vues) sont correctement segmentées grâce au module de segmentation invariant. Ces résultats permettent de générer le premier jeu de données entièrement annoté, créé artificiellement, qui sera bientôt mis à la disposition de la communauté pour faire progresser l'histopathologie numérique — la sixième et dernière contribution de la thèse.

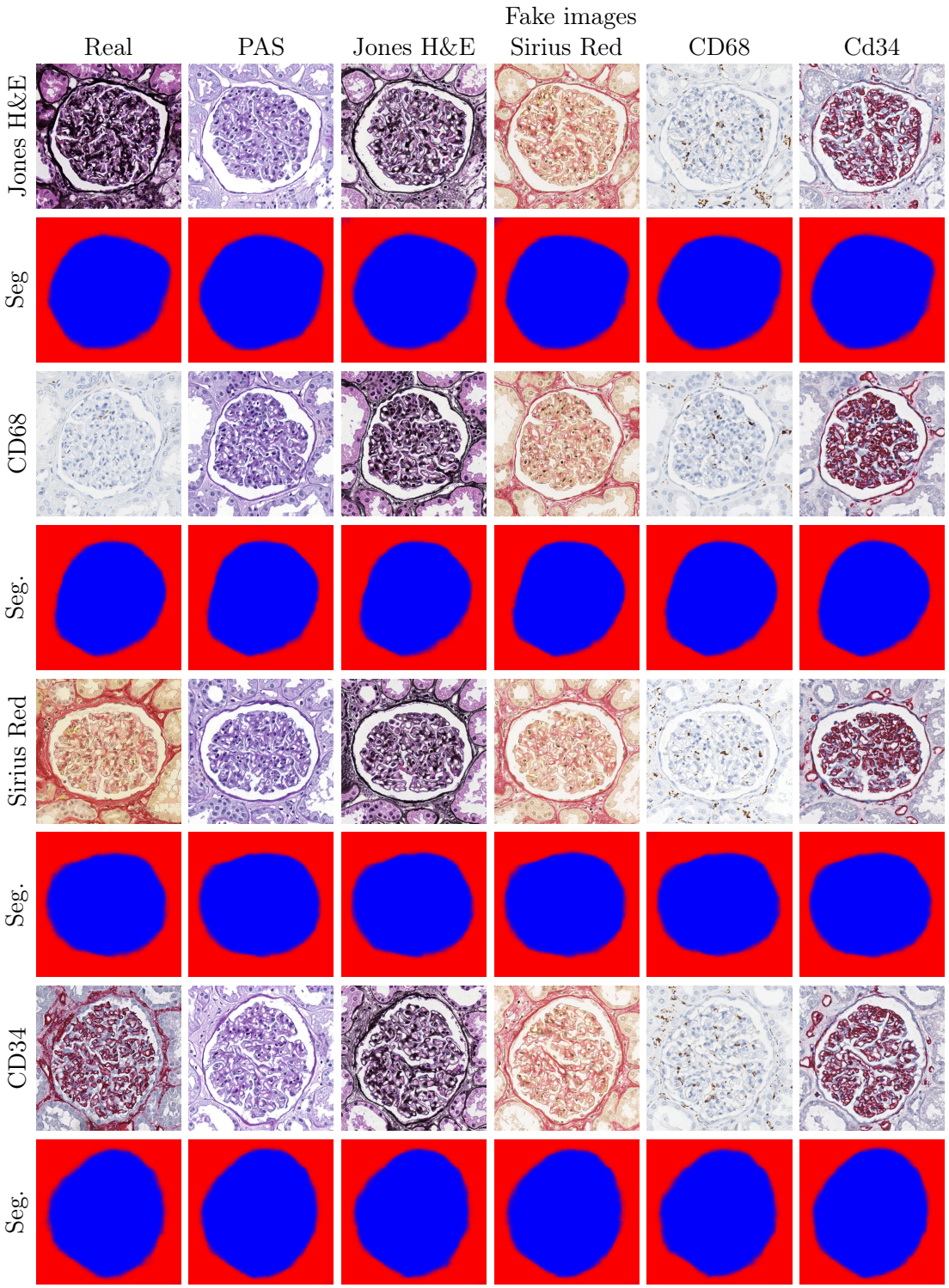


Figure 11: Translations d'HistoStarGAN entre différentes colorations avec les segmentations correspondantes. Chaque translation est obtenue en utilisant différents codes latents.

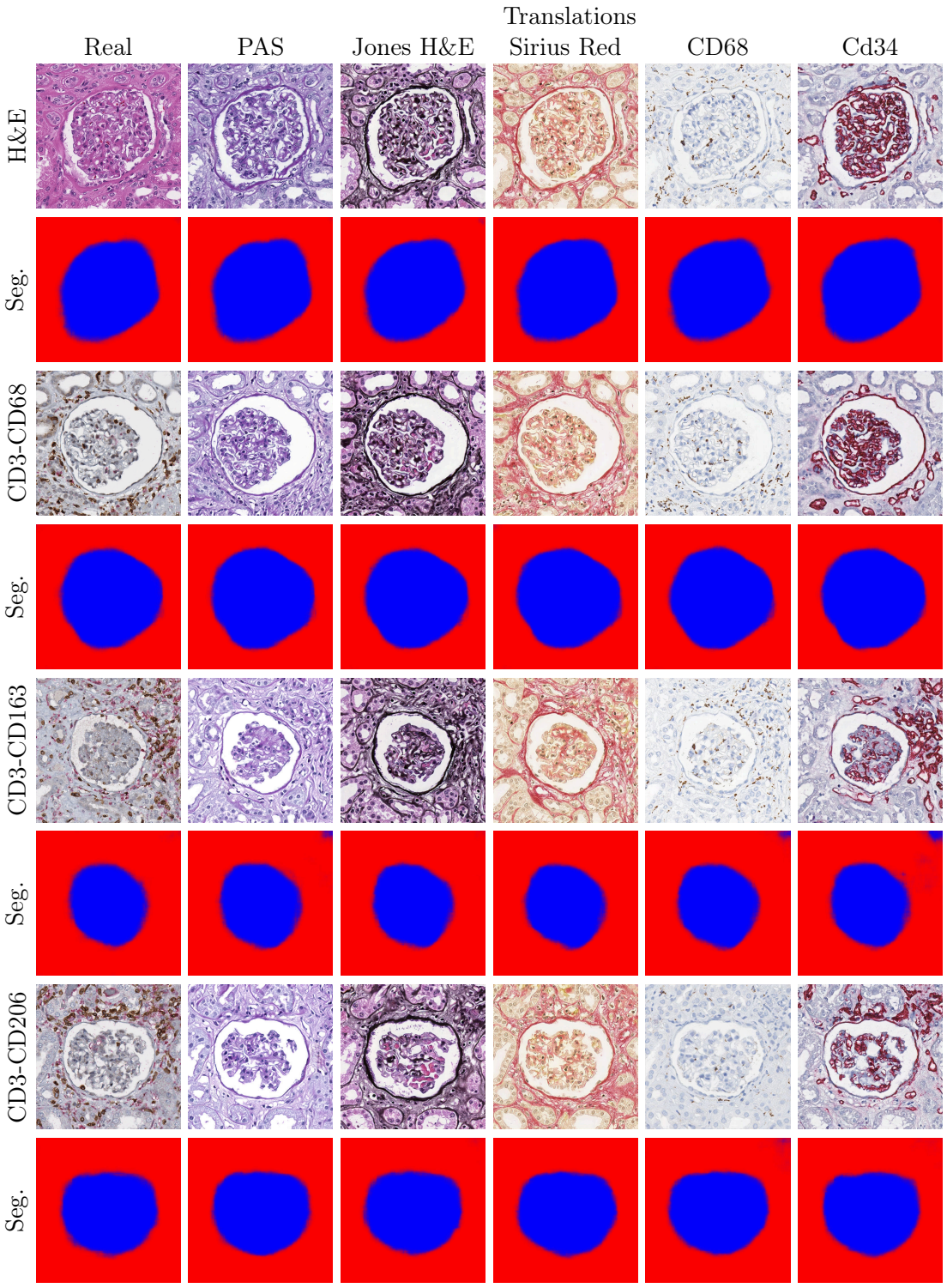


Figure 12: HistoStarGAN — généralisation du transfert de couleurs et de la segmentation à des colorations non vues.

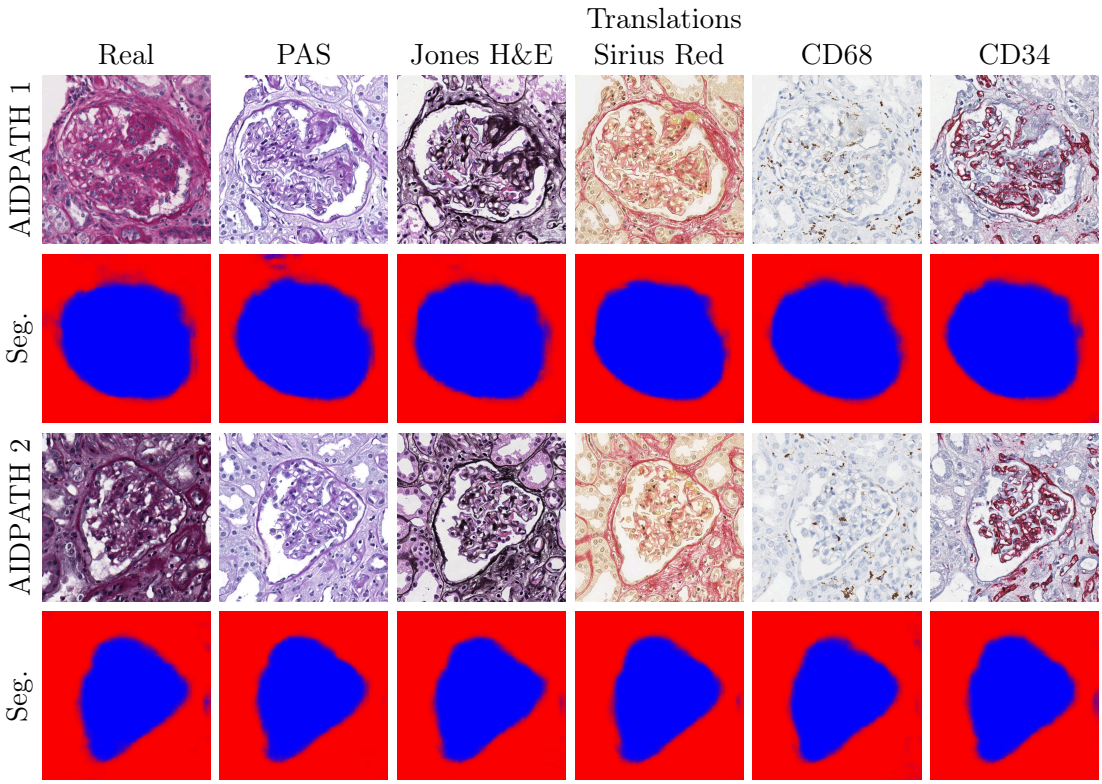


Figure 13: HistoStarGAN a été appliqué pour la normalisation des colorations, le transfert de couleurs et la segmentation des glomérules du jeu de données AIDPATH (basé sur PAS) disponible publiquement.

Jelica Zlatko VASILJEVIĆ

Generative Adversarial Networks in Digital Histopathology: Stain Transfer and Deep Learning Model Invariance to Stain Variation

Résumé

L'histopathologie numérique est un domaine d'innovation très riche, tant dans les applications cliniques que dans la recherche, où les solutions basées sur l'apprentissage profond connaissent un succès remarquable. Cependant, les méthodes actuelles d'apprentissage profond sont des approches gourmandes en données qui nécessitent d'énormes bases de données annotées pour obtenir des modèles performants. Or, le domaine médical est connu pour sa difficulté à obtenir des données et des annotations - la collecte de données relève d'une réglementation stricte et contraignante, tandis que seuls des experts peuvent effectuer des annotations de haute qualité, ce qui est un processus laborieux et coûteux. De plus, compte tenu des variations qui peuvent se produire en raison du processus et des protocoles de coloration, les données déjà collectées et annotées ne peuvent être réutilisées qu'avec un succès limité. Une telle variation de la coloration représente un changement de domaine et affecte considérablement les solutions basées sur l'apprentissage profond dans la pratique. Cela devient plus évident encore lorsque l'apprentissage se focalise sur des structures biologiques visibles avec plusieurs colorations, car les solutions développées en utilisant les données d'une coloration sont susceptibles d'échouer lorsqu'elles sont appliquées à une autre. Cette thèse étudie le potentiel des réseaux adversaires génératifs (GAN) dans deux directions pour résoudre ces problèmes - le transfert de colorations pour permettre la réutilisation de bases de données déjà disponibles et le développement de modèles invariants aux colorations qui réduiraient le besoin d'acquisition de données ou d'annotations supplémentaires. L'application principale de la thèse est la segmentation des glomérules en pathologie rénale avec de multiples colorations.

Résumé en anglais

Digital histopathology has become a rich area of innovation in both clinical application and research, where deep-learning-based solutions have remarkable success. However, current state-of-the-art deep learning methods are data-hungry approaches which require huge, annotated data collections to perform well. Nevertheless, the medical domain is known for its scarcity of data and annotations — collecting data falls under strict low regulations while experts only can perform high-quality annotations, which is a laborious and expensive process. Moreover, considering the variations that can occur due to the staining process and staining protocols, already collected and annotated datasets can only be reused with limited success. Such stain variation represents a source of domain shift and significantly affects deep learning-based solutions in practice. This becomes more evident when a deep learning task tackles problems related to structures visible under multiple stains as solutions developed using the data from one staining are likely to fail when applied to the other. This thesis investigates the potential of Generative Adversarial Networks (GANs) in two directions for addressing these problems — stain transfer to enable reusing already available data collections; and developing stain invariant solutions which would alleviate the need for additional data acquisition or annotations. The application focus of the thesis is glomeruli segmentation in renal pathology with multiple stainings.