Self-supervised learning in the presence of limited labelled data for digital histopathology

Zeeshan Nisar

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE MATHÉMATIQUES, SCIENCES DE L'INFORMATION ET DE L'INGÉNIEUR

Laboratoire ICube – UMR 7357

THÈSE présentée par :

[Zeeshan NISAR]

soutenue le : **24 Septembre 2024**

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

Discipline/ Spécialité : Informatique

Self-Supervised Learning in the Presence of Limited Labelled Data for Digital Histopathology

THÈSE dirigée par :

M. LAMPERT Thomas	Chair of Data Science and AI, Université de Strasbourg, France

RAPPORTEURS :

M. WALTER Thomas	Professeur, Ecole des Mines de Paris, France
M. CAMILLE Kurtz	Professeur, Université Paris Cité, France

AUTRES MEMBRES DU JURY :

M. WEMMERT Cédric	Professeur, Université de Strasbourg, France
M. IENCO Dino	Directeur de Recherche, UMR TETIS, France
Mme. TEMERINAC OTT Maja	Professeur, Fakultät Informatik, Furtwangen University, Germany

Zeeshan NISAR

Self-Supervised Learning in the Presence of Limited Labelled Data for Digital Histopathology

Résumé

Un défi majeur dans l'application de l'apprentissage profond à l'histopathologie réside dans la variation des colorations, à la fois inter et intra-coloration. Les modèles d'apprentissage profond entraînés sur une seule coloration (ou domaine) échouent souvent sur d'autres, même pour la même tâche (par exemple, la segmentation des glomérules rénaux). L'annotation de chaque coloration est coûteuse et chronophage, ce qui pousse les chercheurs à explorer des méthodes de transfert de coloration basées sur l'adaptation de domaine. Celles-ci visent à réaliser une segmentation multi-coloration en utilisant des annotations d'une seule coloration, mais sont limitées par l'introduction d'un décalage de domaine, réduisant ainsi les performances. La détection et la quantification de ce décalage sont essentielles. Cette thèse se concentre sur des méthodes non supervisées pour développer une métrique de détection du décalage et propose une approche de transfert de coloration pour le minimiser. Bien que ces algorithmes réduisent le besoin d'annotations, ils peuvent être limités pour certains tissus. Cette thèse propose donc une amélioration via l'auto-supervision. Bien que cette thèse se soit concentrée sur l'application de la segmentation des glomérules rénaux, les méthodes proposées sont conçues pour être applicables à d'autres tâches et domaines en histopathologie, y compris l'imagerie médicale et la vision par ordinateur.

Résumé en anglais

A key challenge in applying deep learning to histopathology is the variation in stainings, both inter and intra-stain. Deep learning models trained on one stain (or domain) often fail on others, even for the same task (e.g., kidney glomeruli segmentation). Labelling each stain is expensive and time-consuming, prompting researchers to explore domain adaptation based stain-transfer methods. These aim to perform multi-stain segmentation using labels from only one stain but are limited by the introduction of domain shift, reducing performance. Detecting and quantifying this domain shift is important. This thesis focuses on unsupervised methods to develop a metric for detecting domain shift and proposes a novel stain-transfer approach to minimise it. While multi-stain algorithms reduce the need for labels in target stains, they may struggle with tissue types lacking source-stain labels. To address this, the thesis focuses to improve multi-stain segmentation with less reliance on labelled data using self-supervision. While this thesis focused on the application of kidney glomeruli segmentation, the proposed methods are designed to be applicable to other histopathology tasks and domains, including medical imaging and computer vision.

Université de Strasbourg

Doctoral Thesis

Discipline: Computer Science

Self-Supervised Learning in the Presence of Limited Labelled Data for Digital Histopathology

Presented by

Zeeshan Nisar

defended on 24th September 2024

Jury Members

Thesis supervisor:	Thomas Lampert, Chair of Data Science and AI, *Université de Strasbourg*, France
Reviewers:	Thomas Walter, Professor, *Ecole des Mines de Paris*, France Kurtz Camille, Professor, *Université Paris Cité*, France
Examinators:	Maja Temerinac-Ott, Professor, *Fakultät Informatik, Furtwangen University*, Germany Cédric Wemmert, Professor, *Université de Strasbourg*, France Dino Ienco, *Directeur de Recherche*, UMR TETIS, France

ICUBE
Laboratoire des sciences de l'ingénieur de l'informatique et de l'imagerie

Acknowledgements

Finally, the incredible PhD journey of almost four years (January 2021 to September 2024) has come to an end, and I still find it hard to believe. This period has been an extraordinary chapter in my life, taking me from my home in Pakistan to France, a country I had never visited before. It was my first time leaving my city, my country, and most importantly, my home—a leap into the unknown that has changed me in ways I could never have imagined.

Living and studying in France has been a challenge on many fronts, pushing me to grow both personally and professionally. These experiences, though often difficult, have made me a stronger and better person. However, none of this would have been possible without the presence and support of many remarkable individuals, to whom I owe my deepest gratitude.

First and foremost, I would like to express my heartfelt thanks to the most deserving person for this PhD, my supervisor, Prof. Thomas Lampert (Tom). I will never be able to express my gratitude or repay you for all that you have done for me. Throughout this journey, you have been more than just a supervisor; you have been a friend, a colleague who has consistently supported and helped me not only with my academic and research matters but also with administrative and even personal problems. Your unwavering support went far beyond what I could have ever imagined from a supervisor.

Tom, despite being an expert in your field, you have shown exceptional kindness and proved to be an amazing supervisor. I often find myself wishing that every student could have a mentor like you. Thank you for providing me with all the facilities I requested to improve my career, for always listening to my research-related confusions and ideas with patience, and for responding positively to them. Your guidance has been instrumental in shaping my academic journey and future career prospects.

I would like to express my deepest gratitude to my beloved parents. Their unwavering prayers have been a constant source of strength throughout this long and challenging journey. My mother's dream of seeing me become the first PhD doctor in our family and among our relatives has always been an inspiration for me to keep pushing forward, even in the most difficult moments. The motivation and support that both of them have provided, especially during my toughest times, have been invaluable. Without their encouragement, this achievement would not have been possible.

I also want to extend my heartfelt appreciation to my wife, whose sacrifices and unconditional support have been instrumental in my success. Leaving her home country, her family, and everything familiar to her, she embraced a new life with me in France, helping me overcome my loneliness and giving me the strength to focus on my work. Despite the demands of my research, which often kept me in the lab and away from home, especially during the early months of our marriage, she remained understanding and patient. Her support was unwavering, even when she needed me the most during her pregnancy. Together, we were blessed with a beautiful baby

boy who has filled our lives with immense joy and happiness. I cannot thank her enough for her resilience, love, and unwavering support and belief in me.

I am also deeply thankful to my brother and his wife, who have been taking such good care of my parents and continue to do so in my absence. Their love and care have ensured that my parents never felt lonely while I have been away.

I would also like to extend my sincere appreciation to my lab colleagues, Jelica, Mihailo, Jules, Ben, Islem and Ali, with whom I had fantastic collaborations and productive discussions. Working in such a supportive and encouraging environment has been crucial for my growth and success during these years.

Lastly, I want to thank my friends, with whom I spent many weekends playing cricket. Those moments were a wonderful way to relax and recharge, and their friendship has been an important part of this journey.

To all those who have played a part in my journey, whether mentioned by name or not, your contributions have not gone unnoticed. Your encouragement and support have meant the world to me, and for that, I am eternally grateful.

Declaration

I hereby declare that this thesis, entitled "Self-Supervised Learning in the Presence of Limited Labelled Data for Digital Histopathology", submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD) in Computer Science, at the University of Strasbourg, is my original work. It has not been submitted for any other degree or examination in any other institution. The work presented in this thesis gave rise to the following publications:

[a] **Zeeshan Nisar**, Jelica Vasiljević, Pierre Gançarski, and Thomas Lampert. "Towards Measuring Domain Shift in Histopathological Stain Translation in an Unsupervised Manner" In IEEE International Symposium on Biomedical Imaging, pp. 1–5, 2022.

[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): 102420.

[e] Ali Alhaj Abdo, Islem Mhiri, **Zeeshan Nisar**, Barbara Seeliger and Thomas Lampert. "StairwayToStain: A Gradual Stain Translation Approach for Glomeruli Segmentation", MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL) (2024).

[c] **Zeeshan Nisar**, and Thomas Lampert. "Maximising Histopathology Segmentation using Minimal Labels via Self-Supervision", Under Review.

[d] **Zeeshan Nisar**, and Thomas Lampert. "Adversarial Noise Optimisation in Unsupervised Stain Translation", Under Preparation.

I furthermore declare that I have properly acknowledged and cited all sources, including publications, articles, books, and any other materials used or consulted in the preparation of this thesis. All contributions from other individuals or organisations have been duly acknowledged.

I hereby declare the accuracy, originality of the content, and innovative ideas presented in this thesis are entirely my own work. Certain AI-based tools such as, Grammarly and ChatGPT, were used to refine the language, grammar, and clarity of the writing under my careful supervision and verification. These tools were used exclusively for the above-mentioned reasons. Additionally, any assistance received in terms of technical, intellectual, or financial support has been duly acknowledged. I understand that any act of plagiarism or academic dishonesty is a serious offence and may result in severe consequences.

I hereby grant the University of Strasbourg the non-exclusive right to reproduce and distribute copies of this thesis, either in print or electronic format, for scholarly purposes.

Abstract

The integration of artificial intelligence, particularly deep learning, with medical imaging holds tremendous promise and potential. Automated computer-aided diagnostic systems, powered by deep learning, have emerged as one of the most important research area in the medical imaging domain. In such an environment, digital histopathology is not an exception. However, a principle challenge in applying deep learning to histopathology is inter- and intra-stain variations arising from different stainings and protocols. These variations lead to a significant drop in the performance of a deep learning model trained for one stain (aka domain in machine learning) when applied to other stains (even for the same task). Acquiring labels for each stain is a time-consuming and costly process.

To overcome these challenges, researchers have turned their focus towards CycleGAN, an unpaired image-to-image translation framework, based stain transfer methods. These are used to train multi-stain segmentation models using labelled data for a single (source) stain while eliminating the need of labels in the target stains. However, in accordance with the recent advances in CycleGAN, these methods face limitations, notably a drop in performance because of the introduction of additional noise, resulting in domain shift, during stain transfer. A crucial aspect of addressing this domain shift is the ability to detect it. Hence a key contribution of this thesis lies in exploring unsupervised approaches to propose a method that can serve as a metric for detecting and quantifying domain shift in stain transfer. Furthermore, this thesis delves into the exploration of recent advances in the CycleGAN based unpaired image-to-image translation framework to introduce an approach focused on minimising the domain shift in stain transfer.

While existing stain transfer based multi-stain segmentation algorithms have demonstrated their effectiveness in eliminating the need for labels in target stains, their applicability may prove impractical for certain tissue or tumour types that lack substantial amounts of labelled data in the source stain. Nevertheless, histopathology can offer a large amount of unlabelled data and so to overcome these shortcomings, this thesis proposes to use this unlabelled data to reduce the dependence on labelled samples. It therefore introduces a novel approach that integrates visual representation learning methods, particularly self-supervised learning, to enhance multi-stain segmentation algorithms by reducing their reliance on labelled data for the source stain.

In this thesis, we primarily focused on the use case of kidney glomeruli segmentation across multiple stains, but the overall objective is to propose general novelties that can be applied across multiple other related histopathology tasks and domains, including computer vision and medical imaging.

Contents

Chapter 1

Introduction

Artificial intelligence (AI) has rapidly evolved over the past six decades, drawing insights from various scientific domains such as mathematical logic, statistics, computational neurobiology, and computer science. The overall objective has been to replicate the intelligence of humans in machines [1]. The early developments of artificial intelligence were largely centred in the field of computing, facilitating computers to handle increasingly complex tasks. Despite this, automation falls short of true human intelligence, which makes the term "artificial intelligence" open to criticism.

The transformative breakthrough in artificial intelligence emerged around 2010, marked by the advent of revolutionary deep learning (DL) algorithms. This breakthrough can be attributed to several factors [1, 2]: (a) significant improvements in computational power effectively reduced the costs associated with training AI models; (b) increased accessibility to massive amounts of labelled datasets has empowered AI systems to demonstrate capabilities that were once considered unimaginable. In today's world, artificial intelligence has seamlessly integrated itself into the fabric of our daily lives. From the everyday convenience of personal assistants in our smartphones to the remarkable functionality of self-driving cars, AI has become an essential and integral part of our daily routines. Among the multitude of applications, the medical imaging domain emerges as a particularly sensitive and compelling area of concentration, attracting significant attention from researchers and clinical practitioners in recent times.

1.1 Motivation

The integration of artificial intelligence with medical imaging holds tremendous promise and potential. Automated computer-aided diagnostic (CAD) systems, powered by deep learning, have emerged as one of the most important research area in the medical imaging domain. These CAD systems empower clinicians by offering novel perspectives and capabilities in disease diagnosis, prognosis, and treatment planning. By thoroughly analysing and interpreting the patient's condition, these systems offer a level of precision and accuracy [3]. This revolution has opened the door for remarkable artificial intelligence applications in the medical domain [4]. Numerous everyday clinical tasks hold the potential to be fully automated, triggering a staggering amount of research in this field [5–7].

In such an environment, digital histopathology is not an exception. Deep learning algorithms have demonstrated exceptional performance in various histopathological tasks such as cancer detection, disease classification, and transplant assessment

[8]. In a controlled experimental setting, these solutions have attained levels of performance comparable to those of experienced pathologists [7], raising both hope and concern about the potential replacement of many expert jobs by machines. However, it is crucial to emphasise that these developments aim to promote the concept of "AI-assisted human experts" rather than viewing AI as a replacement. Furthermore, the complexity of medical tasks strongly suggests that replacing expert knowledge and experience with a machine is highly challenging [9, 10]. Moreover, the medical domain raises concerns about the responsibility for diagnoses, patient privacy, and the potential for bias or misinterpretations in AI generated results [11, 12]. Consequently, while AI has made significant progress in medical applications, the path towards fully automated medical expertise remains long and complex.

In recent years, the significant growth in publications within the field of artificial intelligence, particularly deep learning methods in digital histopathology, has been remarkable. This surge reflects an increasing interest and anticipation of the widespread adoption of these CAD systems. However, it is crucial to acknowledge the limitations inherent in these developments and to promote awareness about what is reasonable to expect from this technology and its progress. In this thesis, we highlight the common challenges associated with digital histopathology and deep learning based solutions. One of the biggest hurdles in developing effective deep learning solutions stems from the significant variance introduced by the staining process (a detailed description of the staining process and potential sources of variations inherent to this process will be presented in Section 1.2.1). Another notable challenge is the availability of large, high-quality labelled data, as will be outlined in Section 1.2.2.

This thesis seeks to investigate existing state-of-art deep learning methods specifically designed to overcome the challenges posed by staining variations. Additionally, it critically examines the shortcomings of these methods and suggests novel modifications in response to recent advances in deep learning to enhance their overall efficacy. Although the scarcity of high-quality labelled data in digital histopathology poses a significant challenge for integrating cutting-edge deep learning algorithms, digital histopathology boasts an abundance of unlabelled datasets. This abundance creates a significant opportunity for the exploration and development of deep representation learning methods, currently the state-of-art. These methods can be effectively adapted in scenarios where labelled data is limited. Therefore, another objective of this thesis is to investigate recent advances in deep visual representation learning methods and assess their efficacy in medical imaging domain, particularly in digital histopathology.

1.2 Digital Histopathology

Histology (originates from Greek, *histos* — tissue + *logos* — science) is the branch of biology which studies the microscopic structures of healthy animal or human tissues. Histopathology, a subdiscipline of histology, involves the microscopic study of changes that appear in the tissue as a consequence of disease (pathology) [13]. Histopathological examination stands as the gold standard method for diagnosing

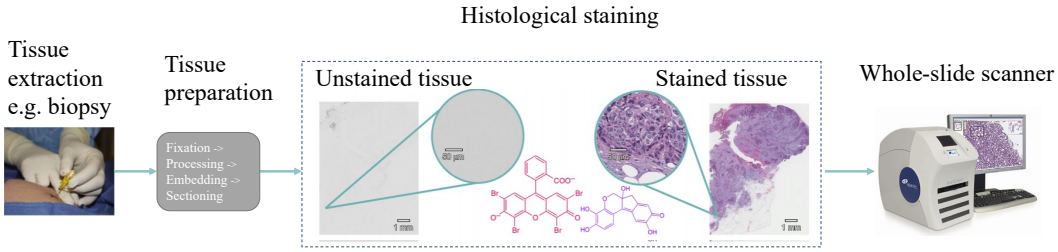


Figure 1.1: Illustration of the routine histological examination process[2]. Source: [17].

a wide range of diseases. The whole process starts with the extraction of a tissue sample from the body, usually through a biopsy or surgery. To facilitate microscopic analysis, the extracted tissue undergoes a series of preparatory steps: fixation, processing, embedding, sectioning and staining [14, 15], followed by scanning, which produces a digital version of the slide, thereby giving the name "digital histopathology[1]". The primary objective of these steps are to preserve the tissue's structure as much as possible, ensuring that the microscopic analysis closely mirrors the actual tissue in the body [16]. The procedural aspects of these preparatory steps are detailed as follows:

Tissue preparation: As depicted in Figure 1.1, the extracted tissue is promptly fixed using fixatives, such as formalin, to prevent decay. Following fixation, the tissue undergoes processing (such as dehydration, clearing and infiltration), and ends with embedding (usually with paraffin). At this step, the hardened block containing the tissue and surrounding embedding medium is placed in a microtome, an instrument designed for sectioning [16]. The microtome extracts very thin tissue sections $(3-10\mu m)$ which are then placed on a glass slide. Once tissue sections are obtained, they appear colourless providing very little detail of tissue structure.

Histological staining: In order to make the tissue components visible, the staining process is performed. Staining is a crucial step in histopathological examination as it visualises the chemical nature of the tissue and cell structures allowing for their detailed microscopic analysis. Further details about the staining process can be found in Section 1.2.1.

Scanning: The stained glass slides, being physical objects, coupled with the risk of scratches, cracks and colour fading due to prolonged exposure to light and environmental factors, necessitates careful preservation and storage to ensure their longevity. To address these challenges, modern histopathological procedures increasingly employ whole slide scanners, which digitally capture the entire glass slide. This process, known as whole slide imaging (WSI), creates a digital representation of the glass slide at the same level-of-detail as viewed through a light microscope

[1] In the rest of the thesis, the terms digital histopathology and histopathology are used equivalently.

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

[18]. For instance, a WSI, where each pixel corresponds to a square of $0.5\mu m$ ($0.25\mu m$) on the slide, is considered to offer an equivalent level-of-detail as observed through a $\times 20$ ($\times 40$) objective on a high-quality microscope [18]. This digitisation has revolutionised the pathology practice by allowing remote examinations and collaborative analysis. Clinicians can now analyse slides using a personal computer and specialised image manipulation software, regardless of their physical location. Additionally, multiple pathologists, even those dispersed across different hospitals, cities, or even countries, can collaborate simultaneously on the analysis of a single WSI.

1.2.1 Staining

Staining is a chemical process crucial for introducing the contrast in tissue sections, offering experts a detailed microscopic analysis of specific tissue components and cells. The basic principle underlying the staining of tissue involves the formation of chemical compounds (e.g. acidic or basic) between the dye and the tissue. This targeted binding allows different stains to highlight different tissue components, allowing pathologists to identify and detect specific patterns, such as the identification and detection of cancer cells and their distribution. The most commonly used stainings in histopathological examination are histochemical (HC) stainings, these are used to highlight general tissue structures. These use chemicals that interact with various tissue components, making them visible from different perspectives. Likewise, many other stain types have been developed to highlight specific tissue components, such as, immunohistochemical (IHC) stainings, which are used to gather more specific information, such as the expression of a particular protein (antigen). The basic working principle of these stainings involves antigen-antibody binding. To achieve this, a solution containing a special antibody is applied to the tissue, so that the antibody binds to the cells with the targeted antigen [13]. In the following, a brief description of the stainings used in this thesis is given:

Hematoxylin and Eosin (H&E): is a classical histochemical staining with a rich history of clinical usage [19]. It serves to highlight various tissue components, facilitating the analysis of a wide range of organs and diseases. It contains two key components: (a) hematoxylin, a basic dye which binds with acidic components such as cell nuclei and eosin; (b) an acidic dye which binds with basic components such as cell stroma or cytoplasm. This pairing results in a striking contrast between the nuclei, which appear blue, and the cytoplasm, which turns pink in the image. Thus, clear nuclear contrast can be achieved to reveal the distribution of cells [20]. This staining is widely recognised as the gold standard for diagnosing various types of cancer, and is routinely conducted in several histopathological examinations.

Jones Hematoxylin and Eosin (Jones H&E): is a histochemical silver staining that is extensively used in renal (kidney) pathology to visualise the basement membranes in black, nuclei in blue, and the background in pink. This staining method is particularly useful for identifying abnormalities in the glomerular basement membrane.

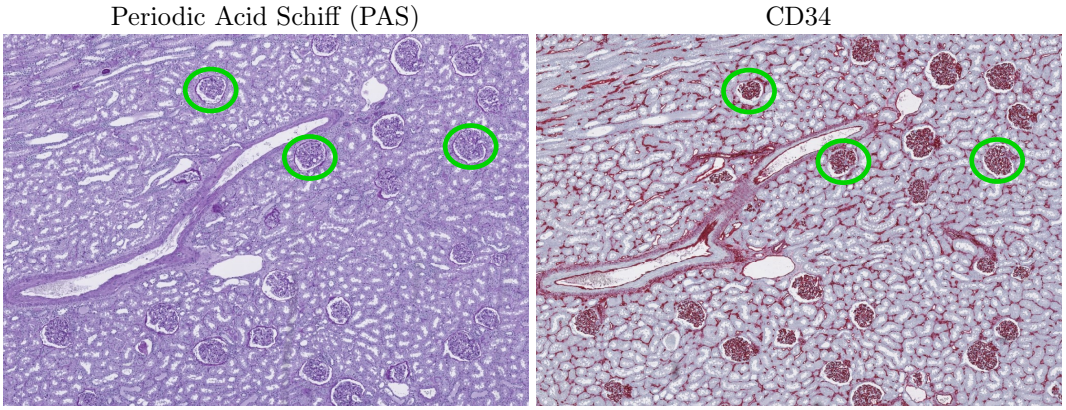


Figure 1.2: An example of two consecutive WSI samples of a kidney nephrectomy stained with different stains. Each stain highlights different information about the tissue structure but some common structures, such as glomeruli, are visible in all stainings (some of them are marked in green circles). Source: [23].

Periodic Acid-Schiff Reaction (PAS): is a histochemical staining used for detecting carbohydrate-rich structures in the tissue. This method involves exposing a tissue section to periodic acid oxidation and subsequently staining it with Schiff's reagent, as outlined in [21]. It visualises carbohydrate-containing cell components in magenta (shades of purplish pink) [16]. PAS is most commonly used to highlight cells filled with glycogen deposits, or the glycocalyx [16]. In kidney pathology, according to the Banff classification scheme [22], PAS staining is particularly valuable for identifying glomerulitis, tubulitis, and tubular basement membrane destruction.

Sirius Red: is a histochemical stain that highlights connective tissue (specifically collagen) in red, while cytoplasm in lighter violet or pink [16].

CD68: is an immunohistochemical stain which highlights 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).

CD3: is an immunohistochemical stain (similar to CD68 in appearance), which serves as a marker for T cells.

In histopathology, an important aspect is the analysis of multiple WSIs from the same tissue stained with different stainings. Consecutive slices from the same biopsy or nephrectomy are stained with different stains to enable the analysis of underlying tissue from different perspectives. While different stains highlight different tissue components, some general analyses can be performed across multiple stainings. For example, in kidney pathology, glomeruli[3] are observable under various stains as il-

[3] A glomerulus is a tiny ball-like structure embedded in the nephron and serves as the kidney's primary filtration unit. It consists of a complex network of specialised capillaries designed for the efficient removal of waste products and excess fluid from the blood, ensuring the maintenance of healthy bodily functions [24].

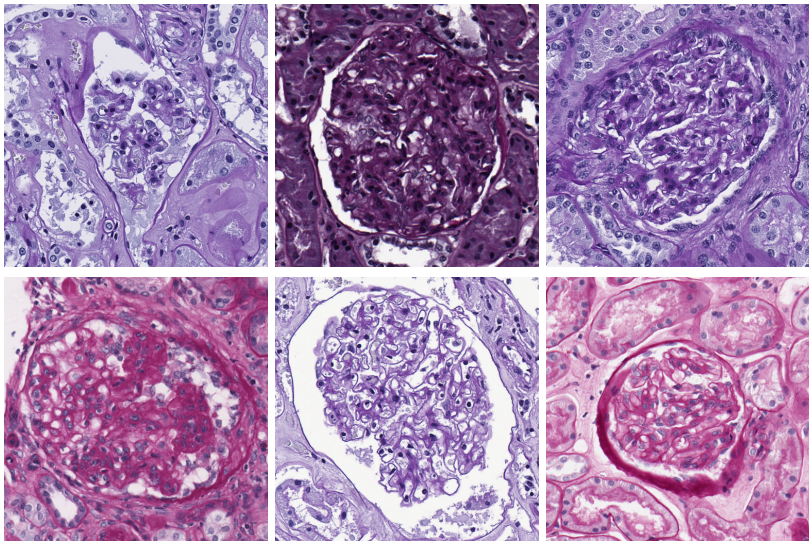


Figure 1.3: Illustration of intra-stain variation in PAS staining in kidney pathology. Each image represents the glomerulus in PAS staining. Source: [23].

lustrated in Figure 1.2. Since each staining highlights different tissue structures, the final WSI can exhibit significant differences in appearance. Additionally, the staining procedure itself is vulnerable to high variability due to inter-subject variations, lab specific techniques, and scanner characteristics. This can can introduce additional differences in the visual appearance of the tissue, even when subjected to the similar staining method. These variations can be attributed to inter-stain and intra-stain variations, which are further explored in the following discussion.

1.2.1.1 Intra-Stain Variation

The preparation of high-quality tissue slides requires careful handling and processing of the tissue at each of the above-mentioned tissue preparatory steps in Section 1.2. Variations in any of these steps can introduce artefacts [25, 26] that lead to intra-stain variation. Besides these artefacts, the most prevalent sources of intra-stain variation include disparities in raw materials, capturing pipeline changes, the quality of the substances used or characteristics of the scanner. Figure 1.3 illustrates the intra-stain variation in the PAS staining in kidney pathology and its impact on the appearance of glomeruli. While such variations can be taken into account during manual analysis by experienced pathologists with specialised training, they pose a considerable challenge for deep learning based automated solutions.

1.2.1.2 Inter-Stain Variation

Despite representing the same anatomical structures, consecutive tissue sections can exhibit significant visual discrepancies when subjected to different stainings, resulting in inter-stain variations as illustrated in Figure 1.4. To effectively analyse and integrate information from different stainings, it is of great importance to focus on

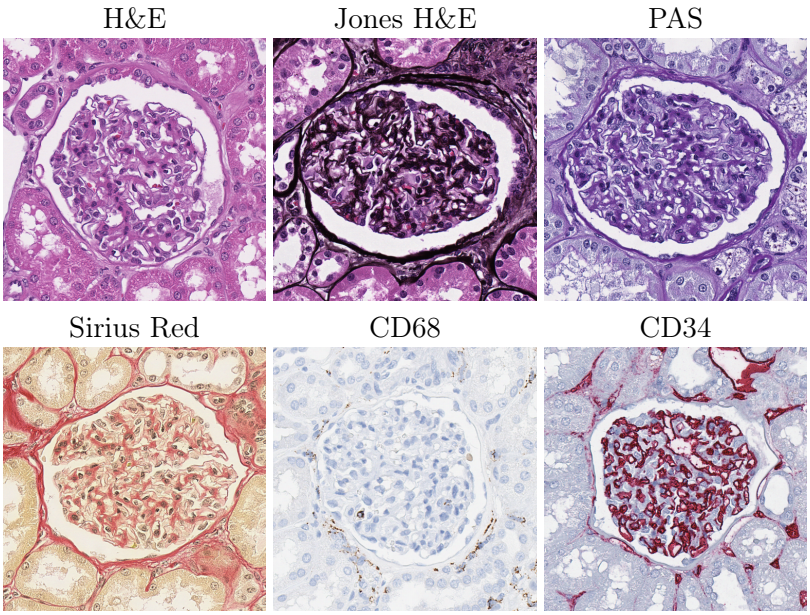


Figure 1.4: Different stains used in kidney pathology. Each image represents a glomerulus and each stain provides specific information about the structure of glomerulus.

Table 1.1: Average segmentation (F_1) scores for 5 different U-Net repetitions (trained on PAS) and applied to full test slides of different stains. Standard deviations are presented in parentheses.

Training Strategy	Test Stainings				
	PAS	Jones H&E	CD68	Sirius Red	CD34
Baseline PAS	0.894 (0.021)	0.062 (0.011)	0.044 (0.098)	0.045 (0.037)	0.056 (0.090)

specific structures of interest, such as glomeruli in the context of kidney analysis. These structures play a critical role in diagnosing pathologies like kidney allograft rejection [27]. To automate this analysis, glomeruli must be detected and/or segmented in each of the consecutive tissue sections, regardless of the employed staining method. However, these inter- and intra-stain variations introduce specific distributional differences, which lead to the problem of *domain shift* across and between stains. Since most deep learning algorithms are sensitive to domain shift it causes a significant drop in the performance of state-of-art algorithms [28], as outlined in Table 1.1, where a deep learning model (i.e. UNet [29]) trained on PAS for the segmentation of glomeruli structures experiences a notable decline in performance when applied to stains other than PAS.

1.2.1.3 Overcoming Stain Variations

While several studies [30–36] have been conducted to address the problem of intra-stain variations, a few address the problem of inter-stain variations [27, 28, 37, 38]. A straightforward, albeit costly, solution to address these challenges involves acquiring a significant number of labels for each stain, followed by the subsequent training of distinct deep learning models tailored for each satin. However, this approach proves highly impractical and inefficient, as it undermines the inherent efficiency and potential generalisation capability of deep learning methods [23]. While each separate model may adeptly learn the specific features of each stain, it would struggle to generalise to other stains even for the same task or same stain collected from other medical centres, limiting the broader applicability of deep learning methods. Moreover, creating and training individual models for each stain can be exceedingly complex, since each model has to be trained on labelled data for each stain. Acquiring these labels is a costly and time consuming process, primarily due to the requirement of highly specialised medical experts to label the data. This further exacerbates the resource constraints associated with this solution.

To address these challenges, it is often preferred to acquire a sufficient amount of labels for a single (source) stain[4] and to train a multi-stain segmentation model that can potentially work across various unlabelled (target) stains, thus mitigating the issues related to data scarcity and facilitating the learning of more generalised features. To achieve this, the most effective method is stain transfer, which will be detailed further in Chapter 2.

1.2.2 Data Availability

The effectiveness of deep learning methods in various tasks is heavily dependent on the availability of large-scale labelled datasets. Numerous studies have consistently demonstrated that increasing the size of training datasets consistently improves the performance of deep learning methods, often surpassing human expertise across various scientific domains [39–41]. In light of these remarkable achievements, substantial advances in automating routine histopathological analysis can be anticipated.

The advent of WSI scanners has facilitated the production of vast amounts of histopathological image data. In recent years, research papers have reported a substantial increase in dataset sizes, often by several orders of magnitude [42]. However, not all the data produced is of sufficient quality to be directly used and requires additional data preparation efforts. Nevertheless, the most tedious task is related to the annotation process, particularly when dealing with structures of interest that can be either sparsely or densely distributed. For instance, structures such as glomeruli are sparse, occupying approximately 2% of the kidney tissue area [43]. Therefore an average kidney WSI, having a size of 100K $\times$ 80K pixels (at 40$\times$ magnification), contains around 500 glomeruli. In contrast, cancerogenic cells can appear in a large portion of WSIs, occupying dense areas within the image. This inherent variability in structure distribution poses a significant challenge for annotation, as it is time-consuming, expensive and task-specific. Additionally, annotation requires

[4] In this thesis, the terms “domain” and “stain” are used equivalently.

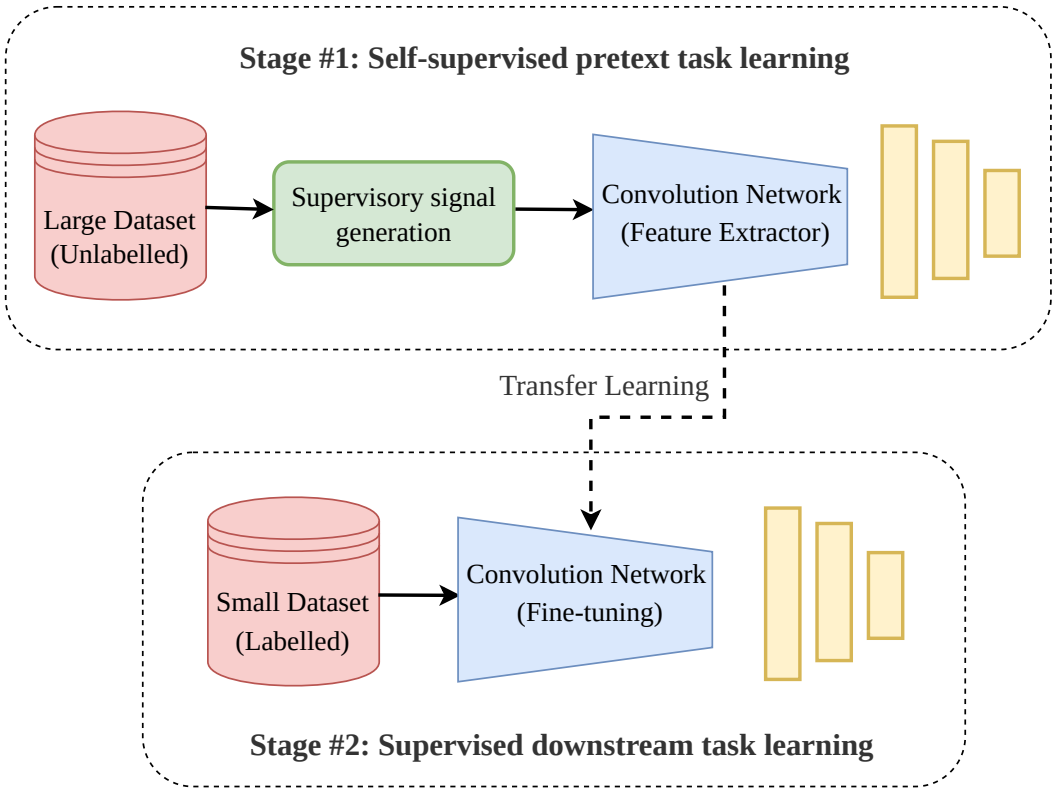


Figure 1.5: Self-supervised learning workflow. (Stage #1): Self-supervised learning involves training an auxiliary pretext task by generating its own supervisory signals for a large pool of unlabelled data. (Stage #2): The learned representations from the pretext task are transferred as initial weights to the specified downstream using a very small amount of labelled data.

specialised expertise [44], making it impractical to obtain high-quality annotations for all the data produced daily in hospitals. In addition to other important concerns related to data privacy, these enormous collected datasets are often left aside and not used in training supervised models.

1.3 Self-Supervised Learning

In situations where unlabelled data is readily available in large quantities and labelled data is limited, a promising approach is to is to leverage this unlabelled data to enhance a model's performance through representation, particularly self-supervised, learning.

Self-supervised learning can be used to learn semantic features by generating its own supervisory signals from a pool of unlabelled data, thereby eliminating the reliance on expert annotations [45]. The learned features from this self-supervision stage can then be effectively employed in various downstream tasks where labelled

data is scarce, as illustrated in Figure 1.5. In essence, self-supervised learning embodies the fusion of both unsupervised and supervised learning approaches. The unsupervised aspect eliminates the need for manual labelling, while the supervised aspect is evident in training the model using labels generated from the data itself [46]. The self-supervised learning pipeline comprises two key stages:

- **Pretext Task:** The initial stage, where self-supervised learning actually occurs, involves a pre-designed task that enables the model to learn meaningful representations without relying on explicit labels, guided by its pre-designed objective function.
- **Downstream Task:** Once the model has learned representations from the pretext task, they are transferred to specific computer vision applications, referred to as downstream tasks. These learned representations serve as initial weights for fine-tuning the specific downstream task with minimal human annotations.

The design of a pretext task is the fundamental aspect of self-supervised learning. Although downstream tasks may vary depending on the application, the pretext task may remain the same. For example, a convolutional auto-encoder can be used to learn visual representations for two different downstream tasks with different data [46]. In recent years, a number of effective self-supervised learning approaches have been proposed in various histopathology tasks such as identification and classification. However, only limited attempts have been made to incorporate the advances of self-supervised learning for histopathology segmentation. This thesis endeavours to investigate the current research directions in self-supervised learning methods and assess their potential impact on the segmentation of glomeruli structures across various stains, while using minimal labelled data.

1.4 Thesis Goals and Contributions

This thesis centres on exploring robust and generalised deep learning algorithms that exhibit stain-invariance and can function seamlessly across multiple stainings for the same task, such as kidney glomeruli segmentation in our case. The primary goal is to achieve this while minimising the need of expert human annotations. This research builds upon prior works [23, 28, 37], which have introduced various multi-stain segmentation approaches to address these challenges, while using labels solely from one (source) stain. These prior works have explored the potential of unsupervised generative adversarial networks (GANs), particularly the CycleGAN framework [47], renowned for its unpaired image-to-image translation approach. This framework is employed to facilitate stain transfer for creating multi-stain segmentation models.

However, in light of recent advances in the CycleGAN [48–50], it will be seen in this thesis that its effectiveness is hampered by imperceptible noise [49] (which will be explored further in Chapter 3) during adversarial image-to-image translation. Therefore, it is essential to exercise appropriate caution when deploying these methods for clinical aid. To mitigate these limitations, this thesis presents several contributions:

- A crucial aspect in addressing the domain shift in stain transfer involves the ability to detect it. Hence, one of the key contributions of this thesis lies in exploring unsupervised approaches to propose a metric to quantify it. Additionally, these findings reveal a strong correlation between domain shift and the segmentation performance of translated stains. These findings therefore pave the way for establishing a mechanism to infer the average performance of a pre-trained model (trained on a source domain) when applied to an unseen and unlabelled target domain.

- Using this measure, we demonstrate the sensitivity of CycleGAN towards subtle architectural modifications. Although, these modifications may not necessarily affect the visual quality of the resulting translations, they significantly effect the overall performance of stain transfer based multi-stain segmentation approaches. This holds true from both a diagnostic and application perspective – highlighting the thesis' second contribution.

- We then propose a novel approach that minimises the addition of noise (domain shift) during stain transfer, thereby enhancing the performance of multi-stain segmentation – highlighting the thesis' third contribution.

- The fourth and last contribution of this thesis involves integrating state-of-art deep representation learning methods, particularly self-supervised learning, to conduct a comprehensive analysis for reducing the number of labels required for histopathological segmentation. Additionally, this contribution strives to improve multi-stain segmentation approaches by reducing their reliance on labelled data for the source stain—which to the best of our knowledge has not been explored previously—paving the way for more cost-effective and scalable solutions for domain adaptation based algorithms. This contribution also puts forth several modifications to enhance the adaptability of self-supervised learning methods across various staining protocols, especially those that are stain-specific and therefore limited to a single type of stain.

The research contributions outlined in this thesis have undergone rigorous training and evaluation by using a private histopathology dataset for kidney glomeruli segmentation across multiple stains. However, the primary objective of these research contributions is to introduce general novelties that hold applicability across other related histopathology tasks and domains, including computer vision and medical imaging, particularly those that face similar challenges. The subsequent section will detail the dataset used within this thesis.

1.4.1 Data

This thesis is focused on renal pathology with a particular emphasis on the segmentation of glomeruli in multiple stainings. The data used in this thesis is private, encompassing tissue samples extracted from a cohort of 10 patients who underwent tumor nephrectomy due to renal carcinoma. The renal tissue was selected as distant as possible from the tumors to represent largely normal renal glomeruli. However, certain samples exhibited varying degrees of pathological modifications, such as

Table 1.2: The Number of glomeruli present in each staining.

Staining	Training	Validation	Test
PAS	662	588	1092
Jones H&E	621	593	1043
Sirius Red	651	579	1049
CD34	565	598	1019
CD68	526	524	1046

complete or partial displacement of functional tissue by fibrotic changes ("scerosis") indicating normal age-related changes or the renal effects of general cardiovascular comorbidity (e.g. cardial arrhythmia, hypertension, arteriosclerosis). Using an automated staining tool (Ventana Benchmark Ultra), the paraffin-embedded samples were sliced into 3µm thick sections and stained with either Jones H&E basement membrane stain, Periodic acid-Schiff reaction (PAS), Sirius Red, in addition to two immunohistochemistry markers, such as CD34 and CD68. An Aperio AT2 scanner was used to capture whole slide images at 40× magnification (a resolution of 0.253 m/pixel). Pathology specialists annotated and verified all of the glomeruli in each whole slide image by labelling them with Cytomine [51]. The whole dataset (WSIs) was split into 4 training, 2 validation, and 4 test patients. The number of glomeruli in each staining dataset is given in Table 1.2.

Kidney glomeruli segmentation is framed as a two class problem: glomeruli (pixels that belong to glomerulus), and tissue (pixels outside a glomerulus). The training set comprised all glomeruli from a given staining's training patients and seven times as many tissue (i.e. non-glomeruli) patches were included to account for the variability observed in non-glomeruli tissue. In order to remove the slide background (non-tissue), each image underwent thresholding based on its mean value, followed by the removal of small objects and closing holes. Throughout the study, the level-of-detail used is 1 (corresponding to 20× magnification) with an image patch size of 508 × 508 pixels, since glomeruli and part of the surrounding area fit within this patch size at the level-of-detail used.

1.4.2 Thesis Outline

The remainder of this thesis is structured as follows:

- Chapter 2 provides a thorough and systematic analysis of the existing literature on stain transfer based multi-stain segmentation approaches in histopathology. Furthermore, the chapter exposes the existing research on self-supervised learning, both within computer vision and specifically in the context of medical imaging, with a focus on histopathology.

- Chapter 3 thoroughly investigates the inherent shortcomings of stain transfer and provides a comprehensive understanding of the specific scenarios in which these stain transfer based approaches may fail. The results from these investigations are published in [52] and [53]. Furthermore, this chapter explores the

recent advances in deep learning to present a range of potential solutions that can effectively address these shortcomings. The results of these findings are currently being documented to submit in a reputable conference or journal.

- Chapter 4 investigates the current state-of-the-art research directions in self-supervised learning methods and assess their potential impact on digital histopathology segmentation (through the use case of glomeruli segmentation) using limited labels. The results of this chapter are currently under review [54].

- Chapter 5 concludes the research presented in this thesis by providing a comprehensive summary of the research findings and identifying promising directions for future investigation.

Chapter 2

Literature Review

Digital histopathology has emerged as a rich area of innovation in both clinical applications and research, where deep learning algorithms have demonstrated remarkable achievements [8]. However, a major challenge lies in the training of these algorithms, since many state-of-the-art deep learning algorithms are data hungry and often demand extensive amounts of labelled data. This can be difficult to collect because medical datasets often require the creation of high quality annotations by field experts [44], which is a very costly and time consuming process. These constraints pose obstacles to the development of deep learning based automated solutions. Additionally, existing datasets (with labels) often have limited reusability due to variations in tissue preparation and staining protocols (as detailed in Chapter 1). Since each stain highlights different tissue structures, even consecutive tissue slides (representing identical anatomical structures) can appear very different (as was illustrated by the glomeruli in green circles in Figure 1.2 of Chapter 1). These discrepancies result in distribution differences in the feature space of each stain, as illustrated in Figure 2.1. The figure clearly shows how each stain occupies a distinct part of the feature space. These distributional differences lead to the problem of domain shift, which significantly impacts the performance of state-of-art deep learning methods, as will be explored in Chapter 3. Therefore, it becomes necessary to explore strategies to address such variations to effectively develop and apply deep learning based automated solutions in digital histopathology.

To address these challenges, stain transfer — where the appearance of an image is artificially modified after its acquisition — has emerged as the state-of-the-art solution. This process aims to transform an image stained with stain A to look like it had been stained with stain B, and vice-versa. An example of the result of stain transfer is illustrated in Figure 2.2, where a stain-B-like image is artificially generated from stain A and a stain-A-like image is artificially generated from stain B. From a computer vision perspective, these artificially generated images can be used to minimise the distribution disparities between stain A and stain B. The underlying hypothesis posits that if stain transfer produces visually convincing translations, it should be able to reduce the distribution differences (domain shift) between different stains, thereby enhancing a model's robustness to different stain variations. For instance, deep learning models trained on real images from stain A should be capable of extracting similar features and performing effectively on stain-A-like images (translated from stain B), and vice-versa. This paves the way for the development of various multi-stain segmentation methods, using labels from only one (source) stain. These methods can be primarily categorised into two different training strategies:

Stain-Specific: Training a segmentation model for a particular stain, referred to

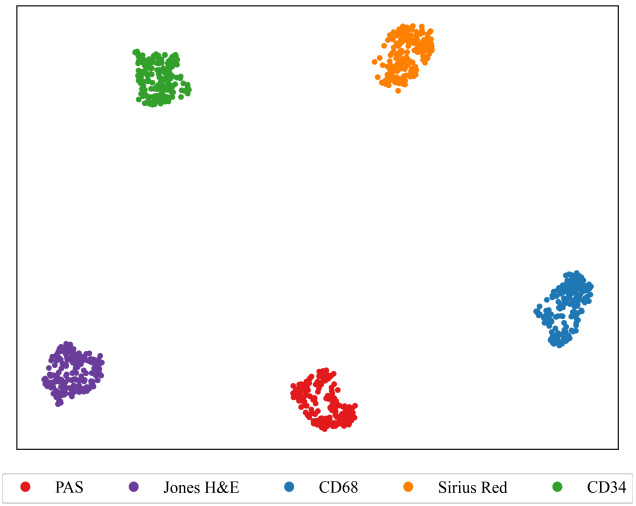


Figure 2.1: PaCMAP [55] visualisations of pre-trained ImageNet-based-ResNet features extracted from 200 randomly selected patches from each stain (1000 patches in total). Each coloured point corresponds to a patch from the respective stain.

as the source stain, for which the labels are available. This model is later applied to various other target stains by translating them to the source stain during test time. Further details are provided in Section 2.2.1.

Stain-Invariant: Training a single stain-invariant segmentation model on all available stains, using labels from only one (source) stain. This stain-invariant model can be directly applied across various other stains, including out-of-distributions stains, without the need for translation during testing. Further details are provided in Section 2.2.2.

Despite the proven effectiveness of existing multi-stain segmentation methods in eliminating the need of labels in the target stains, it is crucial to acknowledge that these methods rely heavily on a large number of labelled data from the source stain, which can still be challenging in the medical domain. For instance, in histopathology, for certain tissue or tumour types, sufficient labelled datasets for the source stain may not exist, preventing the successful training of the aforementioned approaches. However, the advent of whole slide imaging (WSI) scanners has facilitated the production of vast amounts of unlabelled histopathological image data. As such, unlabelled medical imaging datasets are increasing in size by several orders of magnitude [42]. When unlabelled data is accessible in large quantities, it can be used in limited annotation scenarios to enhance model performance through representation, particularly self-supervised, learning. Following this, the remainder of this chapter is organised as follows:

- Section 2.1 presents the concept of stain transfer and explores the current state-of-art methods from the literature for stain transfer.
- Section 2.2 explores the current state-of-art multi-stain segmentation approaches

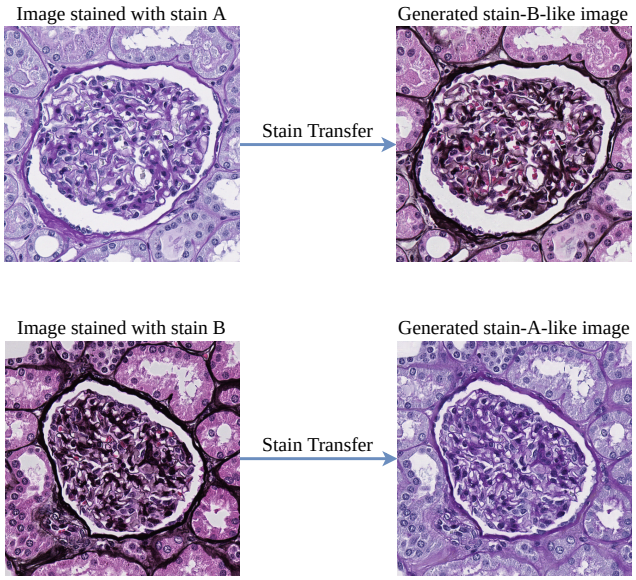


Figure 2.2: The basic principle behind stain transfer where the appearance of an image is artificially modified. In 1st row, an image with the characteristics of stain A is artificially transformed to look like an image with the characteristics of stain B. In 2nd row, an image with the characteristics of stain B is artificially transformed to look like an image with the characteristics of stain A.

based on stain transfer. Furthermore, this section highlights the effectiveness of these methods towards addressing the challenges posed by inter-stain variations.

- Section 2.3 explores cutting-edge research directions in representation learning methods found in the literature.

- Section 2.4 identifies the existing limitations in stain transfer based solutions. Additionally, it serves to motivate the exploration of representation learning methods in histopathology related tasks, aiming to enhance the effectiveness of stain transfer based multi-stain segmentation approaches when confronted with limited labels in the source stain.

- Finally, Section 2.5 summarises the key findings and investigates the prospective research opportunities drawn in this chapter.

2.1 Stain Transfer

Stain transfer can be formulated as an image-to-image translation problem, wherein images originally stained with stain A are translated to appear as if they had been

stained with stain B in a realistic and plausible[1] manner. The primary objective is to transform the visual characteristics of stain A to closely resemble those of stain B, while preserving the underlying image content, as shown in Figure 2.2. Generative adversarial networks (GANs) [56, 57] have emerged as the leading approach in image-to-image translation. The generator network learns to transform an image of stain A to stain B in such a way that the discriminator cannot differentiate between real stain B images and translated stain-B-like images, see Figure 2.2 (1^{st} row), and vice-versa. As such, stain transfer aims to minimise stain distribution differences in image (pixel) space.

While, in recent years, a number of state-of-art unpaired image-to-image translation frameworks [36, 47, 58–67] have been employed for stain transfer in various digital histopathological applications, CycleGAN [47] has emerged as a prevalent choice. This widespread adoption can be attributed to its straightforward applicability [68] and superior performance when compared to others [28, 69]. Furthermore, it has shown remarkable effectiveness in several multi-stain segmentation methods (which are explored later in this chapter).

2.1.1 CycleGAN

CycleGAN [47] is a bi-directional unpaired image-to-image translation framework that has been widely used for stain transfer in digital histopathology [28, 36, 37, 70–73]. The network architecture of CycleGAN is presented in Figure 2.3, where it contains two generators which perform translations between two stains: $G_{AB} : A \rightarrow B$ to translate from stain A to stain B (the output of which is termed stain-B-like) and $G_{BA} : B \rightarrow A$ to translate from stain B to stain A (the output of which is termed stain-A-like); in addition to two discriminators D_A and D_B. The aim of D_A is to differentiate between real stain A images and translated stain-A-like images; while D_B aims to differentiate between real stain B images and translated stain-B-like images. Given an image of a source stain $s \sim S$ and a target stain $t \sim T$, these networks are trained in an adversarial manner using a least-squared loss function ($\mathcal{L}_{\text{adv}}$), such that

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G_{AB}, D_B, G_{BA}, D_A) = \mathbb{E}_{s\sim A}[(D_A(s) - 1)^2] + \mathbb{E}_{t\sim B}[D_A(G_{BA}(t))^2] \\ + \mathbb{E}_{t\sim B}[(D_B(t) - 1)^2] + \mathbb{E}_{s\sim A}[D_B(G_{AB}(s))^2].\end{aligned} \tag{2.1}$$

Additionally, the training is constrained by the cycle-consistency ($\mathcal{L}_{\text{cyc}}$) and identity loss ($\mathcal{L}_{\text{id}}$) functions, such that

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) = \mathbb{E}_{s\sim A}[\|G_{BA}(G_{AB}(s)) - s\|_1] \\ + \mathbb{E}_{t\sim B}[\|G_{AB}(G_{BA}(t)) - t\|_1],\end{aligned} \tag{2.2}$$

[1] In this thesis, the term "plausible" refers to the fact that a histological image, when processed with other staining modalities without the knowledge of adjacent tissue sections and/or patient-specific information (e.g. the underlying disease), looks visually correct to a trained expert with regard to staining characteristics and morphological appearance of the tissue components [23].

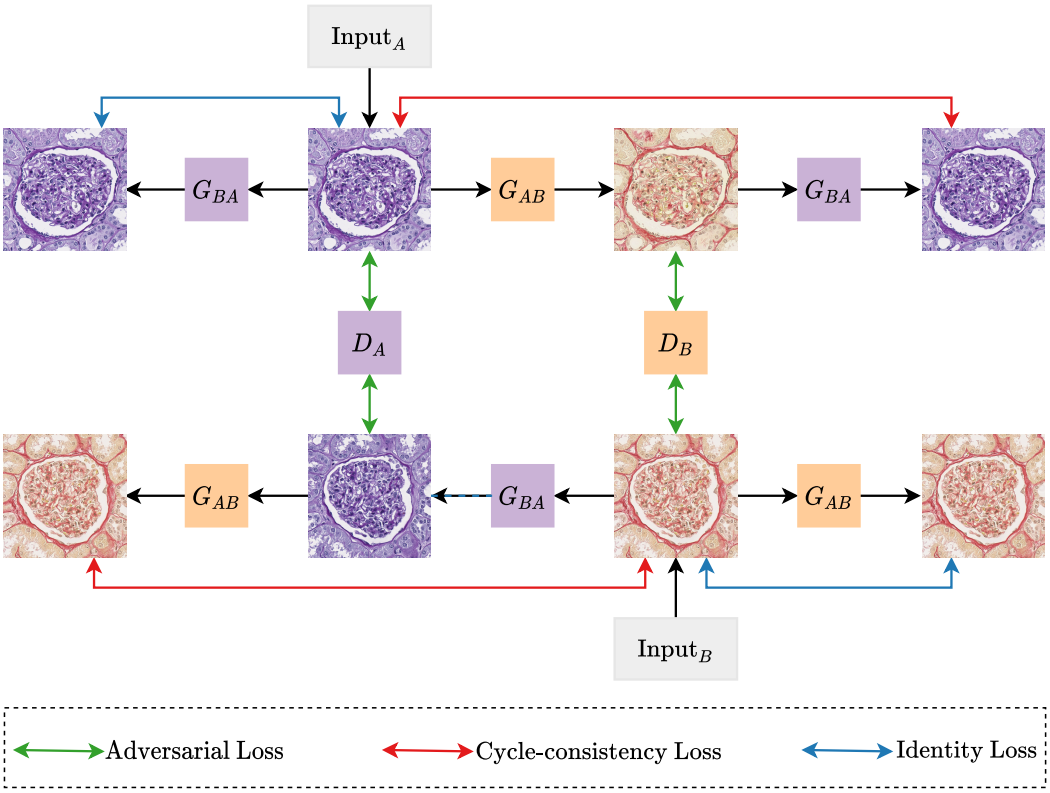


Figure 2.3: Overview of CycleGAN architecture for stain transfer

and

$$\begin{aligned}\mathcal{L}_{\text{id}}(G_{AB}, G_{BA}) = \mathbb{E}_{s\sim A}[\|G_{BA}(s) - s\|_1] \\ + \mathbb{E}_{t\sim B}[\|G_{AB}(t) - t\|_1].\end{aligned} \tag{2.3}$$

Thus, the full objective function becomes

$$\begin{aligned}\mathcal{L}_{\text{CycleGAN}}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{\text{adv}}(G_{AB}, D_B, G_{BA}, D_A) \\ + w_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) \\ + w_{\text{id}}\mathcal{L}_{\text{id}}(G_{AB}, G_{BA}),\end{aligned} \tag{2.4}$$

where w_{cyc} and w_{id} control the relative importance of the cycle-consistency and identity losses, respectively.

Once trained, the CycleGAN model performs translation between source (S) and target (T) stains using the corresponding generators. The outcomes of CycleGAN translations, such as $S \to T$ and $T \to S$ are shown in Figure 2.4, where $S \in \{\text{PAS}\}$ and $T \in \{\text{Jones H\&E, Sirius Red, CD68, CD34}\}$. PAS is used as the source stain throughout this thesis as it is widely used in clinical settings, which results in a substantial amount of readily available data and knowledge associated with this stain compared to others. As visualised in Figure 2.4, all translations appear plausible to

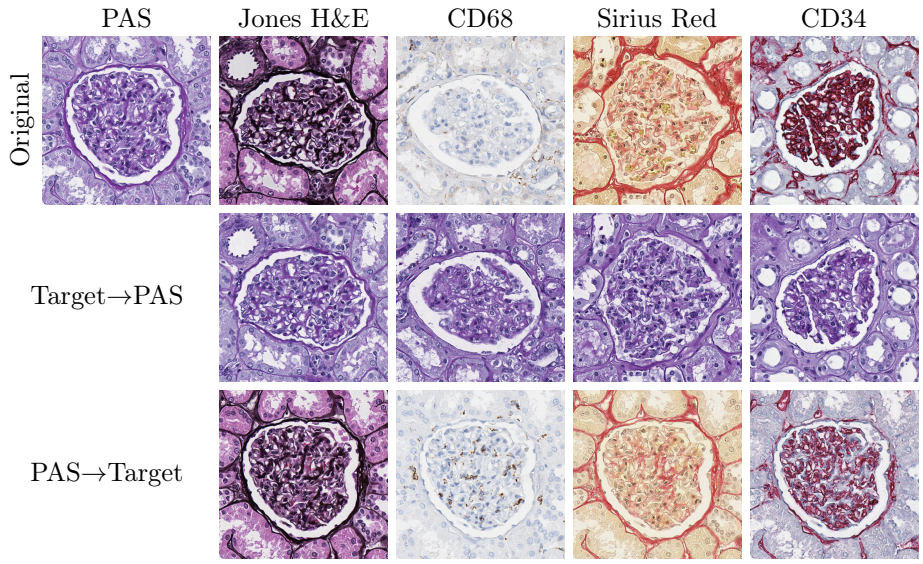


Figure 2.4: Illustration of unpaired stain-to-stain translation using CycleGAN.

a trained pathologist[2].

Additionally, Figure 2.5 demonstrates the ability of CycleGAN based stain transfer to minimise stain variations (distribution discrepancies) across multiple stains. As seen before, the dark coloured points show that each stain exhibits a unique distribution in pixel space, and the translated stains, represented as light coloured points, overlap them. For instance, all Target→PAS translated stains match the distribution space of the real PAS stain, as shown in Figure 2.5(a). Similarly, while translating PAS to each target stain, the PAS→Target translations attempts to match the distribution space of the real target stains, as shown in Figure 2.5(b).

2.2 Multi-Stain Segmentation Approaches

Multi-stain segmentation involves segmenting regions of interest in differently stained histopathological images using labels from only one (source) stain. While numerous multi-stain segmentation approaches have been proposed in the literature, most of them, particularly the state-of-the-art, heavily rely on stain transfer. Commonly an unpaired image-to-image translation framework, such as CycleGAN, is used. For instance, Gadermayr et al. [37] introduced a multi-stain segmentation method which enables a stain-specific segmentation model (trained for the source stain) to be used with multiple other target stains by translating them to the source stain using CycleGAN during test time. Similarly, Lo et al. [74] used CycleGAN translations to achieve multi-stain glomeruli detection, while Wu et al. [75] advocate fine-tuning a CycleGAN generator using a classification network for multi-stain glomeruli classification. Furthermore, Kapil et al. [76] proposed to expand the CycleGAN model

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

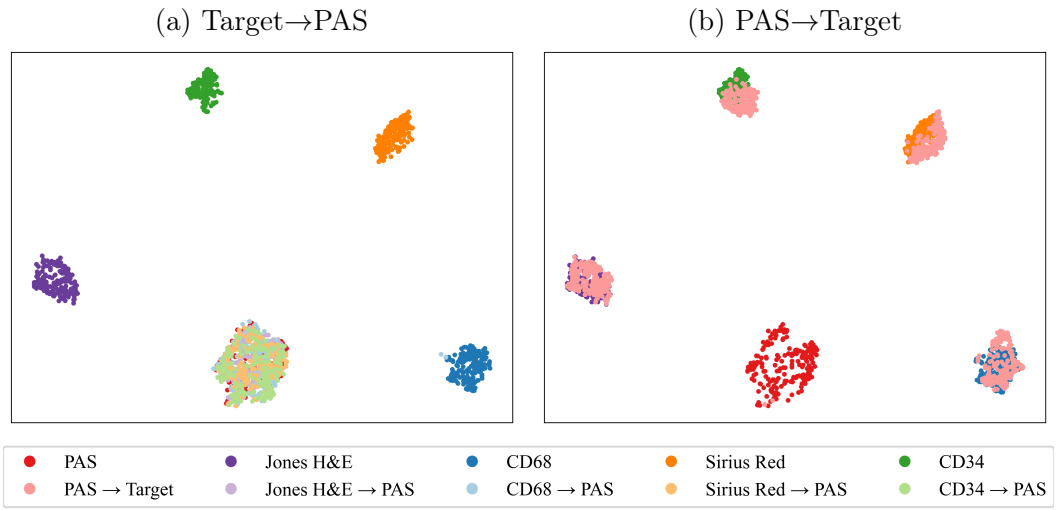


Figure 2.5: PaCMAP [55] visualisations of pre-trained ImageNet-based-ResNet features for real and translated patches. 200 real patches are randomly selected for each stain (1000 patches in total). The translated patches are obtained by translating from each target stain to source (PAS) stain, represented as Target→PAS, and from source stain to each target stain, represented as PAS→Target. The dark coloured points corresponds to the real patches whereas the light coloured points corresponds to the respective translations.

with an auxiliary segmentation task, and Bouteldja et al. [77] proposed to integrate a pre-trained segmentation network to regularise CycleGAN training.

However, all of the aforementioned traditional approaches focus on stain-specific models (the segmentation model only works in one stain, and the target stains are translated to match it), a recent shift has emerged towards developing stain-invariant models to achieve multi-stain segmentation. One of the most successful lines of research in this direction is domain adversarial training [78]. This approach aims to extract features that are both domain-agnostic[3] and task-related. For instance, Mei et al. [79] introduced a GAN based method for glomeruli segmentation across two different stains, while Hou et al. [80] proposed enhancing adversarial training by using two discriminators to adversarially align features across different scales. However, a key limitation of these approaches is that the resulting feature extractor tends to be biased towards the domains encountered during training, leading to potential failures when applied to out-of-distribution stains.

Inspired by the success of stain transfer in generating visually plausible images, augmentation based solutions have been proposed to learn stain-invariant models [28, 68]. These approaches have demonstrated that CycleGAN based translations can be used to augment the annotated dataset, resulting in a robust stain-invariant model capable of working across several stains, including out-of-distribution stains. Additionally, other approaches [75, 81] have employed stain transfer to integrate

[3]A term that refers to systems, tools, or methods that have the potential to solve problems across multiple domains or field of applications without requiring significant modifications.

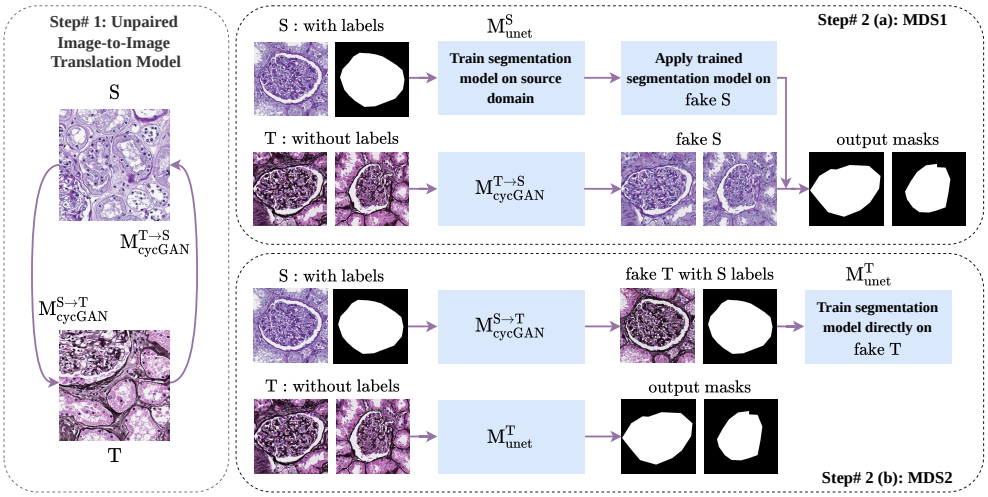


Figure 2.6: Overview of Multi-Domain Supervised architecture.

information from multiple stains, thereby enhancing segmentation or classification performance.

All these methods proposed in the literature demonstrate the effectiveness and widespread application of CycleGAN based stain transfer. However, its inherent limitations are often overlooked or rarely addressed. Therefore, this thesis aims to thoroughly investigate the inherent shortcomings of stain transfer and provides a comprehensive understanding of the specific scenarios in which these stain transfer based approaches may fail. Additionally, this thesis seeks to explore recent advances in deep learning methods to present solutions that can effectively address these shortcomings, ultimately enhancing the performance of multi-stain segmentation methods.

To substantiate and evaluate these objectives, we employ the use case of kidney glomeruli segmentation across multiple stains (including PAS, Jones H&E, CD68, Sirius Red, and CD34), while relying solely on labels from source stain (i.e. PAS). Presently, the prevailing state-of-art approaches in the literature for this application are Multi-Domain Supervised (MDS) [37] and the Unsupervised Domain Augmentation using Generative Adversarial Networks (UDAGAN) approach [28]. These approaches are used in this thesis as benchmarks for kidney glomeruli segmentation, allowing us to evaluate the effectiveness of our proposed contributions. The following subsections will delve into the architectural and training details of these multi-stain segmentation approaches.

2.2.1 MDS

Gadermayr et al. [37] proposed MDS, which results in two separate approaches, MDS1 and MDS2. These approaches aim to generate segmentation masks for images from a domain (stain) that lacks annotations. This is achieved by using a CycleGAN based stain transfer model to translate images from one stain to another. The approach assumes that sufficient labelled data is available in the source

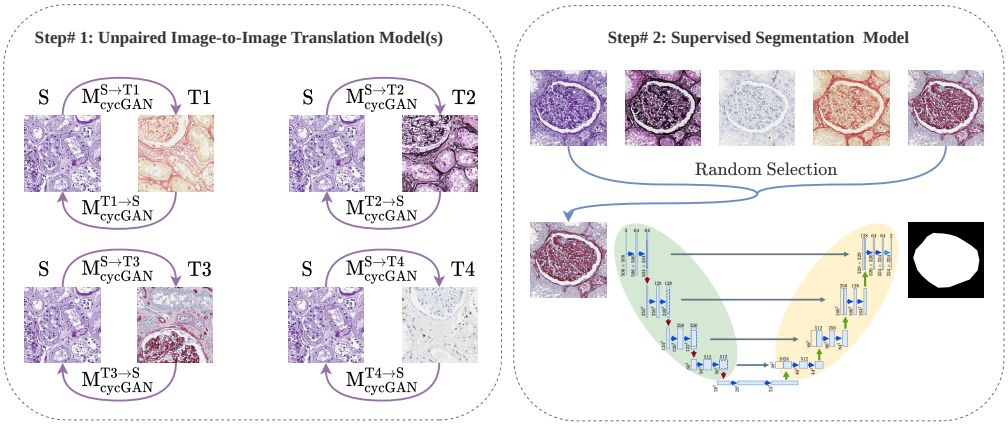


Figure 2.7: Overview of UDAGAN architecture.

stain (S), which is PAS in our case, to train a reliable segmentation network, and unlabelled training data for a target stain (in our experiments these consist of Jones H&E, Sirius Red, CD68, CD34). The primary objective is to generate segmentation masks for images from a target stain without having access to its labels. The UNet [29] model (see Appendix A.1.1 for its architectural details) is employed to train the segmentation model for the PAS stain, because of its remarkable efficacy in segmenting biomedical images [82], specifically for glomeruli segmentation [83]. For both MDS1 and MDS2, a CycleGAN model ($M^{S\leftrightarrow T}_{\text{cycGAN}}$) is trained to translate between the source and target stains, i.e. from $S \rightarrow T$ and $T \rightarrow S$. For each target stain, a separate CycleGAN model must be trained to translate between the source and each target stain.

In MDS1, as shown in Figure 2.6(a), a segmentation model (M^{S}_{unet}) is trained using the training data from source stain and its corresponding labels. Then, the test images from the target stain are translated to match the distribution of the source stain, referred to as the 'fake' source stain (S') using the CycleGAN model ($M^{T\rightarrow S}_{\text{cycGAN}}$) (for an example of these translations, see 2$^{\text{nd}}$ row of Figure 2.4). Finally, the segmentation model (M^{S}_{unet}) is applied to the 'fake' source images, yielding the desired segmentation masks.

In MDS2, Figure 2.6(b), the training data from the source stain is first translated to match the distribution of the target stain, creating 'fake' target stain (T') data (for an example, see 3$^{\text{rd}}$ row of Figure 2.4). The segmentation model (M^{T}_{unet}) is then trained on these fake target stain images using their respective labels (from the source stain). This model is then directly used to segment the original target stain images, resulting in the desired segmentation masks.

2.2.2 UDAGAN

While MDS1 and MDS2 offer promising results for multi-stain segmentation, using labels from only the source stain, they exhibit certain limitations that hinder their broader applicability. Notably, MDS1 requires each target stain to be translated to source stain during test time. Although MDS2 addresses this issue, it requires a

separate segmentation model to be trained for each target stain, which is a time-consuming and computationally expensive procedure. To address these limitations Vasiljević et al. [28] introduced UDAGAN, which aims to combine stain augmentation and adaptation to create a single stain-invariant model that can be directly applied to multiple stains, even those that the model has not seen before (i.e. out-of-distribution stains). Specifically, separate CycleGAN model(s) are trained to enable the translation from the labelled source stain to all of the unlabelled target stains, as depicted in Figure 2.7, Step #1. These CycleGAN model(s) are then used to augment the labelled training set by randomly translating images from the source stain to one of the target stains, Step #2. Given that the translation does not change the overall structure of the image, as depicted in Figure 2.7, the ground truth of the source stain is still valid. As a result, various annotated samples of all available stains are presented to the segmentation model during training, resulting in a single stain-invariant model capable of segmenting various unlabelled target stains. Since UDAGAN generalises and outperforms MDS2 [28], the latter is not evaluated in this thesis.

2.3 Self-Supervised Learning

Despite the aforementioned advances towards reducing the need for labels in the target stains, instead requiring them for only the source stain, their reliance on significant amounts of labelled source data remains a challenge. This becomes particularly pronounced when dealing with specific tissue or tumour types for which an adequate amount of labelled source data may not exist. As a result, the application of existing multi-stain segmentation approaches may prove impractical, creating a significant barrier to the widespread adoption of these methods. In parallel, unlabelled medical imaging datasets are increasing in size by several orders of magnitude [42]. For instance, in histopathology, the advent of whole slide imaging (WSI) scanners has facilitated the production of vast amounts of unlabelled histopathological image data. When unlabelled data is accessible in large quantities, it can be used in limited annotation scenarios to enhance model performance through unsupervised visual representation, particularly self-supervised, learning.

Self-supervised learning was first introduced in 2006 by Bengio et al. [111] but did not become popular until the advent of end-to-end deep neural networks. It laid the groundwork for modern self-supervised learning approaches, which have found application in various fields, such as computer vision, medical imaging, natural language processing, and robotics etc. This approach generates its own supervisory signals (i.e. pseudo labels) and learns useful representations from a pool of unlabelled data by designing a pretext task, thereby eliminating the need for additional human-annotated labels [46]. The representations learned using the pretext task can then be used as initial weights in different downstream tasks where the amount of labelled data is limited. In recent years, a number of self-supervised approaches have been introduced to learn useful representations and they can be grouped into three categories depending on the pretext task: generative, discriminative, and multi-tasking. A categorisation of the most representative self-supervised methods for

Table 2.1: Categorisation of self-supervised learning approaches based on the employed pretext task.

Approach	Authors	Pretext task	Downstream task
Computer Vision			
Generative	Pathak et al. [84]	Image inpainting	Pascal VOC classification, detection, and segmentation
	Zhang et al. [85]	Image colourisation	ImageNet classification, Pascal VOC classification, detection and segmentation
	Donahue et al. [86]	Bidirectional GAN	Pascal VOC classification, detection and segmentation
Predictive	Gidaris et al. [87]	Rotation prediction	ImageNet classification [88], Pascal VOC classification, detection and segmentation [89]
	Noroozi et al. [90]	Jigsaw puzzle	Pascal VOC detection and classification
	Doersch et al. [91]	Relative patch prediction	Pascal VOC detection
Contrastive	He et al. [92]	MoCo	PASCAL VOC and COCO [93] detection and segmentation
	Chen et al. [94]	SimCLR	ImageNet, PASCAL VOC and CIFAR-100 [95] classification
	Grill et al. [96]	BYOL	ImageNet, PASCAL VOC and CIFAR-100 classification
Medical Imaging			
Generative	Prakash et al. [97]	Image denoising	Nuclei segmentation
	Ross et al. [98]	Image colourisation	Surgical instruments segmentation
	Chen et al. [45]	Context restoration	Brain tumour segmentation, Abdominal multi-organ localisation
	Hu et al. [99]	Context encoder	Thyroid nodule segmentation, Liver and Kidney segmentation
Predictive	Taleb et al. [100]	Jigsaw Puzzle	Brain tumour segmentation, Liver segmentation
	Sahasrabudhe et al. [101]	Magnification prediction	Nuclei segmentation
Contrastive	Lu et al. [102]	CPC	Breast cancer classification
	Sowrirajan et al. [103]	MoCo	Tuberculosis detection
	Chen et al. [104]	MoCo	COVID few-shot classification
	Stacke et al. [105]	SimCLR	Histopathology image classification
	Ciga et al. [106]	SimCLR	Histopathology image classification and segmentation
	Stacke et al. [105]	BYOL	Histopathology image classification
	Xie et al. [107]	BYOL extension	Liver and Abdominal organs segmentation, Kidney tumour segmentation
Hybrid	Yang et al. [108]	CS-CO	Histopathology image classification
	Koohbanani et al. [109]	Self-path	Histopathology image classification
	Dong et al. [110]	Multi-task SSL	Heart segmentation

visual representation learning, organised into these three categories, is presented in Table 2.1. This section presents a detailed overview of these categories and the approaches found within them.

2.3.1 Generative Self-Supervised Learning

Generative self-supervised learning aims to model the underlying data distribution $p(x)$ by either reconstructing the original input or by learning to generate new samples from $p(x)$. Auto-encoders and GANs are commonly employed to achieve these objectives. Several such tasks have been proposed, particularly in the field of computer vision and medical imaging, for example image denoising [97, 112], context restoration [45], image colourisation [85, 98], visual field expansion [113], and image inpainting or context encoding [84, 99], etc.

Despite the success of generative self-supervised learning methods, they are often found to be more computationally expensive and complex [94]. Additionally, these methods may not be ideal when the goal is to learn a simple lower-dimensional representation of the data [114]. Moreover, they have an inherent preference for low-level features, which are not effective for discriminative downstream applications [108]. To overcome these limitations, researchers have turned their focus towards discriminative self-supervised learning methods.

2.3.2 Discriminative Self-Supervised Learning

Discriminative self-supervised learning methods focus on learning to distinguish between different transformations or versions of the input data. In earlier stages of developing these methods, researchers focused more on context based (or predictive) methods [115]. These methods aim to learn representations from unlabelled data using a classification or regression based pretext task. Pseudo labels are generated from the data itself and assigned to each image, e.g. by applying a specific transformation, such as rotation. The role of the self-supervised strategy is to accurately predict these pseudo labels. It is important to carefully generate the pseudo labels to facilitate effective feature extraction and to learn useful representations. Numerous predictive pretext tasks, for example rotation prediction [87], jigsaw puzzle [90, 100], relative patch location prediction [91], and magnification prediction [101] etc. have been designed in the field of computer vision, and medical imaging.

Although predictive self-supervised methods have demonstrated significant performance achievements in the computer vision domain, their direct application for medical imaging tasks provides only marginal improvements [46]. To overcome these challenges, contrastive learning has emerged as a powerful discriminative approach, gaining much attention in the field of representation learning in recent years. Its primary objective is to learn representations by comparing pairs of input samples rather than learning from individual samples [114]. It does this by maximising the similarity between similar (positively-paired) samples and minimising it between dissimilar (negatively-paired) samples. Positively-paired samples are generated by applying a set of random augmentations to an input image, resulting in two different augmented views of the same image. Conversely, negatively-paired samples

comprise all other images. The positive pairs are designed to differ but preserve the global features of the input image. This encourages the model to focus on extracting useful representations while discarding irrelevant features. Consequently, the resulting representations tend to be highly discriminative and robust. Contrastive self-supervised learning has been used in computer vision and medical imaging, and is found in methods such as Contrastive Predictive Coding (CPC) [102, 116], Momentum Contrast (MoCo) [92, 103, 104, 117], A Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [94, 105, 106], and Bootstrap Your Own Latent (BYOL) [96], etc.

2.3.3 Multi-Tasking/Hybrid Self-Supervised Learning

In recent years, there has been an increasing trend among researchers to adopt multi-task learning approaches for self-supervised learning. These approaches integrate multiple self-supervised methods, such as predictive, generative, and contrastive, either individually or in a synergistic manner. This integration potentially enhances the model's ability to capture both low-level and high-level features, thereby reducing the limitations and biases inherent in individual self supervision tasks. Moreover, it leads to improved performance in subsequent downstream tasks and allows multiple objectives to be addressed simultaneously. For instance, Graham et al. [118] employed multi-tasking to enhance disease classification, and segmentation within the same framework. Yang et al. [108] employed a combination of generative Cross-Stain (CS) prediction and Contrastive (CO) learning tasks to propose CS-CO to extract more robust representations. Similarly, Zhang et al. [119] combine predictive and contrastive self-supervision tasks in a unified framework and Koohbanani et al. [109] proposed a self-path framework which combines multiple predictive and generative tasks.

2.4 Discussions

This thesis makes two principal contributions: (a) it investigates the inherent limitations of stain transfer-based solutions in digital histopathology, such as multi-stain segmentation, and proposes approaches to address them; (b) it then uses these state-of-the-art approaches in applications where only limited annotations are available for a particular stain. Each of these aspects will be discussed in the subsequent sections.

2.4.1 Stain Transfer

In contrast to natural images, medical images possess complex structures where even minor details can hold significant diagnostic implications. Thus, GAN based artificial image generation and/or translation methods involve the risk of overlooking information necessary for accurate diagnosis. Furthermore, evaluating artificially generated images poses a significant challenge, particularly when assessing their plausibility in clinical applications. For instance, a study performed by Xu et al. [120] provides an interesting perspective on the differences in evaluation between medical

experts (e.g. pathologists) and non-experts (e.g. computer vision researchers). The study found that the medical experts were more adept at identifying errors in the translated stains (images) compared to non-experts, who often perceived the translated stains as nearly indistinguishable from the real ones. As a result, it is suggested that the applicability of these methods in clinical diagnosis is only limited to certain tasks where such risks are acceptable. Particularly, these methods are more suitable for tasks like classification, detection and/or counting of morphologically consistent structures across multiple stains (such as glomeruli in kidney pathology) but not recommended if the decision depends on cell positioning or presence, which could be perturbed during the translation process.

Although GAN based image translation methods, such as CycleGAN, have been widely adopted across various tasks, their significance towards developing multi-stain segmentation approaches is noteworthy. Various extensions to CycleGAN have been proposed to generate more realistic and visually plausible translated images [28, 36, 37, 72, 73] and these studies often rely on visual outputs for comparison [65, 73]. Nevertheless, Vasiljević et al. [53] showed that even visually plausible outputs do not necessarily lead to good downstream task performance, therefore visual assessment alone should not be a criterion for evaluating the quality of the translated images. In parallel, recent studies on the adversarial nature of CycleGAN [48–50] have unveiled their propensity to introduce artefacts imperceptible to humans (even to trained experts) in the translated images (which will be demonstrated in Chapter 3). This underscores the importance for developing measures for translation quality.

2.4.2 Self-Supervised Learning

As discussed earlier, when large quantities of unlabelled data are accessible, it can be leveraged in limited annotation scenarios to enhance the performance of the underlying methods through representation learning, particularly using self-supervised learning [105, 106, 109, 121, 122]. Although numerous self-supervised learning methods have been proposed for the computer vision domain, and both the computer vision and the medical imaging domains (such as digital histopathology) involve the analysis of imagery data, there are significant differences between natural and histopathological images [46]. These differences encompass various aspects, including visual patterns, texture, lighting conditions, and scale. Notably, histopathological images contain information at different magnification levels, whereas the impact of scale on natural images can be largely ignored. Consequently, the direct application of predictive pretext tasks from computer vision may not yield comparable performance. For example, in computer vision, solving a jigsaw puzzle as a predictive pretext task focuses on differentiating between tiles and their positions to learn global semantic representations. However, objects in histopathology are smaller and there is no specific ordering, nor orientation, among them. Therefore, solving a jigsaw puzzle is not relevant [109]. With magnification prediction, the learned model may only focus on size and shape features [108]. Similarly, predicting rotations does not align well with histopathology data, as the arrangement of cells and surrounding structures remain valid even when the image is rotated [106].

To address these aforementioned challenges, researchers have proposed several

contrastive learning based pretext tasks. Recent findings indicate that contrastive self-supervised learning approaches have significantly outperformed both generative and predictive self-supervised learning methods. This trend is evident across various domains, including computer vision [94, 96] and medical imaging, with particular success in histopathology related applications [105, 106, 123]. Moreover, there is a growing trend towards multi-task learning, which aims to integrate strength of various pretext tasks.

Given these developments, this thesis focuses on exploring the potential of state-of-art contrastive and hybrid self-supervised learning approaches. Our primary goal is to significantly reduce the dependence on labelled data for a particular stain, thereby enhancing the applicability of multi-stain segmentation approaches in applications where only limited annotations are available in the source domain. This reduction in label requirements is of huge importance in the field of histopathology, where the acquisition of large-scale annotated datasets is highly challenging, time-consuming, and resource-intensive.

2.5 Conclusions

The development of deep learning solutions in digital histopathology faces significant challenges due to the general scarcity of annotated data and variations in tissue staining. GAN based methods, such as CycleGAN based stain transfer, have made valuable contributions in addressing these challenges and broadening the applicability of these techniques, notably through the development of multi-stain segmentation approaches. However, recent studies in the literature bring forth additional challenges posed by CycleGAN. These include the addition of imperceptible noise which causes domain shift in the translated stains, potentially impacting the final predictions. Research is therefore needed in order to detect and quantify this noise to introduce evaluation metrics for assessing the quality of stain transfer based translated images. These translation approaches can be used to train stain-invariant segmentation models, however they still rely on large amounts of annotated data. State-of-art approaches to self-supervised representation learning offer the potential to reduce this need but work in the literature for digital histopathology segmentation is lacking.

Chapter 3

Stain Transfer Limitations and Optimisation Strategies

While artificially generated stain translations (obtained through stain transfer) visually appear realistic, recent studies [28, 124, 125] have shown the presence of additional information (imperceptible noise) within these translations. This imperceptible noise introduces domain shift in the translated stains (images) [125], potentially impacting the final predictions and thereby raising concerns about the efficacy and deployment of multi-stain segmentation methods in clinical applications. Consequently, there is a pressing need to develop more advanced stain transfer techniques capable of effectively addressing these limitations, ultimately leading to an enhanced performance of multi-stain segmentation approaches. Following this, the rest of the chapter is organised as follows:

- This chapter is motivated in Section 3.1, which explores existing stain transfer based multi-stain segmentation approaches and investigates their efficacy towards addressing inter-stain variations in kidney glomeruli segmentation across multiple stains.
- The properties of common stain transfer approaches that are used in multi-stain segmentation are then explored in Section 3.2 and Section 3.3 to explain the limitations found in the first section.
- Section 3.4 presents several possible techniques to address these stain transfer limitations, thus enhancing the effectiveness of multi-stain segmentation approaches.
- Finally, Section 3.5 summarises the key findings drawn from this chapter.

3.1 Multi-Stain Segmentation Performance

In this thesis, our primary focus lies in exploring stain-specific (MDS1) and stain-invariant (UDAGAN) based multi-stain segmentation approaches. The training details for these approaches are provided in Appendix A.1.3 and A.1.3 respectively. This section delves into evaluating the efficacy of these approaches for the task of kidney glomeruli segmentation across multiple stains using labels from only one (source) stain. To provide a comparison, the segmentation scores of the fully supervised (baseline) models are shown in Table 3.1. These baseline models are trained for a particular stain (e.g. PAS) and applied to a separate unseen test dataset across all stains (e.g. PAS, Jones H&E, Sirius Red, CD68, and CD34). These results reveal

Table 3.1: A comparative analysis between different fully supervised (baseline) and different multi-stain segmentation models. The evaluation is conducted on an independent, unseen test dataset, and the segmentation performance is measured in terms of (F_1) score. For baseline models, each (F_1) score is an average over five different UNet repetitions, whereas for multi-stain segmentation models, each (F_1) score is an average over five different UNet repetitions, each applied to three different CycleGAN repetitions (15 in total); standard deviations are presented in parentheses. The highest (F_1) score achieved across each stain is indicated in bold.

Models	Training Strategy	Test Stains				
		HC Stains			IHC Stains	
		PAS	Jones H&E	Sirius Red	CD68	CD34
Fully Supervised (Baseline)	PAS	0.894 (0.021)	0.062 (0.011)	0.045 (0.037)	0.044 (0.098)	0.056 (0.090)
	Jones H&E	0.009 (0.009)	0.840 (0.029)	0.000 (0.000)	0.000 (0.000)	0.004 (0.007)
	Sirius Red	0.000 (0.000)	0.000 (0.000)	0.865 (0.019)	0.003 (0.003)	0.000 (0.000)
	CD68	0.028 (0.030)	0.016 (0.036)	0.000 (0.000)	**0.836** (0.031)	0.040 (0.081)
	CD34	0.036 (0.025)	0.000 (0.000)	0.000 (0.000)	0.010 (0.012)	**0.888** (0.015)
Multi-Stain Segmentation	MDS1 (Target→PAS)	0.894 (0.021)	0.849 (0.017)	0.870 (0.009)	0.683 (0.043)	0.754 (0.008)
	UDAGAN (Stain-invariant)	**0.901** (0.011)	**0.856** (0.036)	**0.873** (0.025)	0.705 (0.031)	0.799 (0.035)

that glomeruli segmentation is possible for each of the considered stains. However, it is important to note that the features learned by the baseline models are specific to the stain on which they are trained. Consequently, a notable decline in performance is observed when these baseline models are applied to other stains, as shown in Table 3.1. These findings are also illustrated visually in Figure 3.1, where the baseline models trained for a particular stain struggle to segment glomeruli structures across all other stains.

The results obtained with stain transfer based multi-stain segmentation models (including MDS1, and UDAGAN) are also included in Table 3.1. Given that stain transfer is able to produce plausible translations (in accordance with the definition provided on Page 18) and aims to match the distribution of a targeted stain, as shown in Figure 2.5 of Chapter 2, it is reasonable to expect that the baseline model specific to that stain will be able to extract a similar set (or subset) of features in the translated stains. For instance, the baseline model trained on PAS stain should be able to extract similar features in Target→PAS translated stains, which is the

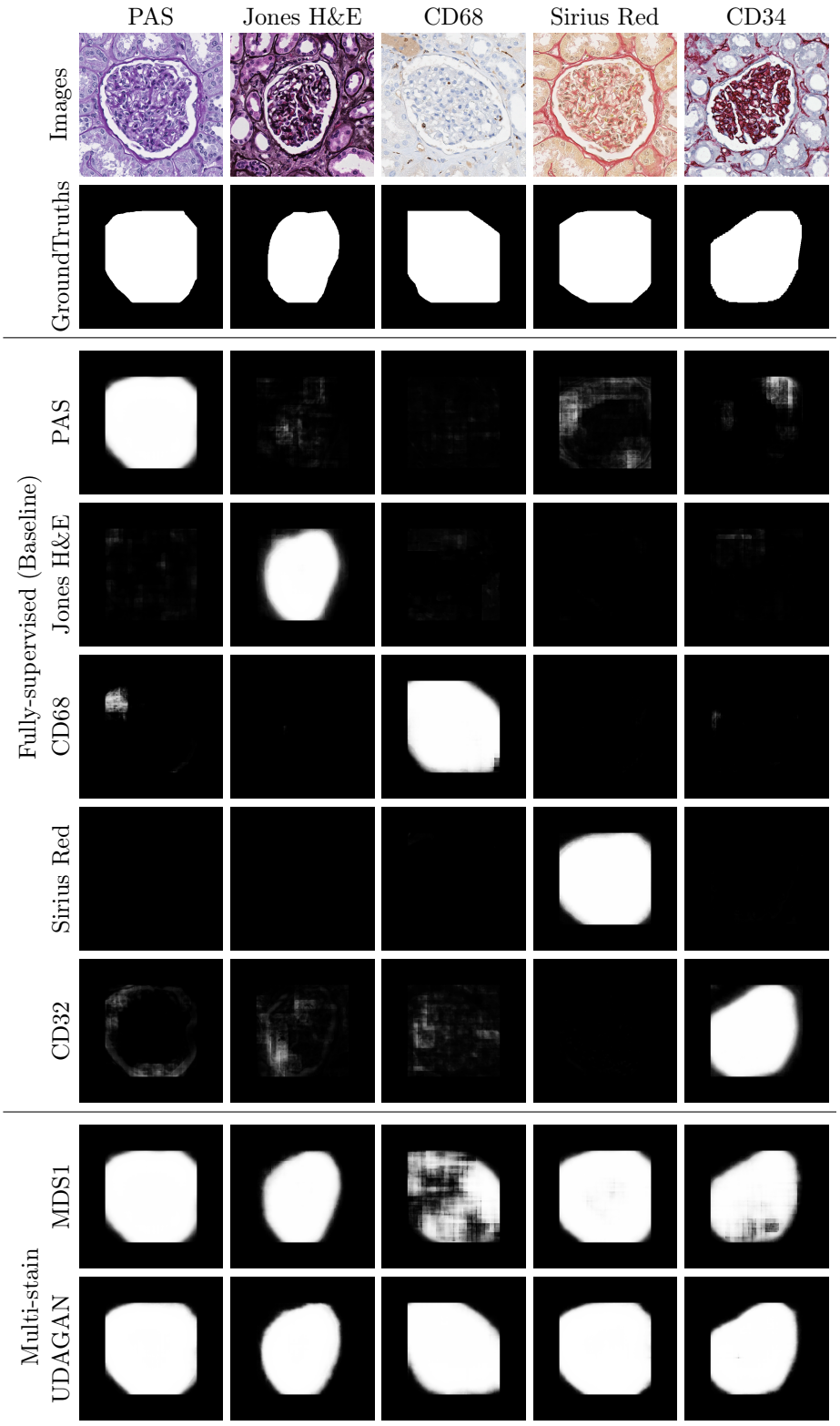


Figure 3.1: Visual comparison of predicted glomeruli segmentation maps across all stains using two different training strategies: Fully supervised (Baseline) and Multi-stain segmentation. Baseline models are trained separately for each stain using their corresponding labels, whereas Multi-stain models are trained using labels from only one (source) stain, demonstrating their ability to generalise across all stains. The input images and their corresponding ground-truths used for this visual comparison comes from a separate unseen test dataset.

case for MDS1. On the other hand, UDAGAN uses translations from the opposite direction (i.e. PAS→Target) to augment the training data which facilitates the learning of more general (stain-invariant) and robust features. As a result, a significant improvement in the segmentation performance of multi-stain segmentation models is observed, as shown in Table 3.1 (6^{th} and 7^{th} row). These finding are also highlighted visually in Figure 3.1, where both MDS1 and UDAGAN demonstrate an ability to effectively segment glomeruli structures across all stains, even when trained using labels from only the PAS stain. This capability sets them apart from baseline models, which do not possess such robustness and generalisability.

While effective, the performance of multi-stain segmentation models significantly varies across different target stains. For instance, for Histochemical (HC) stains, including PAS, Jones H&E, and Sirius Red, these models achieve results comparable to their respective baseline models. However, for Immunohistochemical (IHC) stains such as CD68 and CD34, the multi-stain segmentation models struggle to match the segmentation performance of their corresponding baseline models, despite the plausible translations obtained for these stains as shown in Figure 2.4 of Chapter 2. This discrepancy in performance can be attributed to the distinct characteristics of each staining protocol. For instance, PAS, Jones H&E, and Sirius Red are HC stains that mark general tissue structures, and the translation between, for example, PAS and Jones H&E or Sirius Red is relatively uncomplicated since they are more biologically closer and thus the difference in their highlighted structures is not great. In contrast, IHC stains such as CD68 and CD34 target specific markers such as macrophages and blood vessel (endothelium). In such cases, there exists a significant difference in the highlighted structures of CD68 and CD34 compared to PAS stain, requiring more complicated translations [38]. Given that multi-stain segmentation models are trained using PAS stain labels, these intricacies may hamper the performance of stain transfer based multi-stain segmentation models for CD68 and CD34.

Recent studies [48–50] on the adversarial nature of CycleGAN have provided additional evidence to corroborate these findings. These studies have highlighted that the CycleGAN model is prone to self-adversarial attacks [49].

To understand why, consider the example of PAS and CD68, where a real image from CD68, containing macrophages at specific positions, is translated to PAS using the $G_{CD68 \rightarrow PAS}$ generator. Since the presence of macrophages is irrelevant to the PAS stain, the generator can ignore them. However, when reconstructing CD68 from the translated PAS image using the $G_{PAS \rightarrow CD68}$ generator, the essential details such as macrophage position and their quantity must be preserved, as shown in Figure 3.2, because of the cycle-consistency loss. How is it able to reconstruct these details so precisely when this information is not present in the PAS translation? It must instead embed additional information in the translated image to ensure accurate reconstruction of the CD68 image. This additional information is stored in the form of imperceptible noise.

It is our hypothesis that this noise introduces additional domain shift in the translated images of IHC stains. This domain shift then leads to the drops in performance of multi-stain segmentation models observed in Table 3.1. An important step towards handling this domain shift is the ability to detect and measure it. In the subsequent section, two different methods are proposed to detect and measure

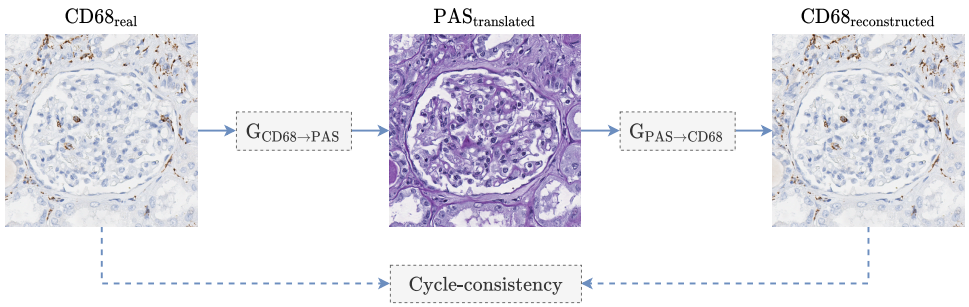


Figure 3.2: Essential details about the position and quantity of macrophages in $\text{CD68}_{\text{real}}$ are preserved in $\text{CD68}_{\text{reconstructed}}$, despite not appearing in $\text{PAS}_{\text{translated}}$.

the domain shift introduced during stain transfer.

3.2 Measuring Domain Shift in Stain Transfer

Given the drop in performance caused by the imperceptible noise (domain shift), it is important to handle this domain shift or at least to estimate when it is likely to affect an algorithm's performance. Therefore, this section concentrates on detecting and quantifying domain shift during stain transfer between a source stain (PAS) and Target→PAS translated stains. To the best of our knowledge, no such work exists for digital histopathology, particularly for kidney glomeruli segmentation.

Two approaches for measuring domain shift are investigated in this regard: (a) the PixelCNN [126] and (b) the Domain Shift Metric [127]. The methodology details for each approach are outlined in the following subsection.

3.2.1 Methods

PixelCNN [126]: is a generative model designed to iteratively generate the pixels of an image. It learns the underlying data distribution in an unsupervised manner by quantifying the pixels of an image x as a product of conditional distributions. As such, it learns to predict the next pixel value given (conditioned on) all previously generated pixels. Formally, this is expressed as

$$p_{\text{CNN}}(x) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1}). \tag{3.1}$$

These conditional distributions are parameterised by a convolutional neural network (CNN) and hence shared across all pixels in the image.

Song et al. [128] have shown that a PixelCNN can be used to detect adversarial attacks in images by visualising differences in the log-likelihood distributions of real (clean) and perturbed images. The authors trained a PixelCNN on a dataset of clean images to estimate their underlying probability distribution. This trained model can subsequently calculate the log-likelihood of any given image, indicating

how well it aligns with the learned distribution of 'clean' images. For this purpose, the authors used bits per dimension (BPD), a normalised measure of log-likelihood. For an image x with resolution $I \times J$ and K channels, BPD is defined as:

$$\text{BPD}(x) \triangleq -\log p_{\text{CNN}}(x)/(I \times J \times K \times \log 2), \tag{3.2}$$

where $p_{\text{CNN}}(x)$ is the probability assigned to the image by the PixelCNN model. Using this formulation, the authors found that the perturbed images consistently exhibited different BPD values compared to the clean images, resulting in their distinct log-likelihood distributions.

We hypothesise that a similar approach can be used to detect the domain shift in the translated images during stain transfer. Specifically, using a PixelCNN model, we aim to visualise the differences in log-likelihood distributions between real PAS stain and translated (Target→PAS) stains. To achieve this, a PixelCNN model is trained on the real PAS stain to model its underlying data distribution (training details are provided in Appendix A.1.5). Once trained, the PixelCNN can be applied to translated Target→PAS stains to determine whether their distributions overlap with that of real PAS stain. We then propose to use the Wasserstein distance [129] ($\mathcal{W}$) to quantify the similarity between the two distributions (PAS and Target→PAS). A smaller $\mathcal{W}$ indicates more similar distributions, thus providing a reliable measure of domain shift.

Domain Shift Metric: The Domain Shift Metric [127] measures the difference between two domains' distributions, referred to herein as Domain Shift Scores or DSS, using the feature representations of a pre-trained model. Consider a CNN with layers $\{l_1, \ldots, l_L\}$. Let $\Phi(x) = \{\phi_{l1}(x), \ldots, \phi_{lk}(x)\}$ such that $\Phi_{lk}(x) \in \{\mathbb{R}^{h \times w}\}$ denotes the filter activations at layer l and filter k. The mean value of each $\Phi_{lk}(x)$ is calculated as

$$c_{lk}(x) = \frac{1}{hw} \sum_{i,j}^{h,w} \Phi_{lk}(x)_{i,j}. \tag{3.3}$$

Let $p^S_{c_{lk}}(x)$ denote a distribution of $c_{lk}(x)$ over the source stain S and $p^T_{c_{lk}}(x)$ denotes the same over the translated (Target→PAS) stain T, then the domain shift metric (DSM) is defined as

$$\text{DSM}(p^S, p^T) = \frac{1}{k} \sum_{i=1}^{k} \mathcal{W}\left(p^S_{c_{lk}}, p^T_{c_{lk}}\right), \tag{3.4}$$

where $\mathcal{W}$ is the Wasserstein distance [129] between $p^S_{c_{lk}}(x)$ and $p^T_{c_{lk}}(x)$, which tends towards zero when S and T are similar.

3.2.2 Results

As was shown in Table 3.1, MDS1 experiences a degradation in performance when applied to translated (Target→PAS) images of stains CD68 and CD34, compared to their respective baseline models. Building upon recent studies [48–50], we hypothesis

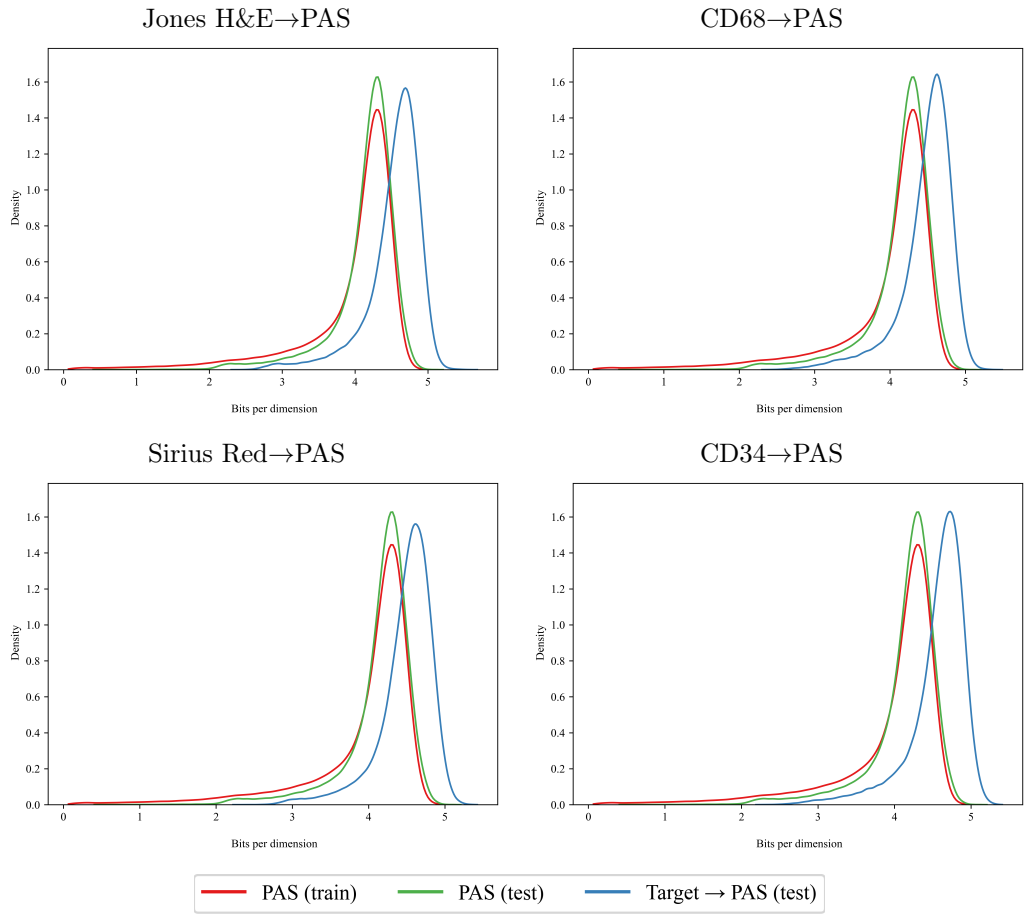


Figure 3.3: PixelCNN based visualisation of domain shift in translated Target→PAS stains w.r.t. real PAS train and test sets.

that this performance degradation is caused by a domain shift introduced in the translated stains during stain transfer.

To test this hypothesis, the PAS trained PixelCNN model is first validated using the PAS training data and an unseen PAS test set, see Figure 3.3. It is found that, their log-likelihood distributions follow the same order of magnitudes, with a low Wasserstein distance of 0.0879 (averaged over 5 sets of 1000 randomly sampled patches), indicating low domain shift. The log-likelihood distributions of the translated Target→PAS stains are also included in this figure and they clearly show that there is a domain shift compared to the overlapping PAS train/test distributions. Consequently, the Wasserstein distance between PAS train and translated Target→PAS stains is observed to be relatively large, see Table 3.2, indicating a significant domain shift in the translated stains.

By using the pre-trained segmentation model to extract feature representations of the source stain (PAS), the domain shift can also be measured using the domain shift metric (previously seen in Equation (3.4)). The respective DSS for all translated (Target→PAS) stains are also included in Table 3.2.

Table 3.2: Average Wasserstein Distance and Domain Shift Scores of 5 sets of 1000 randomly sampled patches for the Target→PAS translated stains; standard deviations are in parentheses.

Methods	Test Stains				
	PAS	Jones H&E→PAS	Sirius Red→PAS	CD68→PAS	CD34→PAS
Wasserstein Distance	0.087 (0.003)	0.537 (0.012)	0.493 (0.004)	0.481 (0.006)	0.580 (0.005)
Domain Shift Scores	0.032 (0.017)	0.097 (0.008)	0.119 (0.003)	0.248 (0.002)	0.138 (0.002)

Now that we can detect and measure what appears to be the domain shift, we investigate whether it is correlated with the full slide segmentation (F_1) scores of MDS1[1] (provided in Table 3.1: 6^{th} row). Figure 3.4 presents the scatter plots of the DSS and the MDS1 based F_1 scores for each translation, revealing a very strong correlation of -0.9135. This strong negative correlation indicates that as the domain shift (measured in terms of DSS) increases, the segmentation performance of MDS1 significantly decreases. In contrast, the PixelCNN based Wasserstein distance demonstrates a moderate correlation (based on the criteria specified by [130]) of $-0.5390^{\star}$ with the full slide segmentation scores of MDS1.

3.2.3 Discussions

Although, a correlation is observed between domain shift measured using PixelCNN and the segmentation performance of MDS1, it is moderate in strength compared to that observed with the DSM. The primary reason for this difference lies in the nature of PixelCNN, which is a completely unsupervised approach and does not use representations specifically tailored for the task to be performed (segmentation in our case). Conversely, DSM uses feature representations from the pretrained segmentation model to measure domain shift and therefore integrates task-specific knowledge. As a result, DSM is more sensitive to the type of domain shifts that affect segmentation performance. Apart for the use of the pre-trained segmentation model in DSM, both of these approaches measure domain shift in an unsupervised manner.

Both approaches exhibit a considerable correlation with whole slide image (WSI) segmentation scores, even though the domain shift is calculated on a small subset of the data. Therefore, these findings offer a way to estimate the average performance of pretrained neural networks trained on source data and applied to unseen target data (for the same task), without requiring any expert opinions or ground-truths. For instance, consider a model trained on a specific task using a dataset from a

[1]MDS1 is a multi-stain segmentation model that allows the applicability of a segmentation model (trained on source stain) to various other target stains by translating them to source stain.

[⋆]In our publication [52], an error was made in calculating the PixelCNN based correlation. The corrected correlation value is provided here.

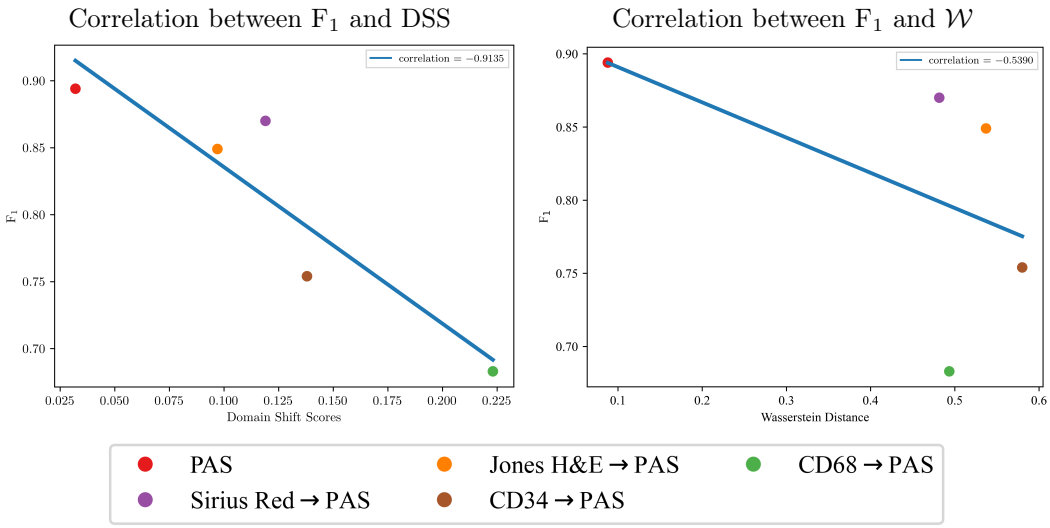


Figure 3.4: Correlation between segmentation (F_1) scores of the whole test slides translated to PAS and the average domain shift (measured in terms of both Domain Shift Scores and Wasserstein Distance) of 5 sets of 1000 randomly sampled test patches.

particular lab or hospital. To assess the model's generalisability and robustness to datasets from other hospitals, traditional methods often depend on expert opinions or ground truth comparisons. However, in situations where ground truths are unavailable, our findings suggest that computing the domain shift between small samples of the two datasets can estimate the trained model's average performance. A minimal domain shift indicates a higher possibility of successful generalisation of the pretrained model to the new data. This can give an indication of whether such results should be relied upon or not. We recommend using DSM to measure the average performance of pretrained models. PixelCNN, on the other hand, is recommended for measuring general domain shifts between source and target data, especially in the absence of task-specific pretrained models.

While our work primarily focused on detecting and/or measuring domain shift introduced during stain transfer, the proposed solution is general and can be applied to multiple other related applications within the field of computer vision and medical imaging. In the next section, we will demonstrate one other use of these measures in the field of digital histopathology.

3.3 Is Visual Inspection Reliable?[2]

In light of recent developments in stain transfer, several improvements to CycleGAN have been proposed [36, 37, 72, 73, 131]. For example, Gadermayr et al. [37] suggested to replace the generator's ResNet [132] architecture with a UNet [29].

[2]This analysis was conducted in collaboration with Jelica Vasiljević (a former PhD student from SDC research team, ICube).

Building on this idea, Cai et al. [131] propose to incorporate additional downsampling and upsampling layers into the UNet architecture to facilitate the learning of higher-level semantic information. Moreover, de Bel et al. [31] proposed the addition of an extra loss function, while Cai et al. [131], Zhang et al. [133] proposed to incorporate different normalisation techniques (more details are explained later in this section). However, due the absence of ground-truth translation targets, these studies often rely on visual outputs for comparison [53, 65, 73]. Despite the plausible visual appearance of these translations, the findings presented in Section 3.2 indicate that visual comparison is unreliable because of the possible presence of imperceptible domain shift within the translations. This ultimately affects the final predictions of a pretrained model. As a result, visual inspection can be misleading and may not determine the difference in the quality of the resulting translations.

Beside visual inspection, several quantitative metrics have been used across various studies [36, 69, 131, 134] to compare different stain transfer methods. These metrics include the Structural Similarity Index Measure (SSIM) [135], Peak Signal to Noise Ratio (PSNR) [136], and Fréchet Inception Distance (FID) [137] etc. However, SSIM and PSNR require ground-truth translations (i.e. paired samples) for evaluation. To fulfil this requirement, researchers have suggested using consecutive tissue slides stained with different stainings as ground-truths [36, 131]. This approach, however, has certain limitations, including variability in tissue structure between slides, inconsistencies in staining procedure, and challenges in slide registration (alignment). Consequently, obtaining ground-truths for stain transfer is not straightforward. FID is capable of evaluating the quality of generated (translated) images without requiring paired samples. Nevertheless, its applicability to medical imaging raises concerns due to its reliance on an ImageNet based pre-trained network, particularly InceptionV3 [138]. To this end, several approaches [139–141] have suggested to adapt FID for medical imaging by using feature extractors pre-trained on medical datasets. However, recent findings by Woodland et al. [142] challenge these approaches and presented novel evidence that medical imaging based pretrained networks do not inherently improve FID performance and may even compromise its reliability. Therefore, there is a pressing need for a metric that can accurately evaluate various stain transfer based translation methods in histopathology without requiring paired samples, while also preserving diagnostically relevant features when using any pretrained model.

Given that our proposed methods provide a promising strategy to measure domain shift during stain transfer (without relying on ground-truths), we propose that it can be used as an evaluation metric to provide a comparison between different stain transfer methods. Notably, this can be achieved by measuring domain shift in their respective translated images, and the method that results in the lowest domain shift will be considered as the most effective in producing high quality and meaningful translations (i.e. from Target→Source and vice-versa).

To further investigate this, we focused on the use case of employing different normalisation techniques within the underlying architecture of CycleGAN and different stain transfer models are created by replacing the normalisation layers in both the discriminators and the generators. Specifically, the original CycleGAN architecture [47] uses Instance normalisation [143], which has been employed in various

state-of-art stain transfer methods [28, 37, 144]. On the other hand, several other studies Shrivastava et al. [73], Cai et al. [131] proposed to use Batch normalisation [145], while Mahapatra et al. [146] proposed to use Group Normalisation [147], and [133, 148] proposed to use CycleGAN without any normalisation. Further details about these normalisation techniques are explained in the subsequent subsection.

3.3.1 Normalisation Techniques

In the case of 2D images, the features computed by a model's layer, denoted as f, is a 4D tensor $f = (N, C, H, W)$ where N denotes batch size, C is the number of channels and H and W are spatial height and width. A normalisation layer normalises f such that

$$\hat{f} = \frac{f - \mu_{norm}}{\sigma_{norm}}, \tag{3.5}$$

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 [145], 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 [149], calculates μ_{norm} and σ_{norm} over the (C, H, W) axes, normalising features for each sample in a batch separately. Instance normalisation [143] computes μ_{norm} and σ_{norm} across the (H, W) axes, thus normalising features for each sample and each channel separately. Similar to Layer normalisation, Group normalisation [147] computes μ_{norm} and σ_{norm} over the (H, W) axes, but instead of normalising over all channels, a specific number of groups of adjacent channels is chosen. Thus, when the number of groups is equal to 1, Group normalisation becomes Layer normalisation, and it reduces to Instance normalisation when the number of groups is equal to the number of channels. Thus, 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 possible due to the minimal number of filters used in the CycleGAN convolutional layers).

3.3.2 Experimental Settings

The above mentioned normalisation techniques, in addition to CycleGAN without using any normalisation (referred to as None) and Layer normalisation [149], are used to train the respective stain transfer models. Once the models are trained, their performance in producing high quality translated images (Target→PAS in our case) is evaluated by measuring domain shift between the real source (PAS) images and the translated (Target→PAS) images. Later, it will be examined whether this domain shift correlates with the segmentation performance of a pre-trained UNet (trained on PAS stain) applied to different stain translations (Target→PAS), commonly known as MDS1 segmentation. Since MDS1 employs a task-specific segmentation pretrained model, we used the DSM, presented in Equation (3.4), in this evaluation.

Table 3.3: Domain shift for stain transfer based Target→PAS translated images using different normalisation techniques. Each score is an average over 5 sets of 1000 randomly sampled patches, with 5 repetitions of pre-trained models each applied to 1 repetition of CycleGAN model. The overall lowest domain shift (averaged across all target stains) is indicated in bold.

Normalisation Techniques	Test Stains				
	Jones H&E→PAS	Sirius Red→PAS	CD68→PAS	CD34→PAS	Overall
Instance	0.097	0.119	0.248	0.138	**0.150**
Batch	0.749	0.560	0.796	0.472	0.644
Group_8	0.158	0.194	0.247	0.215	0.203
Group_{16}	0.139	0.214	0.226	0.197	0.194
Group_{32}	0.127	0.147	0.264	0.155	0.173
Layer	0.154	0.148	0.309	0.170	0.196
None	0.134	0.293	0.490	0.274	0.298

3.3.3 Results

Table 3.3 presents the corresponding domain shift values across various stain transfer models created using different normalisation techniques. These values clearly demonstrate that Instance normalisation based stain transfer achieves the lowest domain shift of 0.150 compared to other normalisations. Given the strong negative correlation between DSM and the average performance of a pretrained model, as highlighted in Section 3.2, it is expected that the overall performance (average over all test stains) for MDS1 segmentation using Instance normalisation will surpass that of other normalisations. Conversely, the highest domain shift value of 0.644 is observed with Batch normalisation, indicating that it will significantly effect the performance of the pretrained model.

Additionally, the results for MDS1 segmentation across various translated stains obtained using different stain transfer models are provided in Table 3.4. These results show that the Instance normalisation achieves the highest overall performance, with an average F_1 score of 0.789 across all target stains, while the lowest performance, with an average F_1 score of 0.312, is observed using Batch normalisation. These results are consistent with the expected outcomes derived from the domain shift values, presenting a very strong correlation of -0.9685 between domain shift and F_1 scores of MDS1 segmentation using different normalisation techniques, as illustrated in Figure 3.5.

To assess the visual quality, the translations (Target→PAS) produced by each stain transfer models are provided in Figure 3.6. Visually, these translations (except Batch normalisation) look plausible and it certainly would not be possible to choose which model produces better or worse performance. However, the results for MDS1 segmentation provided in Table 3.4 demonstrate significant differences in the performance across different types of normalisation. This indicates that, in the absence of ground-truths, relying solely on visual inspection to assess stain transfer model's performance is unreliable and may lead to inaccurate conclusions. However, our

Table 3.4: MDS1 based segmentation (F_1) scores for glomeruli segmentation across various target stains using different normalisation based stain transfer models. The evaluation is conducted on an independent, unseen test dataset. The F_1 score is an average over five different pre-trained segmentation models (UNet), each applied to three different repetitions of stain transfer models (15 in total); standard deviations are provided in parentheses. The overall highest F_1 score (averaged across all target stains) is indicated in bold.

Normalisation Techniques	Test Stains				
	Jones H&E→PAS	Sirius Red→PAS	CD68→PAS	CD34→PAS	Overall
Instance	0.849 (0.017)	0.870 (0.009)	0.683 (0.043)	0.754 (0.008)	**0.789** (0.019)
Batch	0.339 (0.059)	0.508 (0.041)	0.002 (0.001)	0.400 (0.067)	0.312 (0.042)
$Group_8$	0.848 (0.011)	0.810 (0.006)	0.308 (0.101)	0.628 (0.040)	0.648 (0.039)
$Group_{16}$	0.849 (0.011)	0.800 (0.036)	0.486 (0.060)	0.650 (0.039)	0.694 (0.036)
$Group_{32}$	0.815 (0.007)	0.807 (0.017)	0.546 (0.049)	0.737 (0.015)	0.751 (0.022)
Layer	0.816 (0.014)	0.832 (0.005)	0.167 (0.046)	0.754 (0.024)	0.642 (0.022)
None	0.770 (0.003)	0.730 (0.035)	0.250 (0.028)	0.747 (0.047)	0.624 (0.028)

proposed approach provides a reliable measure to evaluate different stain transfer methods.

3.4 Improving Stain Transfer

As highlighted in Section 3.1, CycleGANs can result in the addition of imperceptible noise in the translated images. This leads to a drop in performance of multi-stain segmentation methods, such as MDS1 and UDAGAN, particularly when they are applied to translated immunohistochemical stains. The domain shift metrics detailed in Section 3.2 have shown that this noise can be detected and measured. It remains to be seen therefore whether this metric can be used as a loss function when training the CycleGAN. This loss, which we call Domain Shift Loss (DSL), can act as a novel self-guided strategy towards learning translations with minimal domain shift, ultimately improving stain transfer and thereby enhancing the performance of multi-stain segmentation approaches.

Moreover, several other developments [48–50, 150, 151] in the field of computer vision have been introduced to address the above mentioned limitation of CycleGAN.

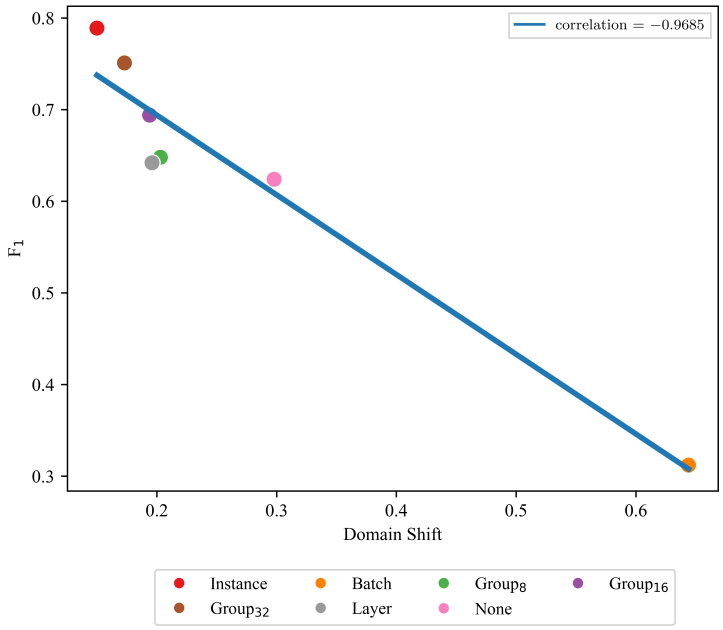


Figure 3.5: Correlation between segmentation (F_1) scores and the measured domain shift for different normalisation based stain transfer models.

Specifically, Wang et al. [50] and Shao et al. [151] proposed to attenuate the cycle-consistency constraint to reduce its potential for inducing noise in the translated images. Conversely, Nizan et al. [150] suggested to learn unpaired image-to-image translations without relying on such constraints. Additionally, Bashkirova et al. [49] adopted the idea from adversarial training [152] to add random Gaussian noise in the translated images to facilitate more accurate translations. Chu et al. [48] proposed to introduce an additional image channel in the training process to embed hidden information (noise) separately. Although, Nizan et al. [150] and Shao et al. [151] have shown superior performance compared to other methods, they are designed for applications involving significant geometric changes and shape deformations between the source and the target domains, which is not our primary focus. Finally, in a model specific to histopathology, Bouteldja et al. [77] suggest to integrate the pre-trained segmentation model into CycleGAN to improve stain transfer.

While these developments have shown significant improvements compared to the original CycleGAN model, they are primarily developed for computer vision applications and have never been applied to histopathology related tasks (except Bouteldja et al. [77]). Therefore, in this section, in addition to our proposed approach of employing DSL in CycleGAN, we propose to use the methods presented by Bashkirova et al. [49], Chu et al. [48], and Bouteldja et al. [77]. The use of these approaches serves two purposes: first, to provide a comparative analysis with our proposed approach of employing DSL; and second, to investigate their effectiveness for enhancing multi-stain segmentation approaches, which to the best of our knowledge, have not been explored previously.

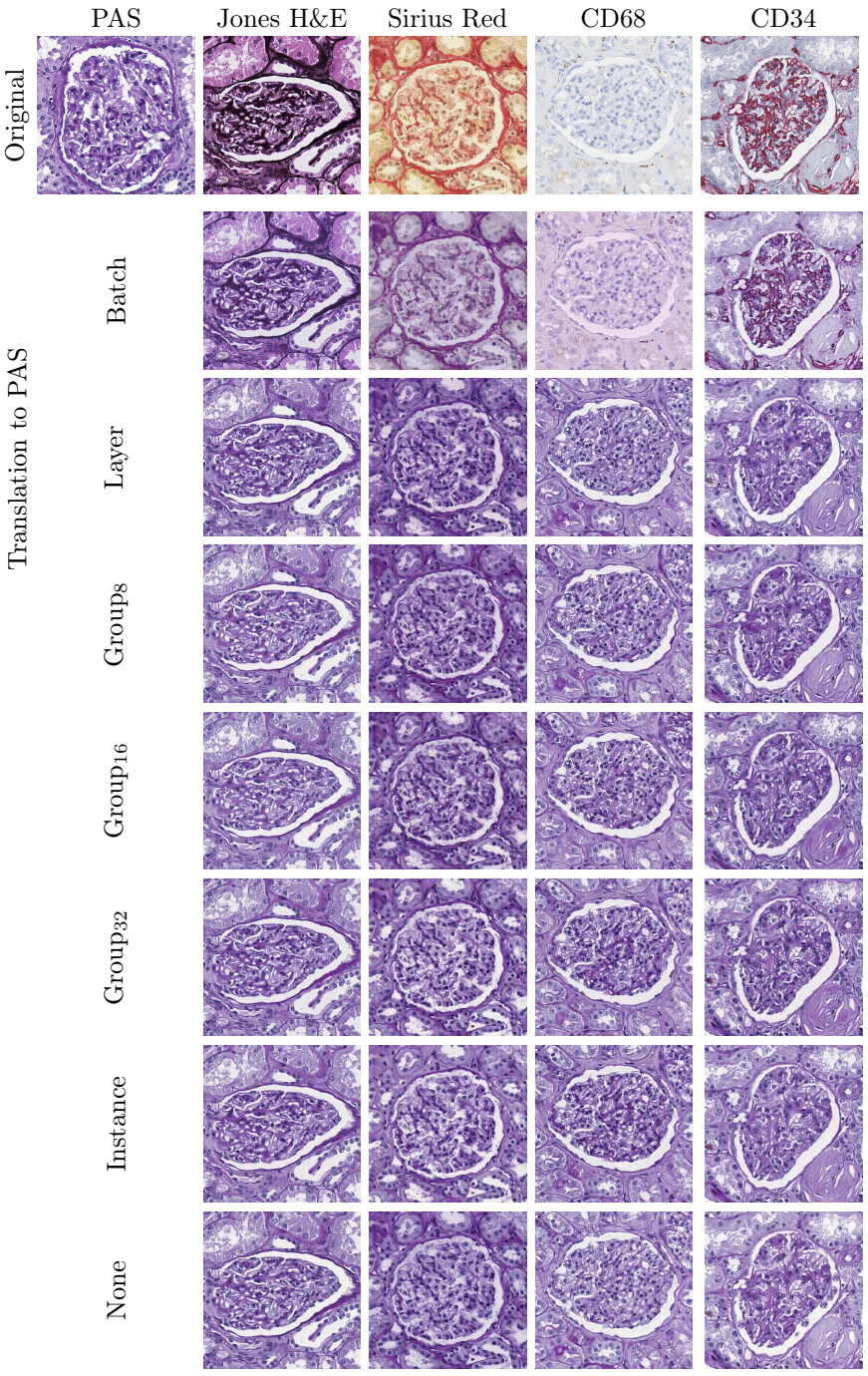


Figure 3.6: Target stain patch translated to PAS using CycleGAN based stain transfer models trained with different normalisation techniques.

3.4.1 Methods

The original architectures, as proposed by the authors [48, 49, 77], were used herein and are provided in Appendix A.1.6, A.1.7, and A.1.8. The architectural details of the Domain Shift Loss are follows.

3.4.1.1 CycleGAN with Domain Shift Loss

The original CycleGAN architecture is modified by integrating the DSM (defined in Equation (3.4)) as a loss ($\mathcal{L}_{\text{dsl}}$) to minimise the impact of domain shift in the translated images. The DSM uses a pretrained segmentation model (trained only for the source stain), therefore, in the bidirectional framework of CycleGAN, it is integrated only in one direction. Particularly, only images from the source stain and translated (target→source) images are provided to the DSM to calculate the loss ($\mathcal{L}_{\text{dsl}}$). Additionally, inspired by the idea of Bouteldja et al. [77], $\mathcal{L}_{\text{dsl}}$ is also calculated between images from the source stain and their respective reconstructions and identity mapping, such that

$$\begin{aligned}\mathcal{L}_{\text{dsl}} &= \mathcal{L}_{\text{dsl,translated}} + \mathcal{L}_{\text{dsl,cyc}} + \mathcal{L}_{\text{dsl,id}} \\ &= \mathbb{E}_{s\sim A}\mathbb{E}_{t\sim B}[\text{DSM}(p^s, p^{G_{BA}(t)}) + \text{DSM}(p^s, p^{G_{BA}(G_{AB}(s))}) + \text{DSM}(p^s, p^{G_{BA}(s)})].\end{aligned} \tag{3.6}$$

This modification results in the following CycleGAN loss function:

$$\begin{aligned}\mathcal{L}_{\text{CycleGAN}}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{\text{adv}}(G_{AB}, D_B, G_{BA}, D_A) \\ + w_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) \\ + w_{\text{id}}\mathcal{L}_{\text{id}}(G_{AB}, G_{BA}) \\ + w_{\text{dsl}}\mathcal{L}_{\text{dsl}}(G_{AB}, G_{BA}).\end{aligned} \tag{3.7}$$

This overall objective function guides the translated images to more closely align to the source domain, thereby reducing their domain shift.

3.4.2 Experimental Setup

While training CycleGAN with Gaussian Noise (referred to as w/ Gaussian Noise), proposed by Bashkirova et al. [49] and defined in Equation (A.1), various levels of noise were tested to identify the best value for maximising segmentation performance across all target stains. Therefore, a separate hyperparameter study was conducted using a range of σ values: 0.0125, 0.025, 0.05, 0.075, 0.1, 0.3., 0.5, 0.9. The results were averaged over 1 CycleGAN and 5 UNet repetitions to select the best value of σ. Once the best value was determined, the models were re-trained, and the segmentation (F_1) scores were evaluated on a separate unseen test set by averaging over 3 CycleGAN and 5 UNet repetitions (15 experimental repetitions in total). Similarly, following range of weight values, 1.0, 5.0, 10.0, were evaluated to select the best w_{dsl} and w_{seg} across all target stains for CycleGANs integrating the DSL and the Segmentation Loss proposed by [77] (see Equation (A.3)).

Table 3.5: MDS1 based segmentation (F_1) scores for glomeruli segmentation across various target stains using different CycleGAN based stain transfer methods. The evaluation is conducted on an independent, unseen test dataset. Each F_1 score is an average over 5 different UNet repetitions, each applied to 3 different CycleGAN repetitions (15 in total), with standard deviations presented in parentheses. The overall highest (F_1) score (averaged across all target stains) is indicated in bold.

Training Strategy	Test Stains				
	HC Stains		IHC Stains		Overall
	Jones H&E	Sirius Red	CD68 PAS	PAS	
CycleGAN* (baseline)	0.844 (0.026)	0.860 (0.023)	0.643 (0.031)	0.747 (0.021)	0.774 (0.025)
w/ Gaussian Noise [49]	0.865 (0.016)	0.878 (0.015)	0.669 (0.026)	0.749 (0.028)	**0.790** (0.021)
w/ Self-supervision [77]	0.840 (0.027)	0.866 (0.021)	0.686 (0.020)	0.753 (0.024)	0.786 (0.021)
w/ Extra-channels [48, 77]	0.862 (0.019)	0.871 (0.020)	0.634 (0.037)	0.669 (0.041)	0.759 (0.029)
Ours	0.849 (0.024)	0.862 (0.022)	0.694 (0.021)	0.763 (0.012)	**0.792** (0.020)

3.4.3 Results

In this section, the previously mentioned CycleGAN based translation methods are evaluated. These methods are compared not only with the original CycleGAN (baseline) method but also with each other using the MDS1 multi-stain segmentation approach.

Table 3.5 presents the results for MDS1 using each translation method. The results show that, for HC stains, all methods demonstrate similar or improved segmentation performance compared to the baseline method. This improvement is more pronounced when using 'CycleGAN with Gaussian Noise' compared to others. Conversely, for IHC stains, performance improvements are observed with all methods except 'CycleGAN with Extra-channels'. Furthermore, these results show that the performance gains are more substantial for IHC stains, particularly CD68, compared to HC stains. This is because the original CycleGAN method struggles with these stains since they are more biologically distinct from the source (PAS) stain, which introduces more noise in the translated images, resulting in reduced baseline performance. However, the proposed methods manage to (to some extent) alleviate this limitation by mitigating such noise from the translated stains. Notably, the highest overall performance (average across all target stains) is achieved using both

*A slight difference is noted in the CycleGAN based MDS1 results reported in Table 3.5 compared to those presented in Table 3.1. This is because the experiments in Table 3.5 are implemented using Tensorflow 2, while the experiments in Table 3.1 are implemented using the Keras framework (which is now deprecated and has been integrated into Tensorflow 2).

Table 3.6: Domain shift measured on Target→PAS translated images using different CycleGAN based stain transfer methods. Each score is an averaged over 5 sets of 1000 randomly sampled patches, with 5 repetitions of pre-trained models each applied to 3 repetition of CycleGAN model.

Training Strategy	Test Stains				Overall
	HC Stains		IHC Stains		
	Jones H&E	Sirius Red	CD68	CD34	
CycleGAN (baseline)	0.108 (0.009)	0.122 (0.020)	0.270 (0.006)	0.155 (0.005)	0.164 (0.010)
w/ Gaussian Noise [49]	0.095 (0.003)	0.118 (0.005)	0.255 (0.016)	0.139 (0.009)	**0.152** (0.008)
w/ Self-supervision [77]	0.109 (0.008)	0.119 (0.009)	0.279 (0.018)	0.156 (0.008)	0.166 (0.011)
w/ Extra-channels [48, 77]	0.128 (0.010)	0.125 (0.005)	0.279 (0.014)	0.172 (0.009)	0.176 (0.010)
Ours	0.099 (0.006)	0.117 (0.006)	0.261 (0.023)	0.146 (0.011)	**0.156** (0.012)

'CycleGAN with Gaussian Noise' and our proposed 'CycleGAN with DSL'.

Moreover, the domain shift (presented in Table 3.6) measured for each method is consistent with the achieved segmentation performance, despite being calculated on a very small subset of the translated stains. This indicates that methods with minimal domain shift tend to achieve the highest performance. Specifically, 'CycleGAN with Gaussian Noise' and our proposed 'CycleGAN with DSL', both showing minimal domain shift, thereby outperforming all other methods.

3.4.4 Discussions

Although, 'CycleGAN with Gaussian Noise' has demonstrated superior performance for HC stains, its efficacy is less pronounced for IHC stains compared to other methods, such as 'CycleGAN with self-supervision' and our proposed 'CycleGAN with DSL'. This discrepancy in performance can be attributed to the Gaussian noise augmentation technique proposed in 'CycleGAN with Gaussian Noise', which is particularly more robust in handling low-amplitude perturbations [49]. Consequently, it appears to be particularly effective for scenarios involving subtle differences, as is the case with HC stains, which are biologically more similar to the source (PAS) stain. Conversely, IHC stains are more biologically distinct from PAS and may induce high-amplitude perturbations, which are less effectively addressed by Gaussian noise augmentation. This observation opens up potential avenues for the research community and could motivate further improvements to enhance this method's robustness.

A similar discrepancy in the performance of 'CycleGAN with Extra-channels' has been observed. While the authors of this approach claimed that it improves

the translation between domains involving significant biological differences (i.e. IHC stains) and is not as beneficial when the translations are straightforward (i.e. with HC stain), we have observed the opposite behaviour. It performs comparatively well for HC stains but significantly reduces performance for IHC stains. We believe that this is because certain shared (common) features between the source and target stains, which should be present in the translated images, could be placed into the additional channel for reconstruction. For instance, when translating from CD68→PAS, common features (i.e. macrophages) present in both PAS and CD68 stains should appear in the resulting CD68→PAS image. However, it is possible that this approach may misinterpret these features as noise and include them in the additional channel, especially when they must also be present in the reconstructed CD68 image. Consequently, a pre-trained model on real PAS images, which have seen these shared features during training, could experience a decline in performance if they are not adequately captured in the translated CD68→PAS images.

Conversely, 'CycleGAN with Self-supervision' and our proposed 'CycleGAN with DSL' tend to achieve a better performance for IHC stains as they use a pretrained segmentation model (trained on the source stain) to provide guidance towards learning features in the translated images that are more closely aligned with the source stain. In conclusion, we recommend using 'CycleGAN with Gaussian Noise' and 'CycleGAN with Extra-channel' methods when the translations are performed between biologically similar stains, and using 'CycleGAN with Self-supervision' and 'CycleGAN with DSL' when the translation are conducted between biologically different stains.

3.5 Conclusions

This chapter investigated the potential of existing stain transfer methods in digital histopathology to address inter-stain variations. Notably, CycleGAN has emerged as a standard approach for stain transfer and has been widely adopted by numerous state-of-art methods [28, 36, 37, 70–73]. While effective, these methods are prone to hallucinating features during stain transfer, thereby introducing additional impacts of domain shift in the translated images, ultimately affecting the final predictions.

An important step towards handling this domain shift is the ability to detect it. To this end, two different approaches, PixelCNN and DSM, were proposed in this chapter. These approaches have shown a great ability to detect and/or measure the domain shift in translated images. DSM, in particular, exhibited a very strong correlation of -0.9135 between the full slide segmentation scores of MDS1 and the domain shift, despite being measured on a very small subset of the data. These findings paves the way to infer the average performance of a pre-trained model when applied to unseen target data (for the same task) without requiring any expert opinion or ground-truth. Although, our proposed solution primarily focused on measuring domain shift in stain transfer, it is general and can detect any kind of domain shift between source and target data.

Furthermore, this chapter highlighted the drawbacks of various strategies commonly employed to evaluate state-of-the-art stain transfer methods and presented a

more reliable measure for this evaluation without the need of ground-truth translation targets and/or paired samples.

Finally, these findings were extended to propose a CycleGAN based stain transfer approach that reduces the domain shift introduced within the translated stains. This approach was shown to be effective in improving the overall segmentation performance across all target stains.

Chapter 4

Self-Supervised Learning

The previous chapter explored the inherent limitation of introducing noise during stain transfer and introduced different strategies for minimising such noise, resulting in the enhanced performance of stain transfer based multi-stain segmentation approaches. While these methods eliminate the need of labels in the target stain, it is crucial to recognise that these methods rely heavily on a large amount of labelled data from the source stain. However, acquiring a sufficient amount of labelled data for source domain[1] is still challenging in various medical disciplines. For instance, in histopathology, for certain tissue or tumour types, sufficient labelled datasets for the source stain may not be readily available. Nevertheless, recent advances in computer vision and medical imaging have led to a significant increase in the size of datasets (usually unlabelled) by several orders of magnitude [42]. For instance, in histopathology, the advent of whole slide imaging (WSI) scanners has facilitated the production of vast amounts of (unlabelled) histopathological image data. In such situations, where unlabelled data is accessible in large quantities, it can be employed in limited labelled scenarios to enhance model performance through unsupervised representation learning [87].

Considering the aforementioned problems, self-supervised learning (SSL) – a subset of unsupervised representation learning – is one of the most feasible solutions and has gained huge attention in recent years. This approach generates its own supervisory signals (i.e. pseudo labels) and learns useful representations from a pool of unlabelled data by designing a pretext text, thereby eliminating the need for additional human-annotated labels [45]. The representations learned using the pretext task can then be leveraged in different downstream tasks where the amount of labelled data is limited. Several studies have demonstrated the effectiveness of self-supervised learning approaches in various histopathology tasks such as detection and classification [102, 105, 108, 109, 113, 121, 122, 153, 154]. However, only limited attempts have been made to incorporate the advances of self-supervised learning for histopathology segmentation [106, 118, 155, 156], particularly multi-stain segmentation.

This chapter uses three approaches to self-supervised representation learning: SimCLR [94], BYOL [96], and our proposed extension to CS-CO [108], called HR-CS-CO. SimCLR is selected due to its widespread adoption in histopathology related downstream tasks [105, 106]. While SimCLR's performance heavily depends on augmentation and the number of negative pairs, demanding huge computing resources [94, 105], recent methods have shown that negative-pairs are not essential for contrastive learning [96, 157, 158]. One such method is BYOL [96]. Finally, to assess

[1] As noted in Chapter 1, the term "stain" and "domain" are equivalent.

the benefits of hybrid SSL strategies, a novel extension to CS-CO [108], called HR-CS-CO is included. These representations will then be refined in several downstream tasks to enhance the performance of single-stain and multi-stain kidney glomeruli segmentation approaches in the presence of limited labelled data.

Following this methodology, the rest of the chapter is organised as follows:

- Section 4.1 presents the architectural and training details of the employed self-supervised learning methods. In the context of learning representations, SimCLR and BYOL are general methods and employ a straightforward learning approach, which makes them adaptable to a wide range of computer vision and medical imaging domains. For instance, in histopathology, both SimCLR and BYOL can be employed to learn representations across various staining protocols. In contrast, CS-CO is a designed for histopathology, particularly for H&E stained images, which makes it stain/domain specific. Therefore, in this section, several modifications are proposed to extend CS-CO, to remove this limitation, making it stain independent.

- Section 4.2 presents the use of the learned representations (through the aforementioned self-supervised learning methods) to enhance the effectiveness of various downstream tasks. These tasks are taken to be: single-stain histopathological segmentation using UNet [29] in the presence of limited labels for each stain, and multi-stain histopathological segmentation using MDS1 [37] and UDAGAN [28] in the presence of limited labels for only one (source) stain.

- Section 4.3 presents an in-depth comparison of these representation learning methods tailored for each specified downstream task. Additionally, it provides valuable insights into the efficacy of these methods, highlighting their strengths and weaknesses.

- Finally, Section 4.4 highlights the key findings drawn from this exploration of self-supervised learning methods.

4.1 Methods

As mentioned earlier, this chapter focuses on three different self-supervised pretext methods to extract the most effective and meaningful representations from a pool of unlabelled data to improve the performance for different downstream segmentation tasks in the presence of limited labels, as illustrated in Figure 4.1. These methods are SimCLR [94], BYOL [96], and a novel extension to CS-CO [108]. This extension (presented in Section 4.1.1.3) overcomes its stain specific formulation (it was proposed for the H&E stain) and is called HR-CS-CO. The original architectures, as proposed by the authors of SimCLR and BYOL, were used in this study, with details provided in Section 4.1.1.1 and 4.1.1.2 respectively.

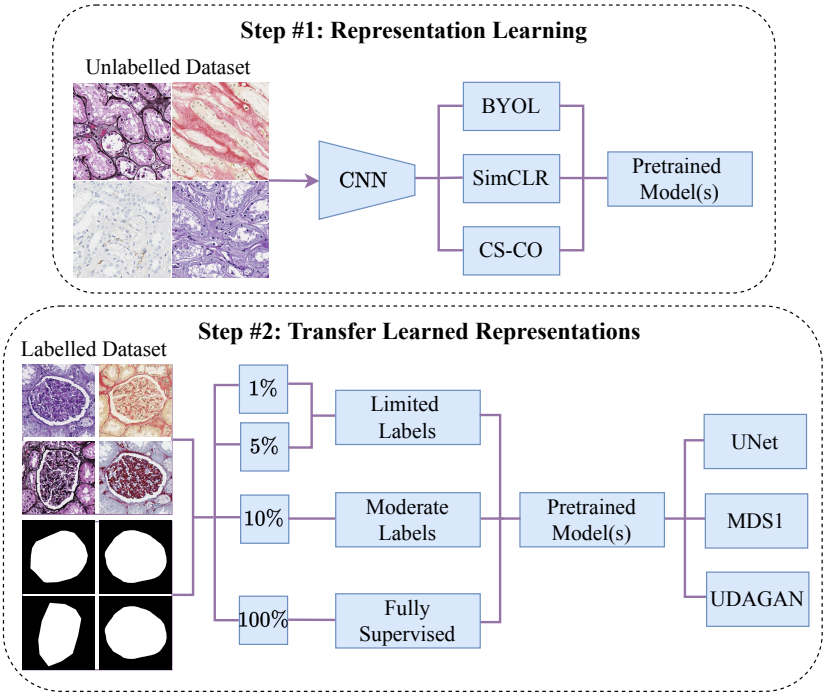


Figure 4.1: Self-supervised learning workflow in histopathology. Step #1: Different self-supervised learning methods are applied to learn representations from a large unlabelled dataset. Step # 2: The learned representations are then refined by fine-tuning on several splits of labelled data for a variety of downstream tasks.

4.1.1 Network Architecture Details

4.1.1.1 SimCLR

A simple framework for contrastive learning of visual representations, or SimCLR in short [94], learns representations by maximising the agreement between two augmented views of the same image via a contrastive loss in the latent space. As illustrated in Figure 4.2, the framework starts with a probabilistic data augmentation module f_{aug} that generates two positively correlated views, x_i and x_j, of a given data sample x. A set of base augmentations [94] are adopted, including random cropping and resizing with a large scale range of (0.1-1.0), flipping, grey-scale, Gaussian blur, and random colour distortions. Based on the findings of [105], two additional augmentations were incorporated, grid distort and grid shuffle, which have demonstrated their effectiveness for histopathology applications. Further details and examples of the augmentations can be found in Appendix B.1. The augmented views, x_i and x_j are then transformed into their corresponding representations, h_i and h_j by employing a convolutional neural network (CNN) base encoder f_θ, where θ is the weight parameters. Subsequently, a projection head g_θ consisting of a multi-layer perceptron (MLP) is employed to map the extracted representations

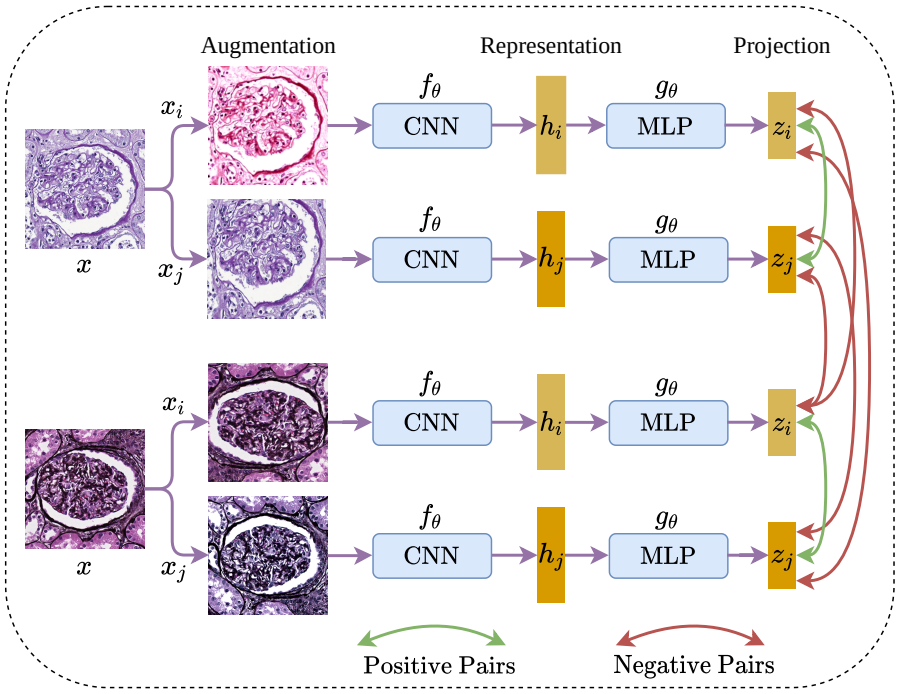


Figure 4.2: Overview of the SimCLR architecture inspired by [94, 106].

into a lower dimensional embedding space in which the contrastive loss is applied. The MLP comprises two dense layers with ReLU activation for the first layer and linear activation for the second layer to obtain $z_i = g_\theta(h_i)$ and $z_j = g_\theta(h_j)$ respectively. In [94], it was observed that comparing z_i and z_j was more effective for learning representations than directly comparing h_i and h_j. Finally, as suggested by the authors of SimCLR, to optimise the entire network *NT-Xent* (the normalised temperature-scaled cross-entropy) loss function is defined, such that

$$\ell_{i,j} = -\log \frac{\exp\left(\text{sim}(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(z_i, z_k)/\tau\right)}, \tag{4.1}$$

where τ is the temperature parameter that weights different samples and facilitates learning from hard negative samples and $\mathbb{1}$ is the indicator function, which outputs 1 when $k \neq i$ and 0 otherwise. The term $\text{sim}(z_i, z_j) = z_i^\top z_j / \|z_i\| \|z_j\|$ represents the dot product between ℓ_2 normalised z_i and z_j, which corresponds to the cosine similarity. This loss function aims to maximise the agreement between positive pairs of augmented images, while minimising it for other images in the same batch (negative pairs). In each training step with a batchsize of $2N$, each augmented image has one positive and $2(N-1)$ negative pairs.

4.1.1.2 BYOL

Bootstrap Your Own Latent Representation (BYOL) is an implicit contrastive learning approach introduced by Grill et al. [96]. Unlike other contrastive methods,

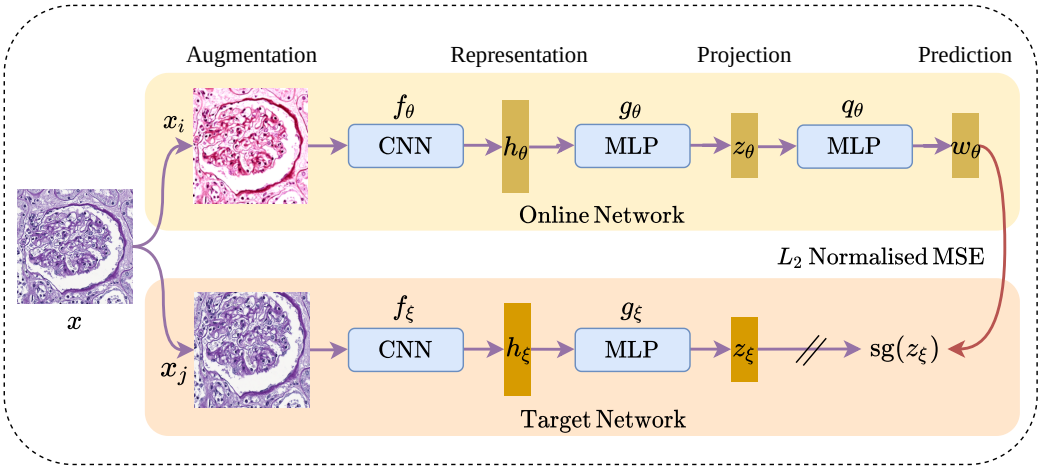


Figure 4.3: Overview of BYOL architecture inspired by [96].

BYOL does not rely on negative pairs and is more robust to the choice of augmentations. The core idea of BYOL revolves around iterative bootstrapping of the network's output to serve as a target for an enhanced representation. To achieve this, BYOL employs two neural networks, Online and Target, which interact and learn from each other. As depicted in Figure 4.3, the Online network is a trainable network comprising a CNN based encoder f_θ, an MLP based projection head g_θ, and a prediction head q_θ. On the other hand, the Target network is a non-trainable network that is randomly initialised. It has the same architecture as the Online network, but has a different set of weight parameters ξ. The Target network provides the regression targets used to train the Online network, and its parameters ξ are updated through an exponential moving average of the Online parameters θ. Considering a target decay rate $\tau \in [0, 1]$, the following update is carried out after each training step:

$$\xi \leftarrow \tau\xi + (1 - \tau)\theta. \tag{4.2}$$

To train the BYOL network, a data augmentation module f_{aug} is used to generate two distinct augmented views x_i and x_j from the input image x. This module incorporates similar augmentations as those used in SimCLR, see Appendix B.1. The Online network processes the first augmented view x_i and outputs a representation h_θ, a projection z_θ, and a prediction w_θ. Similarly, the Target network outputs a representation h_ξ, and a target projection z_ξ from the second augmented view x_j. Notably, the prediction head is solely applied to the Online network, resulting in an asymmetric architecture between the Online and Target pipelines. Following that, both w_θ and z_ξ are normalised using ℓ_2 norm and then fed into a mean squared error (MSE) loss function for optimisation, such that

$$\mathcal{L}_{\theta,\xi} = \|\bar{w}_\theta - \bar{z}_\xi\|_2^2 = 2 - 2 \cdot \frac{\langle w_\theta, z_\xi \rangle}{\|w_\theta\|_2 \cdot \|z_\xi\|_2}. \tag{4.3}$$

The loss $\mathcal{L}_{\theta,\xi}$ is made symmetrical by separately feeding x_j to the Online network and x_i to the Target network. This allows the computation of another loss function

$\tilde{\mathcal{L}}_{\theta,\xi}$. During each training step, a stochastic optimisation step is performed to minimise $\mathcal{L}^{BYOL}_{\theta,\xi} = \mathcal{L}_{\theta,\xi} + \tilde{\mathcal{L}}_{\theta,\xi}$ with respect to θ only, while ξ remains unaffected by applying a stop-gradient (sg), as illustrated in Figure 4.3.

4.1.1.3 CS-CO

CS-CO [108] is a hybrid SSL method, designed particularly for Haematoxylin and Eosin (H&E) stained histopathology images. It contains two stages: cross-stain prediction and contrastive learning. The cross-stain prediction, which is a generative task, captures low-level general features, e.g. nuclei morphology and tissue texture, that are valuable for histopathology analysis [108]. To facilitate this, stain-separation [159] is applied to H&E stained images to extract the single-dye channels, Haematoxylin (H_{ch}) and Eosin (E_{ch}). Afterwards, cross-stain prediction is employed to learn the relationship between H_{ch} and E_{ch}, using two separate auto-encoders, H2E and E2H, where H2E predicts E_{ch} from H_{ch}, and vice-versa.

Nevertheless, CS-CO has certain limitations, restricting its broader applicability. For example, histopathological images often use different staining protocols and reagents to highlight different tissue structures (e.g. PAS, Jones H&E, Sirius Red, CD68, and CD34, as used in this study). The stain separation method integral to CS-CO struggles with ImmunoHistochemical (IHC) stains [159]. Particularly, it fails to accurately extract the individual H_{ch} and DAB_{ch} (Diaminobenzidine) from CD68. Furthermore, in some cases, histopathological stains contain more than two dyes, e.g. Jones H&E, where CS-CO's stain-separation approach would yield three separate channels—J_{ch} (Jones), H_{ch}, and E_{ch}—which cannot be handled in CS-CO's architecture.

To address these limitations and extend the applicability of CS-CO across multiple stainings, we propose to modify its stain-separation strategy as outlined in Figure 4.4. Particularly, we exploit the fact that Haematoxylin is often used as a counterstain in histopathology, and therefore exists in many stains. This was first exploited by Lampert et al. [27] as a strategy for stain invariant segmentation. Here, however, we use it to extract a common Haematoxylin channel, H_{ch}, which highlights cell nuclei, via image deconvolution [160]. The remaining information is retained as a 'Residual' channel (R_{ch}) as illustrated in Fig. 4.4, step #1, and Fig. 4.5, capturing tissue structure highlighted by the other stain components, such as glycogen, collagen, macrophages, and endothelial cells, etc, depending on the staining used. Therefore, all stain (containing H_{ch}) can be included by modelling them as H_{ch} and R_{ch}. In the rest of the chapter, we refer to this modified version of CS-CO as HR-CS-CO.

Moving forward, in the cross-stain prediction stage, two separate auto-encoders H2R and R2H are trained as shown in Fig. 4.4 (Step #2). H2R learns to predict R_{ch} from H_{ch}, and R2H performs the inverse task. Both share the same architecture but have different weights. For simplicity, ϕ_{h2r} and ψ_{h2r} is used to represent the encoder and decoder for H2R (and similarly for R2H). Additionally, the combination of ϕ_{h2r} and ϕ_{r2h}, and ψ_{h2r} and ψ_{r2h}, are denoted as the HR encoder and decoder respectively. The mean square error (MSE) loss is computed to evaluate the dissimilarity

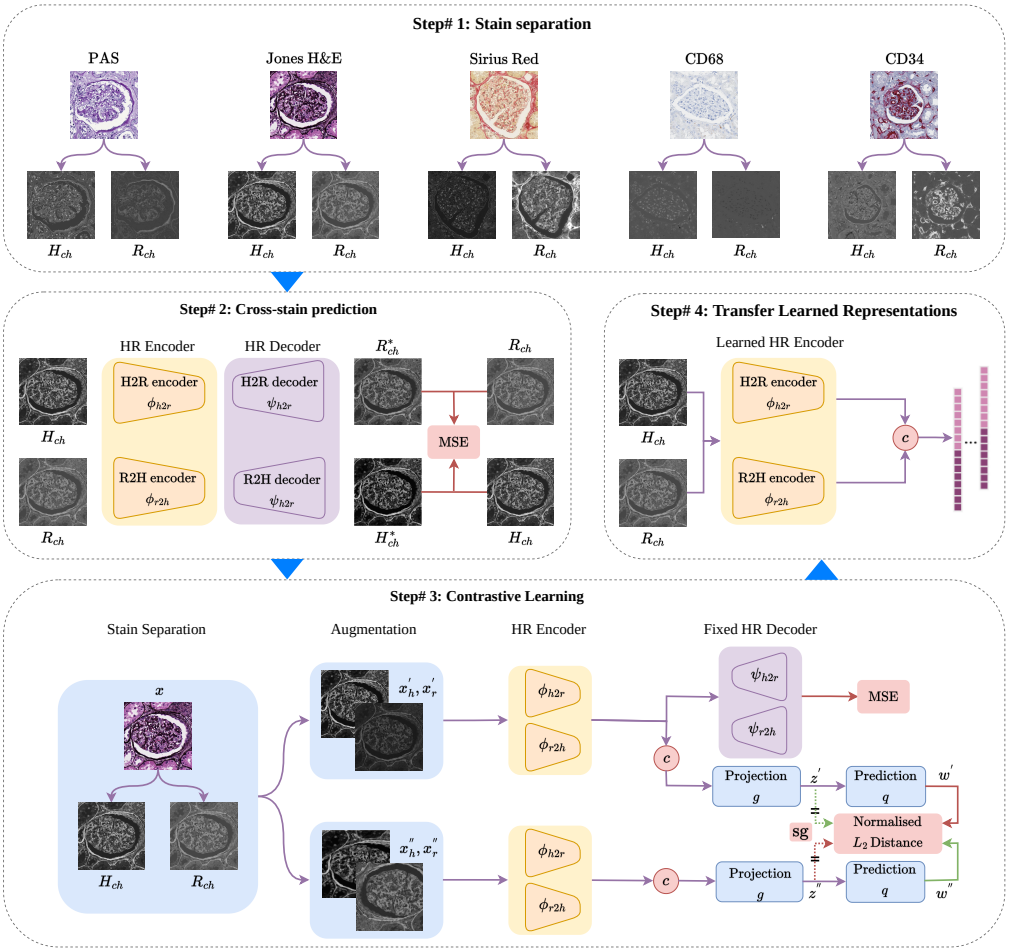


Figure 4.4: Overview of the proposed HR-CS-CO architecture. In Step# 1, stain-separation is applied to separate the H_{ch} and R_{ch} from each each stain. In Step# 2, the cross-stain prediction is employed as a generative task, learning to predict H_{ch} from R_{ch} and R_{ch} from H_{ch}. Lastly, in Step# 3, contrastive learning is used as discriminative task on the augmented views of H_{ch} and R_{ch} to learn the final representations. Here, the weights for ϕ and ψ are initialised to those learnt during cross-stain prediction (i.e. Step #2), thereby combining the strength of generative and discriminative learning.

between the real (H_{ch}, R_{ch}) and predicted (H^*_{ch}, R^*_{ch}) images, such that

$$\mathcal{L}_{cs} = (H_{ch} - H^*_{ch})^2 + (R_{ch} - R^*_{ch})^2 , \tag{4.4}$$

where

$$R^*_{ch} = \psi_{h2r}(\phi_{h2r}(H_{ch})) \text{ and } H^*_{ch} = \psi_{r2h}(\phi_{r2h}(R_{ch})).$$

Once the training of the cross-stain prediction is complete, the trained two-branched

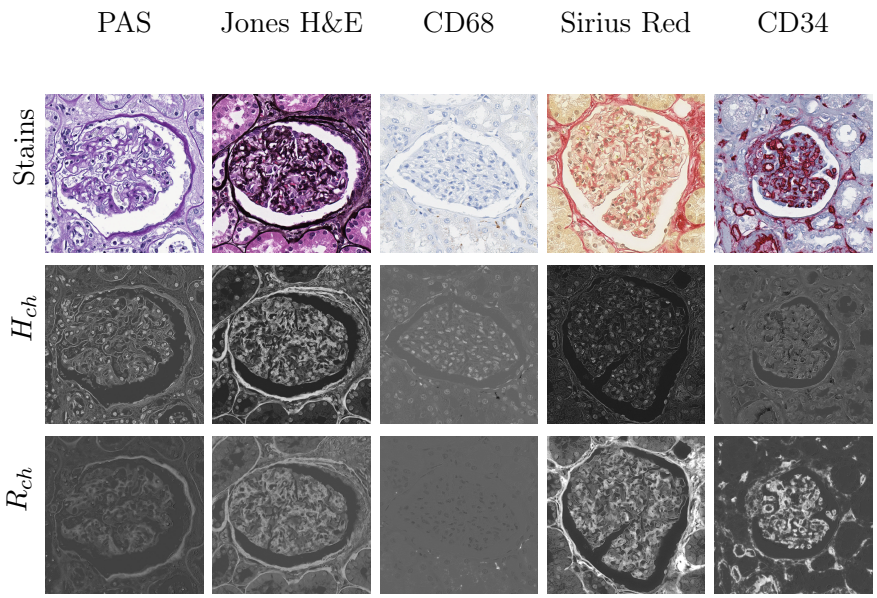


Figure 4.5: Visualisation of Haematoxylin (H_{ch}) and Residual (H_{ch}) channels extracted from each of the stains used in this study.

auto-encoder is expected to be sensitive to low-level features.

Next, contrastive learning is employed as the final step to exploit the benefits of discriminative high level features. Motivated by Chen et al. [157], the model is reorganised into a Siamese architecture [161], consisting of the HR encoder (ϕ), a projection head (g), and a prediction head (q). The parameters are shared between the two branches of the Siamese architecture. Both g and q are multi-layer perceptrons (MLP) with the same architecture. To prevent mode collapse, the HR decoder (ψ) is retained in one branch as a non-trainable regulator. Instead of employing random initialisation, the weights for ϕ and ψ are initialised to those learnt during cross-stain prediction (i.e. Step #2), thereby combining the strength of general low-level and discriminative high level features.

During contrastive learning, H_{ch} and R_{ch} are extracted from a given input image x to give (x_h, x_r). For each data sample, a data augmentation module f_{aug} is used to generate two distinct augmented views: (x'_h, x'_r) and (x''_h, x''_r). This augmentation module includes various augmentations such as flipping, random cropping and resizing, Gaussian blur. Given that the input images are grey-scale, colour-based augmentations are not applicable, however, drawing inspiration from [27], we incorporate an additional augmentation method called stain variation by using colour deconvolution [162]. This augmentation involves modifying the intensities of H_{ch} and R_{ch} using a factor α, sampled from $[-0.25, 0.25]$, and a bias β, sampled from $[-0.05, 0.05]$. These specific values were chosen as they result in realistic output, as depicted in Figure 4.6.

Subsequently, each augmented pair is then separately fed into the Siamese network, where it is encoded by ϕ_{h2r} and ϕ_{r2h}. The resulting outputs are pooled and concatenated to form a single vector. This vector is processed by the projection

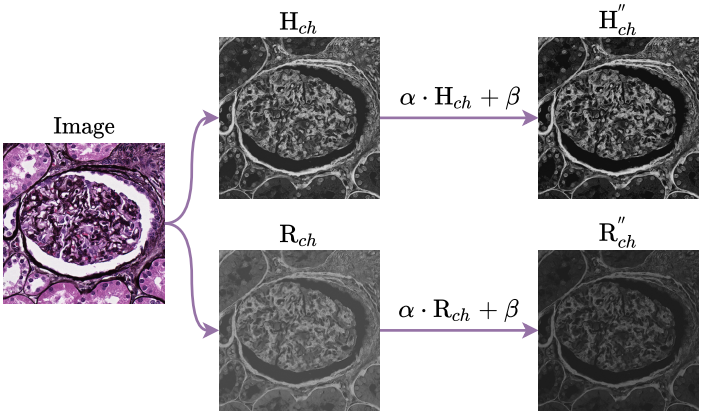


Figure 4.6: Stain-variation augmentation. From left to right: the process begins by decomposing an image into its corresponding hematoxylin (H_{ch}) and residual (R_{ch}) channels. Subsequently, each channel undergoes individual modification using a random factor α and bias β. The modified versions are represented as H''_{ch}, and R''_{ch}.

head g to obtain (z', z'') and the prediction head q to obtain (w', w''). The learning process involves minimising the symmetric loss, such that

$$\mathcal{L}_{co} = \frac{1}{2}\|\bar{w'} - \bar{z''}\|_2^2 + \frac{1}{2}\|\bar{w''} - \bar{z'}\|_2^2, \tag{4.5}$$

where $\bar{w'}$, $\bar{w''}$, $\bar{z'}$, and $\bar{z''}$ represent the ℓ_2 normalised versions of w', w'', z', and z'', respectively. This encourages w' to be similar to z'' and w'' to be similar to z'. Prior to computing the loss, the stop-gradient (sg) operation should be applied to z' and z'' which detaches them from the computational graph. During contrastive learning, the frozen pre-trained HR decoder (ψ) continues to use the outputs of the HR encoder (ϕ) for image reconstruction. To avoid collapse, the HR encoder must maintain the necessary information for image reconstruction, by satisfying Equation 4.4. As a result, the total loss is formulated such that

$$\mathcal{L}_{csco} = \mathcal{L}_{cs} + \gamma\mathcal{L}_{co}, \tag{4.6}$$

where γ represents the weight coefficient.

4.1.2 Training Setup

4.1.2.1 Dataset

The self supervised pre-training dataset includes image patches extracted from the training and validation WSIs, as detailed in Section 1.4.1 of Chapter 1. The image patches are obtained in an unsupervised manner using a uniform sampling strategy. To ensure a representative and balanced distribution of image samples across different patients, 15,000 and 1,000 image patches were randomly sampled from the training and validation WSIs of each stain respectively. This results in a final

dataset of 75,000 training patches and 5,000 validation patches. By using such an extensive dataset, our primary objective is to improve the learning capabilities of the self-supervised models, thereby facilitating more accurate and robust downstream application, particularly in situations where the availability of labelled data is very limited.

4.1.2.2 Training Details

It is a common practice to employ a CNN-based encoder for self-supervised image representation learning, as is the case with SimCLR and BYOL. HR-CS-CO, however, relies on a CNN-based auto-encoder. Considering the proven performance of the UNet, particularly in the context of glomeruli segmentation [83], as shown in Chapter 3, its encoder component was used for SimCLR and BYOL. Similarly, its encoder and decoder components were used as the auto-encoder in HR-CS-CO. In each of these networks, the extracted representations are subsequently projected into a lower-dimensional space using a multi-layer perceptron (MLP). Finally, using the self-supervised validation dataset in an unsupervised scenario, the best trained model is selected based on the validation loss, which is then used for the downstream applications of single-stain and multi-stain kidney glomeruli segmentation. In line with common practice [105], the training process for the self-supervised networks (SimCLR, BYOL, and HR-CS-CO) was performed only once due to computational and time constraints. The training details for each approach are as follows.

SimCLR: The training setup proposed in the original paper [94] was used. For a gradual adjustment of the learning rate, a warm-up period of 10 epochs was used, followed by a decay using the cosine decay schedule. Recently, Stacke et al. [105] have shown that smaller batch sizes are preferable when using SimCLR in histopathology, particularly when dealing with few classes, since it reduces the risk of false negatives and therefore a batch size of 256 was used. This also allowed the higher resolution of histopathological images to be used. Following [105], we trained SimCLR for 200 epochs.

BYOL: A similar training procedure as described in the original BYOL paper [96] was used. The absence of negative samples in BYOL's training paradigm allows it to attain performance parity with SimCLR despite using smaller batch sizes. Therefore, a batch size of 256 was chosen and the model was trained for 200 epochs. Since BYOL exhibits susceptibility to poor initialisation and training collapse [96], batch-normalisation (BN) is incorporated in the encoder to enhance the robustness and stability of the learning process as highlighted by Richemond et al. [163].

HR-CS-CO: Since the concentration of the hematoxylin channel (H_{ch}) can vary between different stainings, we train separate HR-CS-CO models for each stain. This is done in two stages: (1) referred to as cross-stain prediction, the model is trained for 100 epochs with a batch size of 32 using the Adam optimiser (initial learning rate of 0.001, which, based on the validation loss and a patience of 10 epochs, is reduced by a factor of 0.1); (2) referred to as contrastive learning, the model is trained again for 50 epochs and a batch size of 128 using the Adam optimiser (learning rate of

Table 4.1: Training data with different percentages of labelled glomeruli for each staining.

% of Labels	Stainings				
	PAS	Jones H&E	CD68	Sirius Red	CD34
1%	6	5	5	6	5
5%	33	31	26	32	28
10%	66	62	52	65	56
100%	662	621	526	651	565

0.001 and a weight decay of 1×10^{-6}). To prevent over-fitting, early stopping is implemented at both stages.

4.2 Minimally Supervised Histopathology Segmentation

Once the self-supervised model has been trained using one of the aforementioned pretext tasks, it can be refined by fine-tuning to a variety of downstream tasks, such as image classification, object detection, image segmentation, etc. These downstream tasks are usually accomplished by supervised training and need a significant amount of labelled data. Self-supervised pre-training proves particularly valuable when there is limited availability of labelled data, as it enables models to learn representations and gain meaningful insights from extensive amounts of unlabelled data. Similarly, in this Section, pre-training serves as an initial guide to enhance the performance of single-stain and multi-stain glomeruli segmentation in histopathology images with limited labelled data.

Single-Stain Segmentation involves the segmentation of glomeruli regions from histopathological images of each stain using their respective labels. The UNet [29] model is often used for this task due to its proven success in segmenting biomedical images [82], specifically for glomeruli segmentation [83]. However, the UNet is a fully supervised convolutional neural network (CNN) and relies heavily on a substantial amount of labels in each stain. To investigate the potential of self-supervised learning methods to enhance UNet's performance in the presence of limited labels per stain, we employ UNet in a self-supervised learning setting. The architectural and training details of UNet model is provided in Appendix A.1.1. Additionally, to integrate the benefits of the knowledge gained from self-supervised pre-training, the encoder component of the UNet architecture is initialised with the weights learned during pre-training (i.e. with SimCLR, BYOL, or HR-CS-CO).

Following Ciga et al. [106], multiple splits of the overall dataset were created, as shown in Figure 4.1. Each split comprises different percentages of labelled data (1%, 5%, 10%, and 100%) taken from the training patients of each stain, as presented in Table 4.1. Additionally, seven times more tissue (i.e. non-glomeruli) patches were included to account for the variability observed in non-glomeruli tissue. In order to remove the slide background (non-tissue), each image underwent thresholding based on its mean value, followed by the removal of small objects and closing holes. Similar

to Stacke et al. [105], patches extracted from the validation patients of each stain are employed to select the best model based on the validation loss. The performance of the trained models is evaluated by segmenting the full WSIs of the test patients from each stain. The number of glomeruli present in these validation and test stainings are: PAS - 588 (valid.), 1092 (test); Jones H&E - 590 (valid.), 1043 (test); Sirius Red - 576 (valid.), 1049 (test); CD34 - 595 (valid.), 1019 (test); CD68 - 521 (valid.), 1046 (test).

Multi-Stain Segmentation involves the segmentation of glomeruli regions from histopathology images of multiple target stains using the labels for only the source stain. As detailed in Chapter 2, current state-of-art multi-stain segmentation approaches are categorised into: (a) stain-specific methods, such as MDS1 [37]; and (b) stain invariant methods, such as UDAGAN [28]. The architectural and training setup for both MDS1 and UDAGAN are detailed in Section 2.2.1 and Section 2.2.2 of Chapter 2. To assess the effectiveness of the representations learned by pre-trained networks in the context of MDS1 and UDAGAN, we used the same splits of labelled data as presented in Table 4.1, but only for the PAS stain (see Table 4.1, 1st row), since these approaches only require labels for the source stain. The motivation behind selecting PAS as the source stain is also described in Section 2.1.1 of Chapter 2. The principal objective of this study is to leverage the information from the source stain (PAS) to train a model capable of accurately segmenting glomeruli across all target stains, including Jones H&E, Sirius Red, CD68, and CD34. As a result, this could potentially reduce the need for extensive labelling of the source stain and eliminating the need of additional labelling required for target stains.

4.2.1 Results

In this section, the pre-trained models are evaluated for each downstream task (UNet, MDS1, and UDAGAN) in two different settings: fixed-features and fine-tuning. In the fixed-feature setting, the pre-trained weights are frozen to assess the quality of the learned representations from self-supervised pre-trained models, using the same hyperparameters as used for baseline models. When fine-tuning, the pre-trained weights are updated. To determine the optimal hyperparameters for fine-tuning, a separate hyperparameter study was conducted using 1%, and 5% splits of labelled data and the performance was evaluated on the validation set for each task and pre-training method. Five learning rate values, logarithmically spaced between 0.0001 and 0.1, were tested. Additionally, two different settings for weight decay were examined: one with a value of 10^{-4} and one without any weight decay. The learning rate was reduced by a factor of 0.1 at the 90th percentile of training. Based on these experiments, the best hyperparameters were selected, and the fine-tuned models were re-trained for all label splits. The F_1 score is used as the evaluation metric and the results are presented on a separate unseen test set.

Fully supervised models were trained to establish baselines for different label splits, including 100% labels. It was found that the fine-tuned models consistently outperform fixed-feature models, and therefore only fine-tuned results are shown here (fixed-feature results are in B.2).

Table 4.2: A comparison of various self-supervised pre-training methods and respective baselines (randomly initialised without any pre-training) for the downstream tasks of UNet, MDS1, and UDAGAN using various splits of labelled data. For UNet, the labels have been used for all stains, while for MDS1 and UDAGAN, the labels for only source (PAS) stain are used. The evaluation is conducted on an independent, unseen test dataset using F_1 score. Each F_1 score is the average of five different training repetitions (standard deviations are in parentheses). The highest F_1 score for each stain, across different label splits, is in italics, while the overall highest F_1 score averaged across all stains is in bold.

Downstream Tasks	Label Splits	Pre-training	Test Stains					Average
			PAS	Jones H&E	CD68	Sirius Red	CD34	
UNet	1%	None (Baseline)	0.015 (0.031)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.253 (0.059)	0.054 (0.018)
		SimCLR	*0.673 (0.021)*	0.519 (0.040)	0.407 (0.015)	0.472 (0.037)	0.652 (0.018)	0.544 (0.026)
		BYOL	0.660 (0.018)	*0.635 (0.055)*	*0.625 (0.042)*	*0.561 (0.044)*	*0.686 (0.030)*	**0.633 (0.038)**
		HR-CS-CO	0.154 (0.044)	0.188 (0.067)	0.048 (0.083)	0.337 (0.082)	0.463 (0.017)	0.238 (0.058)
	5%	None (Baseline)	0.546 (0.084)	0.593 (0.080)	0.370 (0.188)	0.707 (0.055)	0.782 (0.041)	0.600 (0.090)
		SimCLR	*0.852 (0.019)*	*0.760 (0.017)*	0.599 (0.039)	0.618 (0.042)	*0.802 (0.011)*	0.726 (0.026)
		BYOL	0.768 (0.036)	0.746 (0.076)	*0.736 (0.033)*	*0.721 (0.051)*	*0.800 (0.047)*	**0.754 (0.049)**
		HR-CS-CO	0.756 (0.079)	0.628 (0.086)	0.533 (0.067)	0.406 (0.067)	0.707 (0.037)	0.606 (0.067)
	10%	None (Baseline)	0.730 (0.017)	0.792 (0.024)	0.643 (0.053)	*0.788 (0.022)*	0.827 (0.063)	0.756 (0.036)
		SimCLR	*0.867 (0.019)*	0.813 (0.012)	0.690 (0.057)	0.696 (0.060)	*0.838 (0.007)*	**0.781 (0.031)**
		BYOL	0.794 (0.047)	*0.823 (0.054)*	*0.729 (0.052)*	0.722 (0.044)	0.776 (0.057)	0.769 (0.051)
		HR-CS-CO	0.807 (0.058)	0.748 (0.098)	*0.729 (0.040)*	0.711 (0.074)	0.791 (0.026)	0.757 (0.059)
	100%	None (Baseline)	*0.894 (0.021)*	0.840 (0.029)	0.836 (0.031)	0.865 (0.019)	*0.888 (0.015)*	0.865 (0.024)
		SimCLR	0.884 (0.003)	*0.873 (0.007)*	0.840 (0.011)	*0.881 (0.007)*	0.867 (0.027)	**0.869 (0.011)**
		BYOL	0.867 (0.009)	0.842 (0.035)	0.818 (0.036)	0.847 (0.012)	0.874 (0.021)	0.850 (0.022)
		HR-CS-CO	0.843 (0.033)	0.855 (0.015)	*0.872 (0.006)*	0.842 (0.023)	0.870 (0.011)	0.856 (0.018)
MDS1	1%	None (Baseline)	—	0.030 (0.066)	0.024 (0.054)	0.039 (0.086)	0.036 (0.079)	0.032 (0.071)
		SimCLR		*0.615 (0.015)*	*0.403 (0.031)*	*0.594 (0.026)*	*0.614 (0.028)*	**0.556 (0.025)**
		BYOL		0.516 (0.041)	0.363 (0.027)	0.525 (0.047)	0.494 (0.031)	0.474 (0.037)
		HR-CS-CO		0.326 (0.025)	0.224 (0.045)	0.359 (0.050)	0.384 (0.035)	0.323 (0.039)
	5%	None (Baseline)	—	0.711 (0.032)	0.526 (0.041)	0.685 (0.031)	0.613 (0.050)	0.634 (0.038)
		SimCLR		*0.798 (0.005)*	0.534 (0.015)	0.767 (0.008)	*0.729 (0.016)*	**0.707 (0.011)**
		BYOL		0.713 (0.051)	*0.538 (0.047)*	0.733 (0.032)	0.605 (0.061)	0.647 (0.048)
		HR-CS-CO		0.760 (0.028)	0.335 (0.084)	*0.773 (0.015)*	0.607 (0.044)	0.619 (0.043)
	10%	None (Baseline)	—	0.776 (0.017)	*0.575 (0.025)*	0.778 (0.023)	0.656 (0.030)	0.696 (0.024)
		SimCLR		*0.784 (0.026)*	0.541 (0.029)	0.752 (0.040)	*0.722 (0.016)*	**0.700 (0.028)**
		BYOL		0.706 (0.063)	0.541 (0.060)	0.731 (0.084)	0.650 (0.043)	0.657 (0.062)
		HR-CS-CO		0.771 (0.037)	0.433 (0.059)	*0.804 (0.041)*	0.633 (0.033)	0.660 (0.042)
	100%	None (Baseline)	—	0.849 (0.017)	*0.683 (0.043)*	0.870 (0.009)	*0.754 (0.008)*	**0.789 (0.032)**
		SimCLR		0.826 (0.033)	0.638 (0.056)	0.836 (0.034)	0.712 (0.030)	0.753 (0.038)
		BYOL		0.833 (0.032)	0.632 (0.042)	0.864 (0.028)	0.652 (0.066)	0.745 (0.042)
		HR-CS-CO		*0.863 (0.017)*	0.614 (0.067)	*0.878 (0.018)*	0.730 (0.040)	0.771 (0.036)
UDAGAN	1%	None (Baseline)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SimCLR	0.477 (0.015)	0.403 (0.025)	0.261 (0.053)	0.408 (0.010)	0.518 (0.016)	0.413 (0.024)
		BYOL	*0.647 (0.062)*	*0.504 (0.083)*	*0.401 (0.099)*	*0.513 (0.088)*	*0.598 (0.064)*	**0.533 (0.079)**
	5%	None (Baseline)	0.669 (0.038)	0.498 (0.056)	0.352 (0.056)	0.618 (0.072)	0.692 (0.024)	0.566 (0.049)
		SimCLR	0.719 (0.018)	0.616 (0.020)	0.524 (0.014)	0.632 (0.015)	0.716 (0.015)	0.641 (0.016)
		BYOL	*0.815 (0.027)*	*0.730 (0.071)*	*0.603 (0.028)*	*0.732 (0.028)*	*0.726 (0.055)*	**0.721 (0.042)**
	10%	None (Baseline)	0.816 (0.031)	0.687 (0.014)	0.614 (0.019)	0.750 (0.069)	0.770 (0.022)	0.727 (0.031)
		SimCLR	0.781 (0.013)	0.712 (0.013)	0.606 (0.015)	0.706 (0.026)	0.768 (0.012)	0.715 (0.016)
		BYOL	*0.834 (0.035)*	*0.767 (0.051)*	*0.654 (0.040)*	*0.742 (0.090)*	*0.781 (0.037)*	**0.755 (0.051)**
	100%	None (Baseline)	*0.901 (0.011)*	0.856 (0.036)	0.705 (0.031)	0.873 (0.025)	0.799 (0.035)	0.827 (0.027)
		SimCLR	0.892 (0.008)	*0.866 (0.018)*	*0.777 (0.013)*	*0.888 (0.015)*	*0.844 (0.003)*	**0.853 (0.011)**
		BYOL	0.883 (0.019)	0.854 (0.039)	0.722 (0.051)	0.818 (0.068)	0.792 (0.036)	0.814 (0.042)

The results presented in Table 4.2 indicate that, in the majority of limited label scenarios (1%, 5%), the fine-tuned models consistently outperformed the baselines, while with moderate (10%) and fully labelled (100%) data, they result in similar or better performance across all stains.

On average, in the limited label cases, which are equivalent to 5–6 (1%) and 26–33 (5%) labelled glomeruli per stain, the fine-tuned UNet models significantly outperform the respective baseline UNet models (see last column). This outperformance is not uniform over all stains however, notably Sirius Red and CD34 with 5% labels do benefit from pre-training but not as considerably as the other stains. For some stains, it can be observed that pre-training with 100% labels can even outperform the baseline fully supervised models, however, the benefits are not evident when averaging over all stains. As our goal is to find a labelling level that minimises labelling effort while maximising performance, 5% labels offers a good balance between the two (10% giving only a small increase in performance, while 1% a considerable drop). At this level of labelling, a 11% drop in performance is observed with BYOL pre-trained UNet in comparison to the fully (100%) supervised model. This highlights that the number of labels required for training can be reduced by 95%. If SSL had not been used in this case, a 26.9% drop in performance would have been observed (5^{th} row, last column of Table 4.2).

In MDS1 multi-stain segmentation, the same pattern can be observed. Using 1% and 5% labels (but in this case only from the source, PAS, stain) results in a considerable average performance increase over the baseline models. Focusing on 5% labels, SimCLR pre-training enables MDS1 to achieve an average F_1 score of 0.707, which is only 8.2% lower than the 100% supervised MDS1 baseline (0.789), while reducing the labelling requirement by 95%. Moreover, this is only 5% lower than the best average UNet single-stain performance with pre-training, which requires labels for all stains, whereas MDS1 requires them for only the source stain.

This trend continues in the stain invariant UDAGAN model's results, where on average pre-training and fine-tuning with 1% and 5% labels (again, for only the source stain) considerably outperforms the baselines in all stains. HR-CS-CO pre-training is not evaluated as UDAGAN is a stain-invariant single-model multi-stain segmentation approach and HR-CS-CO is trained separately for each stain. In this case, we observe a 10.6% performance drop when fine-tuning with 5% labels (and pre-training with BYOL) compared to the 100% supervised baselines. If the model had been trained in a fully supervised manner with this amount of labels, a 26.1% drop would have been observed, thus fine-tuning is able to minimise the impact of the lack of labels.

A visual confirmation of these findings is shown in Fig. 4.7, in which glomeruli segmentation maps (for models trained with 5% labels) for each stain are presented.

4.2.1.1 Omitting Validation Data

As shown above, a balance between minimising labels and maximising performance is achieved using 5% labels. Nevertheless, when training the final models, the results were obtained using a fully labelled validation set. Therefore, Table 4.3 evaluates whether the validation set is necessary or whether this labelling requirement can also

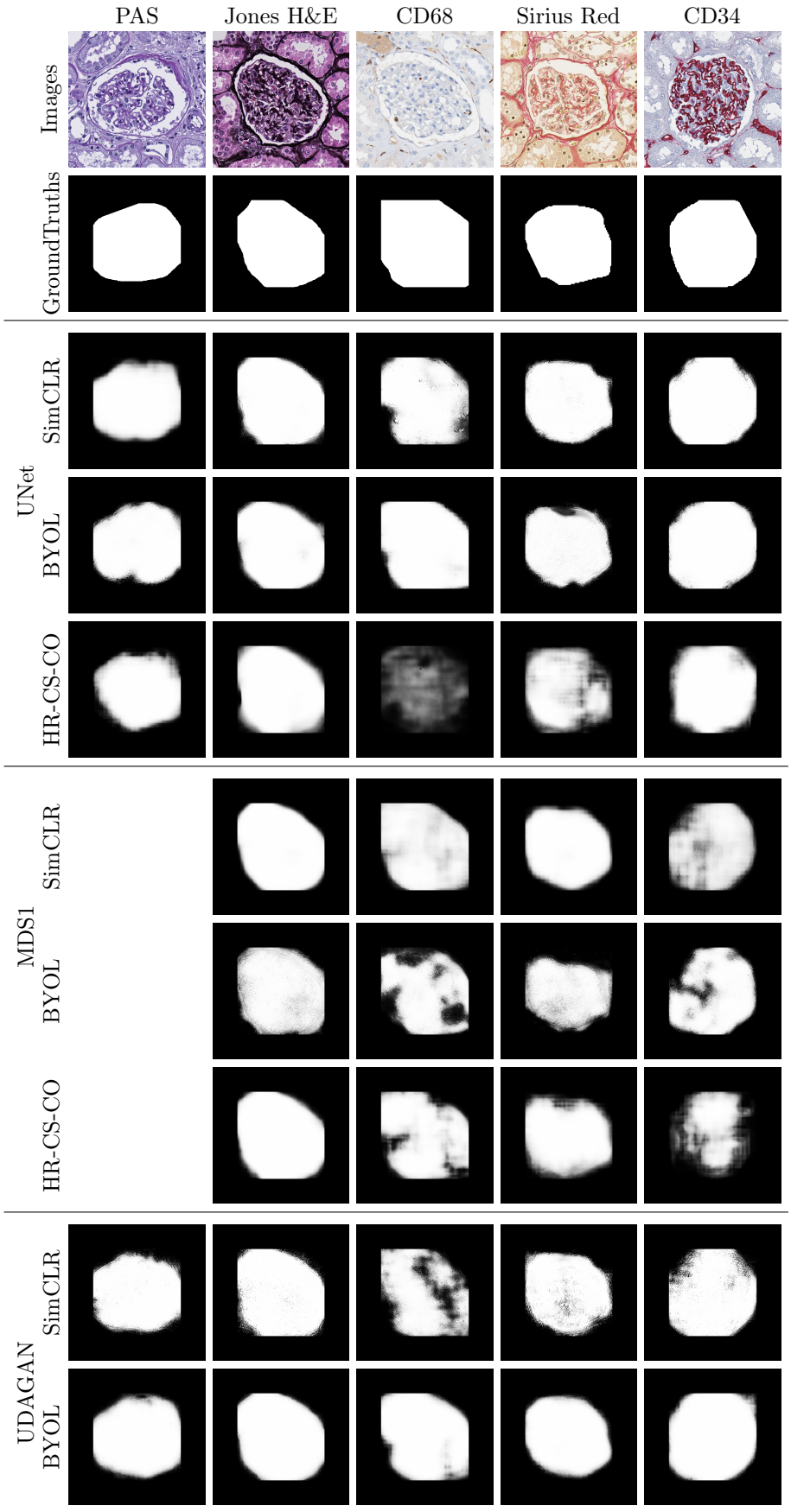


Figure 4.7: Visual comparison between predicted glomeruli segmentation maps and real ground-truths for each test stain using fine-tuned UNet, MDS1, and UDAGAN models (trained with 5% labels).

Table 4.3: Downstream task performance with 5% training labels, without a validation set. UNet, 5% labels are used for all stains, MDS1 and UDAGAN, 5% labels are used for only source, PAS, stain. The evaluation is conducted on test set. Each F_1 score is the average of five different training repetitions (standard deviations in parentheses). Highest F_1 score for each stain is in italics, overall highest F_1 score averaged across all stains is in bold.

Downstream Tasks	Pre-training	Test Stains					Average
		PAS	Jones H&E	CD68	Sirius Red	CD34	
UNet	SimCLR	*0.812 (0.019)*	0.795 (0.034)	0.575 (0.146)	0.612 (0.066)	0.810 (0.020)	0.720 (0.057)
	BYOL	0.786 (0.020)	*0.839 (0.025)*	*0.771 (0.027)*	*0.788 (0.021)*	*0.870 (0.003)*	**0.810 (0.019)**
	HR-CS-CO	0.777 (0.032)	0.695 (0.092)	0.428 (0.086)	0.425 (0.094)	0.700 (0.060)	0.605 (0.072)
MDS1	SimCLR	—	0.787 (0.016)	0.608 (0.015)	0.770 (0.021)	*0.704 (0.022)*	0.717 (0.018)
	BYOL		*0.813 (0.037)*	*0.646 (0.038)*	*0.823 (0.037)*	0.695 (0.038)	**0.744 (0.037)**
	HR-CS-CO		0.776 (0.013)	0.251 (0.051)	0.812 (0.007)	0.599 (0.026)	0.609 (0.024)
UDAGAN	SimCLR	0.402 (0.193)	0.389 (0.078)	0.000 (0.000)	0.072 (0.120)	0.359 (0.260)	0.244 (0.130)
	BYOL	*0.850 (0.008)*	*0.822 (0.021)*	*0.650 (0.029)*	*0.815 (0.026)*	*0.771 (0.011)*	**0.765 (0.022)**

be reduced. It is shown that in many cases, the performance without a validation set outperforms that obtained when using a labelled validation set. This is explained by the fact that in the dataset used, there is a lower domain shift (measured by following [52]) between the train and test set distributions, which is 0.0655 (averaged across all stains), compared to the train and validation set distributions, 0.1857. This allows the models trained without validation data to outperform (on the test data) those selected using the validation loss. Although, this behaviour is specific to datasets with the above-mentioned characteristic, it only affects the difference in performance between the two experimental settings and not the findings themselves. Let us imagine that there were a lower domain shift between the validation and training sets, in this case removing the validation set would only eliminate the increase in performance observed here. It therefore does not invalidate the findings presented herein, that the validation set can be removed to further minimise labelling requirements.

With SimCLR and UDAGAN, however, a considerable drop in performance is observed. This is likely because of overfitting in the absence of validation data. The model is trained in two stages: (1) pre-training using SimCLR on original image patches; (2) translation (using CycleGAN models) from PAS to all other stains during fine-tuning. During the second stage, imperceptible noise caused by the CycleGAN transfer [52] is introduced into the training patches. This causes a domain shift between the training data and test images, reducing performance. This is exacerbated by the absence of a validation set, which would normally prevent overfitting to this 'noisy' training data. In contrast, BYOL is not affected because it uses batch-normalisation, which helps to stabilise the training process and prevent overfitting the noisy inputs. This can be visualised in Fig. 4.8, where there is a noticeable lack of class boundary between test glomeruli and negative patches when training SimCLR-UDAGAN without a validation set, see Fig. 4.8(a). Such a boundary exists in the BYOL-UDAGAN representation without validation data,

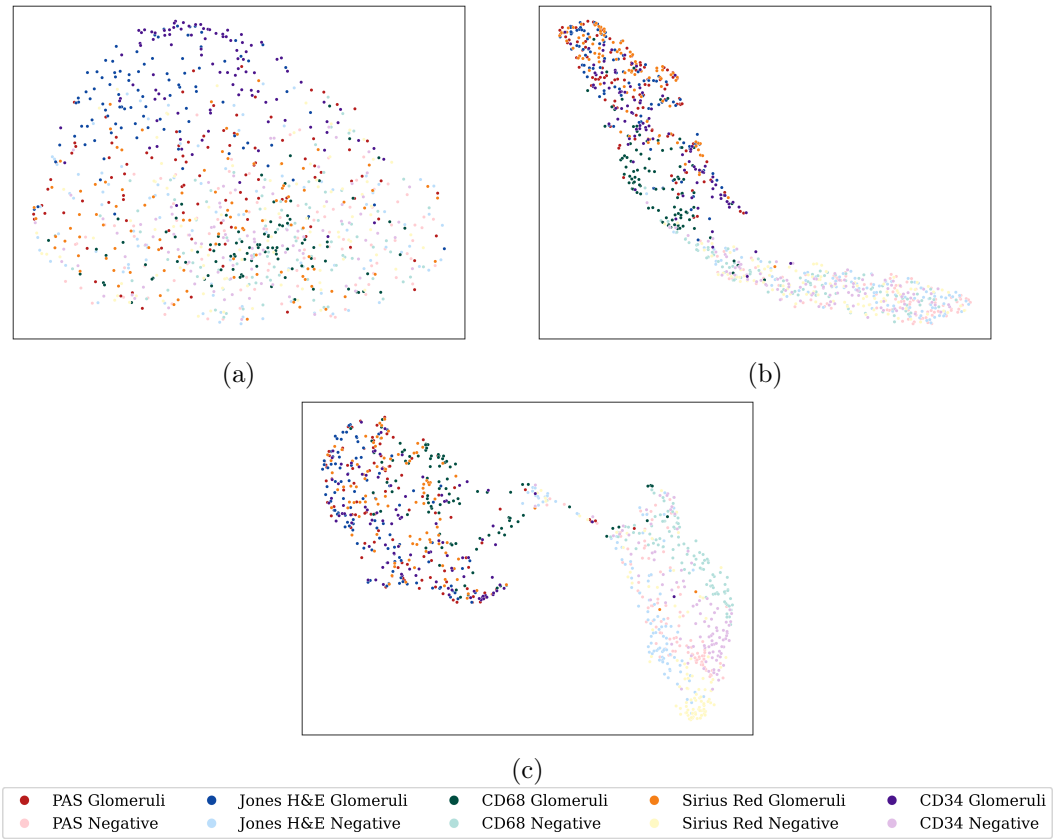


Figure 4.8: Two-dimensional UMAP embeddings of the representations learned by: (a) SimCLR and (b) BYOL based UDAGAN models, trained without a validation set, and (c) SimCLR UDAGAN with a 5% labeled validation set. Models randomly chosen, representations sampled from the penultimate convolutional layer, 100 patches per stain per class from the unseen test set. Each point is a patch from the respective class and staining.

see Fig. 4.8(b), and a SimCLR-UDAGAN trained with 5% validation labels, see Fig. 4.8(c) (for comparison, this model achieves an average F_1 score of 0.686, vs. 0.244 without the validation set).

4.3 Discussion

The previous section showed the effectiveness of SSL in combating a lack of labelled segmentation data in histopathology, approaching fully-supervised performance (e.g. with BYOL pre-training) in both single-stain UNet and multi-stain UDAGAN models (e.g. with ∼30 labels per training stain).

We can observe however that not all self-supervised learning approaches are equal. When fewer labels are available (1% and 5%), general computer vision (CV) approaches such as SimCLR and BYOL perform best. Even though HR-CS-CO is specifically designed for histopathology, it only becomes competitive and/or outper-

forms the CV approaches when provided with moderate (10%) to larger amounts of labelled data. It is particularly successful when applied to the CD68 stain, outperforming even the baseline models. CD68 is an immunohistochemical stain in which haematoxlyn highlights the main structural component and specific immune cells are highlighted in brown. It is therefore particularly suited to an approach such as HR-CS-CO. There are many other similar immunohistochemical stainings, and more complicated double stainings (e.g. CD3-CD68, CD3-CD163, CD3-CD206, etc [164]) that should be suitable for such an approach (including the H&E stain CS-CO was originally developed for). In some of the other stains used in this study (e.g. Sirius Red), it appears that the superposition of staining components (and weak haematoxyln staining) prevents the haematoxlyn channel from being efficiently extracted, limiting the effectiveness of HR-CS-CO (the difficulty of extracting this component from Sirius Red has been previously noted in the literature [27]).

The final intended tasks of the pre-trained model often dictate the type of SSL that should be used. In UDAGAN, BYOL consistently outperforms SimCLR, especially with highly limited labels, likely due to its robustness to noise during fine-tuning. Unlike SimCLR, which relies on negative pairs, BYOL uses only positive samples and keeps a moving average for regularisation, making it less sensitive to noisy (i.e. translated) data during fine-tuning [165]. Nevertheless, SimCLR outperforms BYOL in MDS1, despite both being applied to 'noisy' translated stains, notably CD68 (MDS1's UNet is trained on the source stain's real, noise-free, data making it sensitive to any noise in the target→PAS data during testing [52]). It is known that this is particularly evident in immunohistochemical stainings such as CD68 and CD34 [28, 52], which is confirmed in this study where the noise degrades the performance of all pre-training methods equally, including downstream segmentation.

The role of validation data was shown to strongly impact the success of fine-tuning pre-trained models. Surprisingly, omitting a validation set greatly improved the success of fine-tuning, reaching performance levels approaching those of fully supervised models. This means that almost state-of-the-art performance can be achieved while reducing labelling requirements by 95%.

Finally, the benefits of self-supervised pre-training are not just restricted to limited label situations. This study has shown that the performance of fully-supervised stain-invariant models such as UDAGAN can be improved—pre-training the UDAGAN model before fully-supervised training lead to a 2.6% increase in F_1 score. This offers a new SOTA performance in stain-invariant glomeruli segmentation without any architectural nor labelling changes.

Moving away from renal histopathology, these results are consistent with other histopathology studies found in the literature and extend upon existing efforts to reduce the need for extensive manual annotations. Particularly, Prakash et al. [97] showed that for nuclei segmentation in the Broad Bioimage Benchmark Collection dataset, a self-supervised fine-tuned UNet using only 5% labels (32 images) demonstrated only 3% reduction in IoU score compared to a full supervised UNet. Similarly, Punn et al. [166] reported that a self-supervised fine-tuned UNet using 20% labels (134 images) for nuclei segmentation on the Kaggle Datascience Bowl Challenge 2018 dataset only lost 5.1% in F_1 score compared to a fully supervised UNet.

Combined with the results presented herein, these demonstrate minimal performance degradation despite significant reductions in label requirements. The findings presented herein, however, go further. Not only do they show the benefit of integrating pre-training into fully-supervised approaches, but also into multi-stain segmentation strategies and removing the need for labelled validation datasets. This more than reduces the labelling requirement to the source stain (a reduction of at least n times, where n is the number of stains to be segmented).

This discussion has already outlined the limitations of HR-CS-CO and so it remains to address SSL limitations in general. Foremost, there is a risk of introducing false negatives when training SimCLR on datasets with few classes because mini-batches are likely to contain several samples from one class. This can lead to a model that fails to distinguish between semantically "similar" and "dissimilar" images, reducing downstream performance. BYOL, however, overcomes this limitation by not using negative pairs. Moreover, contrastive SSL in general relies on augmentation to create "similar" pairs. As outlined by Garcea et al. [167], medical imaging is sensitive to augmentation since it contains subtle, easily distorted features.

4.4 Conclusions

This chapter has shown how to significantly reduce the need for labelled data (>95%) in histopathology image segmentation. To achieve this, self-supervised pre-training techniques— SimCLR, BYOL, and a novel histopathological SSL approach, HR-CS-CO—were used to learn general features from unlabelled data. These features were then fine-tuned for single stain and multi-stain segmentation tasks using UNet, MDS1, and UDAGAN models, making them robust to training scenarios with limited labels.

These approaches demonstrated consistently superior performance compared to their respective baselines, and were able to approach the performance of fully supervised models. These findings underscore the potential and significance of incorporating these advanced learning techniques in histopathology. The results also demonstrated that self-supervised learning combined with fine-tuning is most effective without a validation set, further reducing the labelling requirement. However, some methods, such as SimCLR, are more susceptible to domain shifts and may benefit from some labelled validation data to ensure generalisation.

Furthermore, this study advanced the recent trend in histopathology towards creating multi-stain segmentation models by demonstrating that it is possible to train a stain-invariant segmentation model with as few as ∼30 labelled positive patches from one stain. This model closely matches the performance of a fully supervised UNet trained with ∼3000 positive patches.

Chapter 5

Conclusions and Perspectives

Deep learning algorithms have shown impressive achievements across various digital histopathology tasks such as cancer detection, disease classification, and transplant assessment. However, a significant challenge arises in the training of these algorithms, as many state-of-the-art deep learning algorithms are data hungry, often demanding extensive amounts of labelled data. This problem is further compounded by variations in tissue preparation and staining protocols. Consequently, existing datasets with annotated labels often exhibit limited reusability, even for similar tasks. To address these variations, researchers have proposed various CycleGAN based stain transfer methods. These methods facilitate the development of cutting-edge multi-stain segmentation and/or classification models that can be used across various stains without requiring the labels for each stain, as detailed in Chapter 2. Instead, they tend to only use the labels from a single (source) stain.

While CycleGAN based stain transfer has emerged as a standard approach for addressing stain variation, its inherent limitations are often overlooked. It struggles to learn appropriate mappings when translating between stain pairs with significant differences, such as between histochemical (i.e. PAS) and immunohistochemical (i.e. CD68/CD34) stains. This limits the use of CycleGAN in downstream tasks, since it must include additional information (in the form of imperceptible noise) in order to complete the transfer, which leads to domain shift in the translated stains and this can affect the final downstream task predictions. The application that this thesis focused on to investigate these limitations was the segmentation of glomeruli structures in renal pathology across multiple stains, all of which have been labelled by trained experts (pathologists) for evaluation purposes.

A crucial step towards handling this domain shift is the ability to detect and measure it. Therefore, Chapter 3 of this thesis explored the state-of-art in unsupervised deep learning techniques and proposed an approach to detect and quantify this domain shift. While this focused on detecting domain shift introduced by CycleGAN based stain transfer models, the proposed solution is general and can detect other forms of domain shift. Moreover, it was shown that the proposed measures have a strong correlation with the segmentation performance in the translated stains. Consequently, these findings offer a mechanism to estimate the average performance of deep neural networks (trained for a source domain) when applied to the same task in unseen and unlabelled data. This measure was used to demonstrate that relying on visual assessment, which is widely adopted in practice, is ill-advised and should not be the sole criterion for evaluating the quality of CycleGAN based translated images. Finally, the measure was integrated into the CycleGAN model in order to prevent the noise from being introduced in the translated images and this was shown to enhance multi-stain segmentation performance.

Despite the proven effectiveness of existing stain transfer based multi-stain segmentation methods in eliminating the need of labels in the target domain(s), it should be acknowledged that these methods still rely on a large number of labelled data from the source domain. This can be challenging in various medical disciplines, in which the application of existing multi-stain segmentation approaches may prove impractical, creating a barrier to their widespread adoption. Therefore, Chapter 4 of this thesis shows that the amount of labelled source data for multi-stain histopathological segmentation can be reduced by a significant amount with little or no loss of performance. This is achieved by integrating self-supervised representation learning (i.e. SimCLR, BYOL, and our proposed HR-CS-CO) using large amount of unlabelled data.

Moreover Chapter 4 several approaches to use these self-supervised learning methods in different staining protocols. Methods such as CS-CO are stain specific, and therefore limited to only a single type of stain (i.e. H&E). The proposed modifications, which we referred to as HR-CS-CO, not only improved its performance over baseline models but also expanded its applicability across multiple other stains.

Bringing all the above together, it was demonstrated that the number of required labels for histopathological segmentation can be reduced by upto 95%, allowing multi-stain segmentation with as little as approximately $\sim$30 labelled glomeruli in a single (source) stain.

In conclusion, the proposed contributions demonstrate significant efficacy in digital histopathological tasks, particularly those hindered from a scarcity of labelled data. By enabling models to adapt and generalise effectively in scenarios with limited labels, these contributions stand to relieve medical experts from time-intensive labelling efforts and allow them to focus more on the patient care. The proposed contributions also have the potential to be applied across various other related applications, particularly in emerging applications, where labels are completely absent in all domains. By adopting this approach, it becomes feasible to significantly save time and cost by acquiring few labels for only one specific domain, rather than for all domains. This capability not only enhances the potential integration into clinical workflows but also facilitates the generalisation of deep learning models across diverse medically related tasks.

Collectively, these advances contribute to the broader application of deep learning in medical settings, promoting more efficient, scalable, and cost-effective diagnostic solutions.

5.1 Perspectives

The work presented in this thesis opens up several possible research directions, some of which are concerned with directly improving the proposed methods, while others involve applying the developed methodology to different but related fields.

The thesis demonstrated that it is possible to segment glomeruli across multiple stains using only a few labels from a single stain. To further substantiate these findings, additional exploration should be made to confirm that the same approaches can be used to classify, detect or segment other diagnostically relevant structures

in digital histopathology (e.g. tubules, etc), irrespective of the staining modality. Successful application requires that the target structures maintain consistent morphology across different stains, even if the textural and colour information varies. This approach could also extend to other medical imaging modalities, such as MRI or CT scans, where anatomical structures of interest retain their general appearance despite different imaging techniques, and to broader computer vision problems where objects keep their general appearance despite changes in context, texture, or colour.

The proposed advances in the CycleGAN architecture enhanced its robustness and generalisation capabilities when used for stain transfer based multi-stain segmentation models. This lays the ground for exploring several other state-of-the-art deep learning methods that work on similar principles. Recently, diffusion based deep learning models have been introduced for generating high quality images [168] and are gaining significant attention in digital histopathology tasks, particularly for stain transfer [169, 170]. As a result, one of our future research directions is to investigate diffusion models to further enhance the effectiveness of multi-stain segmentation methods.

Despite the strides made by the self-supervised learning (SSL) methods in Chapter 4, our work acknowledges certain challenges. Notably, when training SimCLR on datasets with limited class diversity, there is an increased risk of generating false negatives. Additionally, the augmentations used in contrastive based SSL methods are specifically designed for natural images and medical images are highly sensitive to these augmentations [123, 167]. To overcome these challenges, our future work will focus on transformer based masked image modelling (MIM) [171, 172]. This approach represents a more robust approach to self-supervised learning because MIM aims to learn representations by generating missing parts of an image, thereby forcing the model to learn relationships between different elements within an image. Notably, these methods do not need any additional augmentation steps and have demonstrated great scalability and robustness.

Appendices

Appendix A

Stain Transfer

A.1 Network Architectures and Training Details

A.1.1 UNet

The UNet [29], presented in Figure A.1, is a highly effective CNN architecture that has demonstrated remarkable efficacy in segmenting biomedical images [82], specifically for glomeruli segmentation [83]. It adopts an encoder-decoder structure, forming a U-shaped network, which effectively handles both local and global information. The encoder path, also known as the contracting path, comprises repetitive blocks, each encompassing two consecutive 3×3 convolutions followed by ReLU activation and a max-pooling layer. Conversely, the decoder path, or expanding path, gradually upsamples the feature maps using 2×2 transposed convolution layers. Subsequently, the corresponding feature map from the contracting path is cropped and concatenated with the up-sampled feature map, followed by two consecutive 3×3 convolutions and a ReLU activation. Finally, a 1×1 convolution is applied to reduce the feature map to the desired number of channels (classes), generating the segmentation map. The cropping step is necessary since pixels at the edges contains less contextual information and therefore should be discarded.

The UNet is trained for 250 epochs (following [28]) using a batch size of 8 and a learning rate of 0.0001. To ensure consistency, all patches are standardised to $[0, 1]$ and normalised by the mean and standard deviation of the training set. Five different repetitions of the UNet model were trained for each stain and for each split of labelled data. As cropping is performed to discard the less contextual information, the predicted segmentation output has a size of 324×324 pixels, see Figure A.1 for more details. The same augmentation as suggested by Lampert et al. [27], are applied with an independent probability of 0.5 ("on the fly"). Specifically, these augmentations with their respective parameters are listed as follows:

elastic deformation: with parameters ($\sigma = 10$, $\alpha = 100$)

affine: random rotation sampled from [0°, 180°], random shift sampled from [-205, 205] pixels, random magnification sampled from [0.8, 1.2], and horizontal/vertical flip

noise: additive Gaussian noise[1] with $\sigma \in [0, 2.55]$

blur: Gaussian filter[1] with $\sigma \in [0, 1]$

brightness: enhance[2] with a factor sampled from $[0.9, 1.1]$;

contrast: enhance[2] with a factor sampled from [0.9, 1.1];

[1] Using the appropriate scikit-image functions.

[2] Using the appropriate PIL functions.

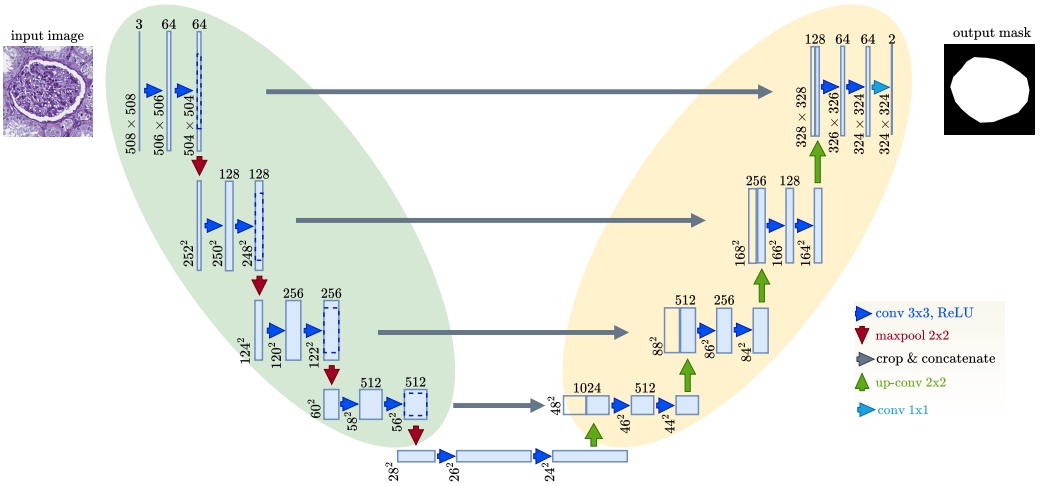


Figure A.1: Overview of the UNet architecture.

stain variation: by colour deconvolution [162], α sampled from [-0.25, 0.25] and β from [-0.05, 0.05].

A.1.2 CycleGAN

To train CycleGAN, the network architecture and loss weights ($w_{cyc} = 10, w_{id} = 5$) are taken from the original paper [47], since they produced realistic output, as can be seen in Figure 2.4 of Chapter 2. To deal with large patch sizes (i.e. above 256×256 pixels), a translation network with 9 ResNet blocks is employed, as suggested by the authors. The model is trained for 50 epochs, with a learning rate of 0.0002 using the Adam optimiser, and a batch size of 1. Starting from the 25^{th} epoch, the learning rate linearly decayed to 0, and the weights for cycle-consistency (w_{cyc}) and identity (w_{id}) are halved. In all experiments, the translation model from the final (50^{th}) epoch is used. Additionally, to reduce model oscillation, the strategy proposed by [173] is adopted that updates the discriminator using the 50 previously generated samples.

A.1.3 MDS1

MDS1 requires both a UNet and CycleGAN network to be trained. The UNet and CycleGAN models are trained as described in Appendix A.1.1 and A.1.2 respectively. Once the models are finished training, the trained UNet model is then applied to 'fake' source stains, referred herein as Target→Source, to generate the segmentation mask for Target stain.

To account for random variations, the CycleGAN network was trained three times for each target stain, and five repetitions of the UNet were trained for each CycleGAN (resulting in 15 repetitions).

A.1.4 UDAGAN

The training process of UDAGAN follows a similar approach to MDS1, incorporating both UNet and CycleGAN networks. Specifically, the first step involves training of a separate CycleGAN network for each target stain, enabling the translation from Source→Target. Subsequently, a training patch is translated into randomly selected stain (with a probability of $\frac{N-1}{N}$, where N is the number of stains) using the pre-trained CycleGAN network(s). Thus, all available stains (including the source stain) are presented to the network with equal probability, $\frac{1}{N}$, forcing the network to learn stain invariant features. The same augmentations as for the UNet are used when training UDAGAN.

A.1.5 PixelCNN

The PixelCNN [126] architecture is used to model the underlying distribution of PAS stain. The architectural configurations are formalised as: the model employs 3 Resnet [132] blocks consisting of 5 residual layers in the encoding phase, with 2×2 downsampling between the ResNet blocks. In the decoding phase, the same architecture is employed, 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, we extracted 1280000 train, validation, and test patches from the corresponding patients. 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 with the lowest bits-per-dimension score [174] on the validation set. We use 128000 patches as the validation set, extracted randomly from the validation patients. We employed the original publicly available implementation[3].

A.1.6 CycleGAN with Gaussian Noise

Bashkirova et al. [49] stated that it is impossible to separate the pure structured noise from the translated images. Consequently, classical adversarial defence methods [175–178] can not be used. Nevertheless, Inspired by [179], Bashkirova et al. [49] introduced a novel mechanism to enhance the CycleGAN model's robustness by adding random noise to the translated images before the reconstruction of original image, as illustrated in Figure A.2. This additional noise disrupts the embedded noisy signal in the translated image, leading to a significant increase in the reconstruction error. As a result, the generator should start to learn the accurate image reconstruction without relying on the embedded noise. This modification is integrated into the cycle-consistency loss (defined in Equation (2.2) of Chapter 2), which

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

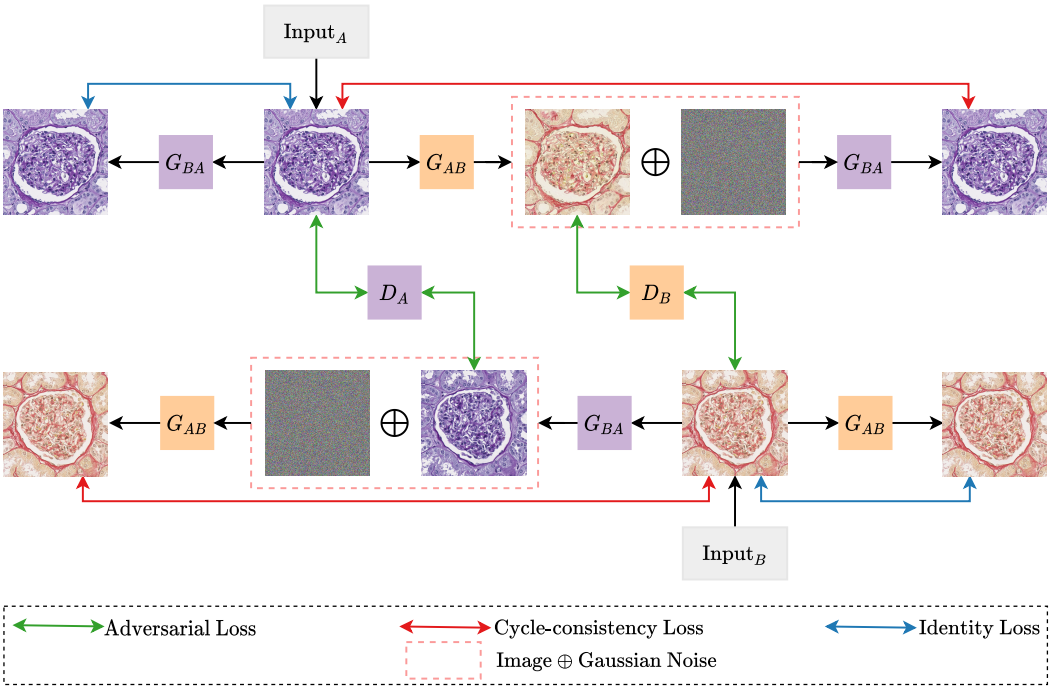


Figure A.2: Overview of CycleGAN with Gaussian Noise for stain transfer

is modified such that

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) = \mathbb{E}_{s\sim A}[\|G_{BA}(G_{AB}(s) + \mathcal{N}(0,\sigma)) - s\|_1] \\ + \mathbb{E}_{t\sim B}[\|G_{AB}(G_{BA}(t) + \mathcal{N}(0,\sigma))) - t\|_1],\end{aligned} \tag{A.1}$$

where $\mathcal{N}(0,\sigma)$ denotes zero-mean Gaussian noise with standard deviation σ, where σ lies within the range of $[0,1]$.

A.1.7 CycleGAN with Extra-Channels

Inspired by the idea of [48], Bouteldja et al. [77] introduced another approach that incorporates an additional-feature channel, similar to image size (i.e. 508×508 pixels), into both the input and output of each generator. The input is zero-padded with three extra channels (RGB). While translating Source→Target (and vice versa), the output of the generator (G_{AB}) now comprises the usual three-channel translation, which is then propagated through the respective discriminator, alongside three additional channels. The hope is that these additional channels will be used by the generator to store the artificial meta-information (i.e. hidden noise) that are crucial for its subsequent reconstruction. Thus, this approach presents an opportunity for the generators to implicitly separate the imperceptible noise from the translations, leading to the noise-free translations.

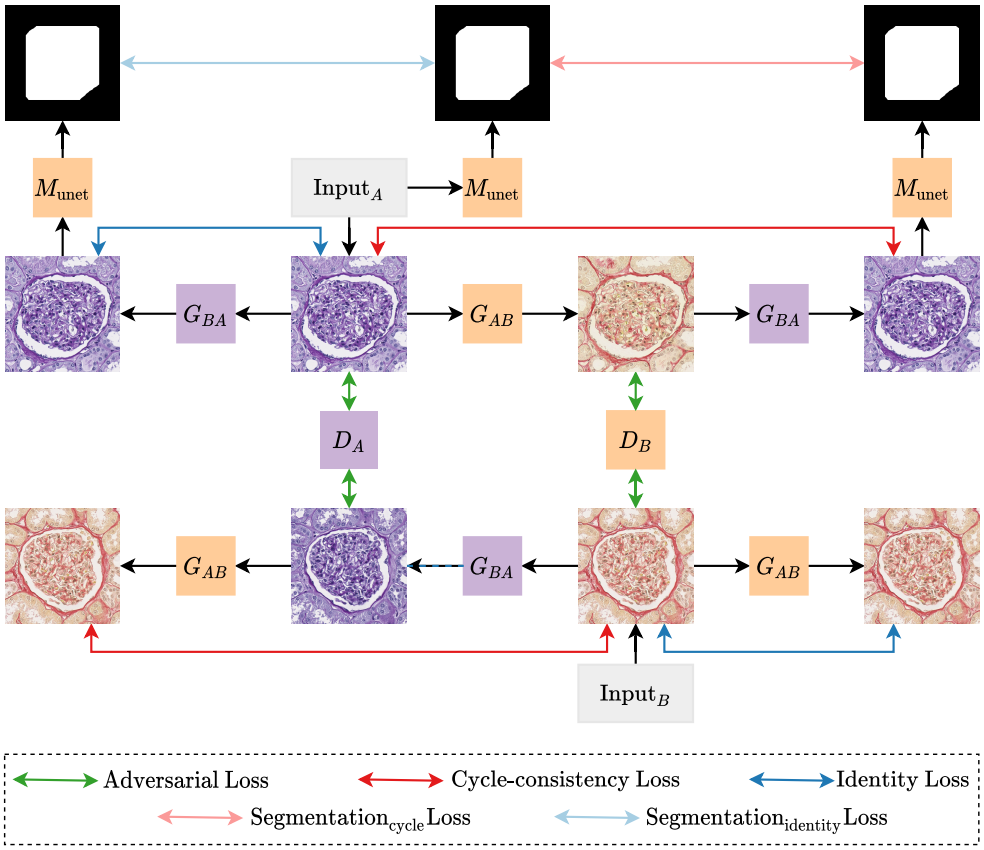


Figure A.3: Overview of CycleGAN with Self-supervision for stain transfer.

A.1.8 CycleGAN with Self-Supervision

Bouteldja et al. [77] introduced a novel approach tailored to improve the CycleGAN architecture for specific applications, such as stain-transfer based segmentation in histopathology, by integrating a pre-trained segmentation network. This approach aims to optimise the hidden information (imperceptible noise) within the translated images through semantic guidance in a self-supervised manner. The pre-trained segmentation network is initially trained on the source domain, where labels are readily available. Throughout training, as depicted in Figure A.3, only images from the source domain, alongside their reconstructions and identity mappings, are fed through the pre-trained segmentation network. Since the segmentation network is pre-trained on real samples from the source domain, the segmentation output of real images from source domain serves as ground-truth. These ground-truths are then used as targets for self-supervision, wherein discrepancies between the network's predictions (for reconstructed and identity images), and the ground-truth are penalised

using a segmentation loss ($\mathcal{L}_{\text{seg}}$), which is defined as:

$$\begin{aligned}\mathcal{L}_{\text{seg}} &= \mathcal{L}_{\text{seg,cyc}} + \mathcal{L}_{\text{seg,id}} \\ &= \mathbb{E}_{s\sim A}[\|M_{\text{unet}}(G_{BA}(G_{AB}(s)) - M_{\text{unet}}(s))\|_1 \\ &+ \mathbb{E}_{s\sim A}\|M_{\text{unet}}(G_{BA}(s)) - M_{\text{unet}}(s))\|_1] \end{aligned} \tag{A.2}$$

where M_{unet} is a pre-trained source segmentation network. Consequently, by optimising $\mathcal{L}_{\text{seg}}$, the overall CycleGAN framework learns to better translate the features to give consistent segmentation outputs. The overall loss function of CycleGAN (presented in Equation (2.4)) can now be formulated as:

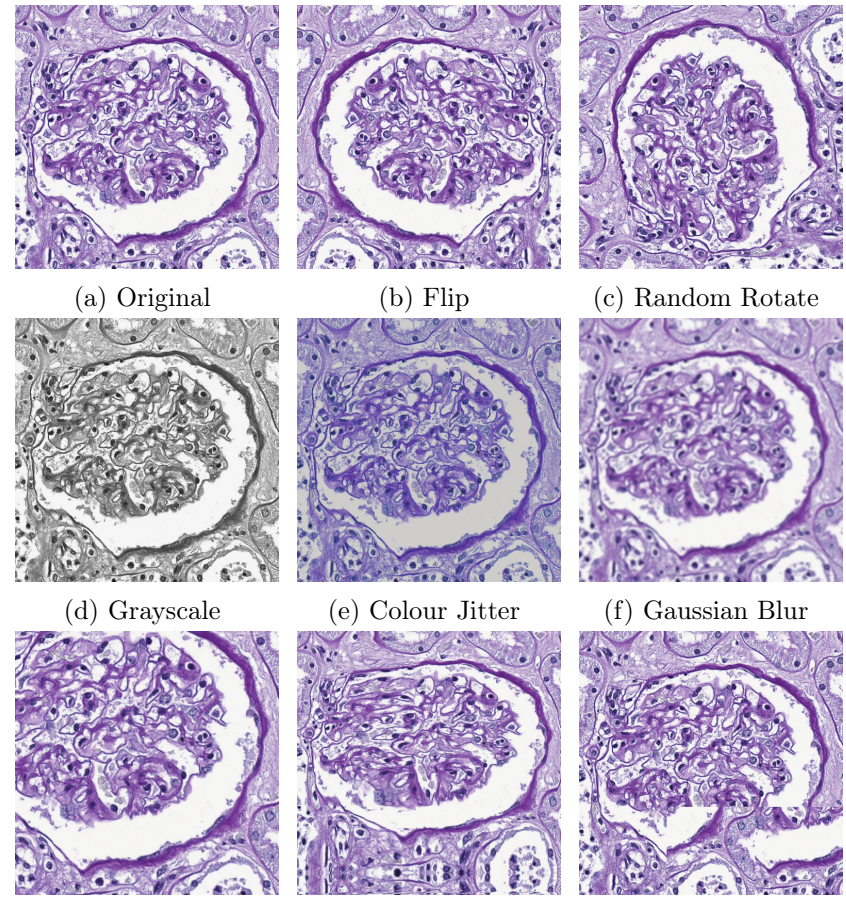
$$\begin{aligned}\mathcal{L}_{\text{CycleGAN}}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{\text{adv}}(G_{AB}, D_B, G_{BA}, D_A) \\ + w_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) \\ + w_{\text{id}}\mathcal{L}_{\text{id}}(G_{AB}, G_{BA}) \\ + w_{\text{seg}}\mathcal{L}_{\text{seg}}(G_{AB}, G_{BA}).\end{aligned} \tag{A.3}$$

Appendix B

Self-supervised Learning

B.1 Augmentations

Albumentations library [180] is used to apply different augmentations for the pre-training of various self-supervised methods. Given an input image of size 508×508, a visual example for each augmentation is provided in Figure B.1. Table B.1 provides the implementation details and specified augmentation parameters in our study.



(a) Original (b) Flip (c) Random Rotate

(d) Grayscale (e) Colour Jitter (f) Gaussian Blur

(g) Random Crop Resize (h) Grid Distort (i) Grid Shuffle

Figure B.1: Visualising different methods to augment the input data for self-supervised learning.

Table B.1: List of applied image augmentations including their specific parameters and the probability of application.

Augmentation	Parameters	Probability
Flip	-	0.5
Grayscale	-	0.5
Random Rotate	-	0.5
Grid Shuffle	grid : (3, 3)	0.6
Gaussian Blur	blur_limit : (3, 7) sigma_limit : (0.1, 2.0)	0.5
Random Crop Resize	height : 508 width : 508 scale : (0.2, 1.0)	0.8
Grid Distort	num_steps : 9 distort_limit : 0.3 border_mode : 2.0	0.6
Colour Jitter	brightness : 0.8 contrast : 0.8 saturation : 0.8 hue : 0.2	0.8

B.2 Results with Fixed Representations

To evaluate whether self-supervised learning methods are able to learn meaningful representations and generalise to downstream tasks, it is important to use a fixed-feature setting. Therefore, in this setting, the representations learned are used as feature vectors in the downstream tasks of UNet, MDS1, and UDAGAN and the results are provided in Table B.2, evaluated across different splits of labelled data. Since these self-supervised pre-training methods are not explicitly designed for learning representations well-suited for the tasks of UNet, MDS1, and UDAGAN. Consequently, fixed-feature settings exhibit a significant drop in performance when compared to fine-tuned models (as presented in Table 4.2), and this is particularly noticeable in the case of HR-CS-CO. This highlights the need for a more effective stain separation methods beyond the classical, matrix decorrelation based approach employed in our study. This is why, during fine-tuning, HR-CS-CO's representation is able to better adapt to the specific characteristics of the downstream task,

Table B.2: Performance evaluation of various self-supervised learning based UNet methods in a fixed-feature scenario for glomeruli segmentation. The performance is evaluated in terms of segmentation (F_1) score, averaged over five different training repetitions, with the standard deviations presented in parentheses.

Downstream Tasks	Label Splits	Pre-training	Test Stains					Average
			PAS	Jones H&E	CD68	Sirius Red	CD34	
UNet	1%	SimCLR	0.575 (0.043)	0.472 (0.022)	0.348 (0.073)	0.376 (0.070)	0.700 (0.032)	0.494 (0.048)
		BYOL	0.478 (0.086)	0.556 (0.075)	0.190 (0.075)	0.471 (0.071)	0.688 (0.011)	0.477 (0.064)
		HR-CS-CO	0.191 (0.049)	0.079 (0.039)	0.030 (0.061)	0.170 (0.070)	0.312 (0.075)	0.156 (0.059)
	5%	SimCLR	0.800 (0.009)	0.724 (0.013)	0.538 (0.087)	0.526 (0.093)	0.809 (0.004)	0.679 (0.041)
		BYOL	0.734 (0.068)	0.763 (0.035)	0.435 (0.044)	0.656 (0.047)	0.745 (0.040)	0.667 (0.047)
		HR-CS-CO	0.469 (0.056)	0.546 (0.020)	0.228 (0.051)	0.252 (0.054)	0.499 (0.036)	0.399 (0.043)
	10%	SimCLR	0.850 (0.005)	0.794 (0.017)	0.698 (0.033)	0.509 (0.077)	0.820 (0.019)	0.734 (0.030)
		BYOL	0.785 (0.042)	0.765 (0.015)	0.600 (0.042)	0.731 (0.036)	0.789 (0.032)	0.734 (0.033)
		HR-CS-CO	0.563 (0.051)	0.644 (0.011)	0.279 (0.016)	0.461 (0.083)	0.540 (0.034)	0.498 (0.039)
	100%	SimCLR	0.881 (0.006)	0.848 (0.016)	0.794 (0.011)	0.786 (0.024)	0.876 (0.009)	0.837 (0.013)
		BYOL	0.878 (0.007)	0.849 (0.011)	0.781 (0.012)	0.800 (0.018)	0.867 (0.011)	0.835 (0.012)
		HR-CS-CO	0.578 (0.077)	0.711 (0.009)	0.619 (0.032)	0.683 (0.028)	0.675 (0.013)	0.653 (0.032)
MDS1	1%	SimCLR	—	0.540 (0.045)	0.348 (0.013)	0.526 (0.034)	0.513 (0.030)	0.482 (0.031)
		BYOL		0.471 (0.034)	0.329 (0.029)	0.472 (0.042)	0.466 (0.026)	0.435 (0.033)
		HR-CS-CO		0.165 (0.034)	0.085 (0.043)	0.207 (0.054)	0.186 (0.051)	0.161 (0.046)
	5%	SimCLR	—	0.746 (0.013)	0.529 (0.012)	0.732 (0.012)	0.645 (0.021)	0.663 (0.014)
		BYOL		0.702 (0.015)	0.438 (0.020)	0.657 (0.018)	0.638 (0.022)	0.609 (0.019)
		HR-CS-CO		0.410 (0.014)	0.145 (0.019)	0.352 (0.014)	0.362 (0.015)	0.317 (0.015)
	10%	SimCLR	—	0.806 (0.017)	0.632 (0.016)	0.731 (0.015)	0.780 (0.016)	0.709 (0.026)
		BYOL		0.745 (0.013)	0.612 (0.014)	0.715 (0.015)	0.729 (0.013)	0.622 (0.023)
		HR-CS-CO		0.435 (0.012)	0.355 (0.017)	0.482 (0.016)	0.501 (0.014)	0.365 (0.024)
	100%	SimCLR	—	0.921 (0.005)	0.844 (0.010)	0.853 (0.010)	0.896 (0.010)	0.744 (0.009)
		BYOL		0.896 (0.008)	0.831 (0.013)	0.843 (0.010)	0.880 (0.011)	0.742 (0.015)
		HR-CS-CO		0.724 (0.011)	0.610 (0.016)	0.683 (0.015)	0.705 (0.013)	0.427 (0.037)
UDAGAN	1%	SimCLR	0.529 (0.038)	0.463 (0.055)	0.315 (0.078)	0.491 (0.048)	0.589 (0.043)	0.477 (0.053)
		BYOL	0.534 (0.020)	0.427 (0.018)	0.281 (0.042)	0.473 (0.051)	0.560 (0.021)	0.455 (0.031)
	5%	SimCLR	0.752 (0.007)	0.664 (0.042)	0.524 (0.067)	0.689 (0.008)	0.753 (0.010)	0.677 (0.027)
		BYOL	0.779 (0.023)	0.683 (0.043)	0.462 (0.047)	0.694 (0.026)	0.701 (0.044)	0.664 (0.037)
	10%	SimCLR	0.775 (0.019)	0.691 (0.045)	0.608 (0.035)	0.733 (0.026)	0.768 (0.011)	0.715 (0.027)
		BYOL	0.830 (0.027)	0.743 (0.029)	0.518 (0.035)	0.728 (0.031)	0.764 (0.017)	0.717 (0.028)
	100%	SimCLR	0.835 (0.018)	0.755 (0.036)	0.637 (0.056)	0.794 (0.031)	0.772 (0.032)	0.758 (0.035)
		BYOL	0.860 (0.020)	0.819 (0.022)	0.618 (0.025)	0.810 (0.022)	0.791 (0.025)	0.780 (0.023)

and therefore compensate for the limitations caused by the loss of information resulting from the stain separation used during training. Nonetheless, it is essential to acknowledge that even though fixed-feature models experience a decline in performance, they show improved results in comparison to baseline models, especially when employing SimCLR and BYOL as pre-trained models. This improvement is particularly evident when the models are subjected to limited labelled data, such as 1% and 5%. Moreover, when provided with moderate (10%) to fully (100%) labelled data, the fixed-feature models approach the performance levels of baseline models. This highlights the effectiveness of self-supervised learning methods in the context of their capacity to learn meaningful representations.

Bibliography

[1] Council of Europe. History of Artificial Intelligence, 2023. URL https://www.coe.int/en/web/artificial-intelligence/history-of-ai. Accessed: 2023-12-13. (Cited on page 1.)

[2] Peter H. Diamandis. WHY AI IS EXPLODING NOW!, 2023. URL https://www.diamandis.com/blog/ai-exploding-now. Accessed: 2023-12-13. (Cited on page 1.)

[3] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227, 2020. (Cited on page 1.)

[4] Terrence J. Sejnowski. *The Deep Learning Revolution*. The MIT Press, 2018. ISBN 9780262038034. (Cited on page 1.)

[5] 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. (Cited on page 1.)

[6] 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. (Not cited.)

[7] 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. (Cited on pages 1 and 2.)

[8] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67, 2021. (Cited on pages 2 and 15.)

[9] 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. (Cited on page 2.)

[10] 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. (Cited on page 2.)

[11] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics Guidelines for Trustworthy Ai*. Publications Office, 2019. (Cited on page 2.)

[12] Adrian P. Brady. Artificial intelligence in radiology: An exciting future, but ethically complex. *EMJ Radiology*, pages 54–57, 2021. (Cited on page 2.)

[13] 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, 2014. (Cited on pages 2 and 4.)

[14] 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. (Cited on page 3.)

[15] John D Bancroft and Marilyn Gamble. *Theory and Practice of Histological Techniques (sixth Edition).* Elsevier Health Sciences, sixth edition, 2008. (Cited on page 3.)

[16] Anthony Mescher. *Junqueira's Basic Histology: Text & Atlas.* McGraw-Hill Education, sixth edition, 2018. ISBN 978-1-26-002618-4. (Cited on pages 3 and 5.)

[17] 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. (Cited on page 3.)

[18] Kim S. Suvarna, Christopher Layton, and John D. Bancroft. *Bancroft's Theory and Practice of Histological Techniques.* Elsevier, eighth edition, 2019. ISBN 978-0702068645. (Cited on page 4.)

[19] Michael Titford. The long history of hematoxylin. *Biotechnic & Histochemistry*, 80(2):73–78, 2005. (Cited on page 4.)

[20] 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. (Cited on page 4.)

[21] Kurt E. Johnson. *Histology and Cell Biology.* The National medical series for independent study. Harwal Pub. Co., 1991. ISBN 978-0683062106. (Cited on page 5.)

[22] 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. (Cited on page 5.)

[23] Jelica Vasiljević. *Generative adversarial networks in digital histopathology : stain transfer and deep learning model invariance to stain variation.* Theses, Université de Strasbourg, 2022. (Cited on pages 5, 6, 8, 10, 18 and 109.)

[24] 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. (Cited on page 5.)

[25] 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. (Cited on page 6.)

[26] 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. (Cited on page 6.)

[27] Thomas Lampert, Odyssée Merveille, Jessica Schmitz, Germain Forestier, Friedrich Feuerhake, and Cédric Wemmert. Strategies for Training Stain Invariant CNNS. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 905–909, April 2019. doi: 10.1109/ISBI.2019.8759266. ISSN: 1945-8452. (Cited on pages 7, 8, 56, 58, 68 and 77.)

[28] 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, October 2021. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.07.005. (Cited on pages 7, 8, 10, 18, 21, 22, 24, 28, 31, 41, 49, 52, 62, 68, 77, 108 and 109.)

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28. (Cited on pages 7, 23, 39, 52, 61 and 77.)

[30] Aïcha Bentaieb and G. Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE Transactions on Medical Imaging*, 37:792–802, 2018. URL https://api.semanticscholar.org/CorpusID:3668673. (Cited on page 8.)

[31] Thomas de Bel, Meyke Hermsen, Jesper Kers, Jeroen van der Laak, and Geert Litjens. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *International Conference on Medical Imaging with Deep Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:146033099. (Cited on page 40.)

[32] 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 *Deep Learning in Medical Image*

Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3, pages 83–91. Springer, 2017. (Not cited.)

[33] Gaoyi Lei, Yuanqing Xia, Di-Hua Zhai, Wei Zhang, Duanduan Chen, and Defeng Wang. Staincnns: An efficient stain feature learning method. *Neurocomputing*, 406:267–273, 2020. (Not cited.)

[34] Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, and Mitko Veta. Learning domain-invariant representations of histological images. *Frontiers in Medicine*, 6, 2019. URL https://api.semanticscholar.org/CorpusID:196611135. (Not cited.)

[35] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, 2009. URL https://api.semanticscholar.org/CorpusID:15008471. (Not cited.)

[36] 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. (Cited on pages 8, 18, 28, 39, 40, 49 and 108.)

[37] 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(10): 2293–2302, October 2019. ISSN 1558-254X. doi: 10.1109/TMI.2019.2899364. Conference Name: IEEE Transactions on Medical Imaging. (Cited on pages 8, 10, 18, 20, 22, 28, 39, 41, 49, 52, 62 and 108.)

[38] Caner Mercan, Germonda Reijnen-Mooij, David Tellez Martin, Johannes Lotz, Nick Weiss, Marcel van Gerven, and Francesco Ciompi. Virtual staining for mitosis detection in breast histopathology. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1770–1774, 2020. URL https://api.semanticscholar.org/CorpusID:212737036. (Cited on pages 8 and 34.)

[39] 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. (Cited on page 8.)

[40] China Focus: AI beats human doctors in neuroimaging recognition contest, 2018. URL http://www.xinhuanet.com/english/2018-06/30/c_137292451.htm. (Not cited.)

[41] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, December 2017. (Cited on page 8.)

[42] 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. (Cited on pages 8, 16, 24, 51 and 117.)

[43] 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. (Cited on page 8.)

[44] 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. (Cited on pages 9 and 15.)

[45] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, December 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.101539. (Cited on pages 9, 25, 26, 51 and 117.)

[46] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022. (Cited on pages 10, 24, 26 and 28.)

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, October 2017. doi: 10.1109/ICCV.2017.244. ISSN: 2380-7504. (Cited on pages 10, 18, 40, 78 and 107.)

[48] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography, December 2017. (Cited on pages 10, 28, 34, 36, 43, 44, 46, 47, 48, 80, 110, 113, 115 and 117.)

[49] 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. (Cited on pages 10, 34, 44, 46, 47, 48, 79, 115 and 117.)

[50] Tongzhou Wang and Yihan Lin. CycleGAN with Better Cycles, 2018. (Cited on pages 10, 28, 34, 36, 43, 44, 110 and 113.)

[51] 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, May 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw013. (Cited on page 12.)

[52] Zeeshan Nisar, Jelica Vasiljević, Pierre Gançarski, and Thomas Lampert. Towards Measuring Domain Shift in Histopathological Stain Translation in an Unsupervised Manner. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. doi: 10.1109/ISBI52829.2022.9761411. (Cited on pages 12, 38, 66, 68, 110, 121 and 122.)

[53] Jelica Vasiljević, Zeeshan Nisar, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. CycleGAN for virtual stain transfer: Is seeing really believing? *Artificial Intelligence in Medicine*, 133:102420, 2022. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2022.102420. (Cited on pages 12, 28, 40 and 111.)

[54] Zeeshan Nisar and Thomas Lampert. Maximising Histopathology Segmentation using Minimal Labels via Self-Supervision. *(under review)*. (Cited on pages 13 and 111.)

[55] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021. (Cited on pages 16 and 21.)

[56] 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. (Cited on page 18.)

[57] 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. (Cited on page 18.)

[58] Ming-Yu Liu, Thomas M. Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:3783306. (Cited on page 18.)

[59] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:4883312. (Not cited.)

[60] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:260444154. (Not cited.)

[61] 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. (Not cited.)

[62] 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. (Not cited.)

[63] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. URL https://api.semanticscholar.org/CorpusID:220871180. (Not cited.)

[64] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJlZ5ySKPH. (Not cited.)

[65] Hongtao Kang, Die Luo, Weihua Feng, Li Chen, Junbo Hu, Shaoqun Zeng, Tingwei Quan, and Xiuli Liu. Stainnet: A fast and robust stain normalization network. *Frontiers in Medicine*, 8, 2020. URL https://api.semanticscholar.org/CorpusID:229363501. (Cited on pages 28 and 40.)

[66] Xinyang Li, Guoxun Zhang, Hui Qiao, Feng Bao, Yue Deng, Jiamin Wu, Yang fan He, Jingping Yun, Xing Lin, Hao Xie, Haoqian Wang, and Qionghai Dai. Unsupervised content-preserving transformation for optical microscopy. *Light, Science & Applications*, 10, 2021. URL https://api.semanticscholar.org/CorpusID:232090954. (Not cited.)

[67] Yi-Ting Lin, Bowei Zeng, Yifeng Wang, Yang Chen, Zi yu Fang, Jian Zhang, Xiang Ji, Haoqian Wang, and Yongbing Zhang. Unpaired multi-domain stain transfer for kidney histopathological images. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:249301062. (Cited on page 18.)

[68] 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. *Knowledge-Based Systems*, 277:110780, 2023. ISSN 0950-7051. doi: https://doi.

org/10.1016/j.knosys.2023.110780. URL https://www.sciencedirect.com/science/article/pii/S0950705123005300. (Cited on pages 18 and 21.)

[69] Igor Zingman, Sergio Frayle, Ivan Tankoyeu, Segrey Sukhanov, and Fabian Heinemann. A comparative evaluation of image-to-image translation methods for stain transfer in histopathology. In *International Conference on Medical Imaging with Deep Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:257834181. (Cited on pages 18 and 40.)

[70] Lorenzo Veronese, Isabella Poles, Eleonora D'Arnese, and Marco D. Santambrogio. Stain transfer using cyclegan for histopathological images. In *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, pages 1–5, 2023. doi: 10.1109/EUROCON56442.2023.10199027. (Cited on pages 18, 49 and 108.)

[71] Joseph Boyd, Irène Villa, Marie-Christine Mathieu, Eric Deutsch, Nikos Paragios, Maria Vakalopoulou, and Stergios Christodoulidis. Region-guided Cyclegans for Stain Transfer in Whole Slide Images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 356–365. Springer, 2022. (Not cited.)

[72] 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. (Cited on pages 28 and 39.)

[73] 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. (Cited on pages 18, 28, 39, 40, 41, 49 and 108.)

[74] 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. *Appl. Soft Comput.*, 98:106822, 2020. URL https://api.semanticscholar.org/CorpusID:226327875. (Cited on page 20.)

[75] Bingzhe Wu, Xiaolu Zhang, Shiwan Zhao, Lingxi Xie, Caihong Zeng, Zhihong Liu, and Guangyu Sun. G2c: A generator-to-classifier framework integrating multi-stained visual cues for pathological glomerulus classification. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://api.semanticscholar.org/CorpusID:49652811. (Cited on pages 20 and 21.)

[76] 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 Computational Pathology Workshop COMPAY*, 2019. URL https://openreview.net/forum?id=SkxOZOh2gr. (Cited on page 20.)

[77] Nassim Bouteldja, Barbara M. Klinkhammer, Tarek Schlaich, Peter Boor, and Dorit Merhof. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. *Journal of Pathology Informatics*, 13:100107, 2022. ISSN 2153-3539. (Cited on pages 21, 44, 46, 47, 48, 80, 81, 115 and 117.)

[78] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *Journal of machine learning research*, 2015. URL https://api.semanticscholar.org/CorpusID:2871880. (Cited on page 21.)

[79] Ke Mei, Chuang Zhu, Lei Jiang, Jun Liu, and Yuanyuan Qiao. Cross-stained segmentation from renal biopsy images using multi-level adversarial learning. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1424–1428, 2020. URL https://api.semanticscholar.org/CorpusID:211204773. (Cited on page 21.)

[80] Xianxu Hou, Jingxin Liu, Bolei Xu, Bozhi Liu, Xin Chen, Mohammad Ilyas, Ian O. Ellis, Jonathan Mark Garibaldi, and Guoping Qiu. Dual adaptive pyramid network for cross-stain histopathology image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019. URL https://api.semanticscholar.org/CorpusID:202749971. (Cited on page 21.)

[81] Laxmi Gupta, Barbara Mara Klinkhammer, Peter Boor, Dorit Merhof, and Michael Gadermayr. Gan-based image enrichment in digital pathology boosts segmentation accuracy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019. URL https://api.semanticscholar.org/CorpusID:204027370. (Cited on page 21.)

[82] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, December 2017. ISSN 1361-8415. doi: 10.1016/j.media.2017.07.005. (Cited on pages 23, 61 and 77.)

[83] Thomas de Bel, Meyke Hermsen, Bart Smeets, Luuk Hilbrands, Jeroen van der Laak, and Geert Litjens. Automatic segmentation of histopathological slides of renal tissue using deep learning. In *Medical Imaging 2018: Digital Pathology*, volume 10581, pages 285–290. SPIE, March 2018. doi: 10.1117/12.2293717. (Cited on pages 23, 60, 61 and 77.)

[84] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, June 2016. (Cited on pages 25 and 26.)

[85] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful Image Colorization. In *Computer Vision – ECCV 2016*, pages 649–666, 2016. (Cited on pages 25 and 26.)

[86] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJtNZAFgg. (Cited on page 25.)

[87] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, February 2018. (Cited on pages 25, 26, 51, 110 and 117.)

[88] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014. URL https://api.semanticscholar.org/CorpusID:2930547. (Cited on page 25.)

[89] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98 – 136, 2014. URL https://api.semanticscholar.org/CorpusID:207252270. (Cited on page 25.)

[90] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *Computer Vision – ECCV 2016*, pages 69–84. Springer International Publishing, 2016. (Cited on pages 25 and 26.)

[91] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, December 2015. (Cited on pages 25 and 26.)

[92] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. (Cited on pages 25 and 27.)

[93] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767. (Cited on page 25.)

[94] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *In-*

ternational conference on machine learning, pages 1597–1607. PMLR, 2020. (Cited on pages 25, 26, 27, 29, 51, 52, 53, 54, 60 and 117.)

[95] Alex Krizhevsky. Learning multiple layers of features from tiny images. Theses, University of Toronto, 2009. (Cited on page 25.)

[96] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. (Cited on pages 25, 27, 29, 51, 52, 54, 55, 60 and 117.)

[97] Mangal Prakash, Tim-Oliver Buchholz, Manan Lalit, Pavel Tomancak, Florian Jug, and Alexander Krull. Leveraging Self-supervised Denoising for Image Segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 428–432, April 2020. (Cited on pages 25, 26 and 68.)

[98] Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, Hannes Kenngott, Stefanie Speidel, Annette Kopp-Schneider, Klaus Maier-Hein, and Lena Maier-Hein. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, 13:925–933, June 2018. (Cited on pages 25 and 26.)

[99] Szu-Yen Hu, Shuhang Wang, Wei-Hung Weng, JingChao Wang, XiaoHong Wang, Arinc Ozturk, Quan Li, Viksit Kumar, and Anthony E. Samir. Self-Supervised Pretraining with DICOM metadata in Ultrasound Imaging. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, pages 732–749. PMLR, September 2020. (Cited on pages 25 and 26.)

[100] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal Self-supervised Learning for Medical Image Analysis. In *Information Processing in Medical Imaging*, pages 661–673, 2021. (Cited on pages 25 and 26.)

[101] Mihir Sahasrabudhe, Stergios Christodoulidis, Roberto Salgado, Stefan Michiels, Sherene Loi, Fabrice André, Nikos Paragios, and Maria Vakalopoulou. Self-supervised Nuclei Segmentation in Histopathological Images Using Attention. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 393–402, 2020. (Cited on pages 25 and 26.)

[102] Ming Y Lu, Richard J Chen, and Faisal Mahmood. Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (conference presentation). In *Medical imaging 2020: digital pathology*, volume 11320, page 113200J. SPIE, 2020. (Cited on pages 25, 27 and 51.)

[103] Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models. In *Medical Imaging with Deep Learning*, pages 728–744, February 2021. (Cited on pages 25 and 27.)

[104] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. *Pattern Recognition*, 113:107826, May 2021. (Cited on pages 25 and 27.)

[105] Karin Stacke, Jonas Unger, Claes Lundström, and Gabriel Eilertsen. Learning representations with contrastive self-supervised learning for histopathology applications. *Machine Learning for Biomedical Imaging*, 1:1–33, 2022. (Cited on pages 25, 27, 28, 29, 51, 53, 60 and 62.)

[106] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7: 100198, 2022. (Cited on pages 25, 27, 28, 29, 51, 54, 61 and 118.)

[107] Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen. Pgl: Prior-guided local self-supervised learning for 3d medical image segmentation. *ArXiv*, abs/2011.12640, 2020. URL https://api.semanticscholar.org/CorpusID:227162415. (Cited on page 25.)

[108] Pengshuai Yang, Xiaoxu Yin, Haiming Lu, Zhongliang Hu, Xuegong Zhang, Rui Jiang, and Hairong Lv. CS-CO: A hybrid self-Supervised Visual Representation Learning Method for H&E-stained Histopathological Images. *Medical Image Analysis*, 81:102539, 2022. (Cited on pages 25, 26, 27, 28, 51, 52, 56 and 117.)

[109] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-Path: Self-Supervision for Classification of Pathology Images With Limited Annotations. *IEEE Transactions on Medical Imaging*, 40(10):2845–2856, October 2021. ISSN 1558-254X. doi: 10.1109/TMI.2021.3056023. Conference Name: IEEE Transactions on Medical Imaging. (Cited on pages 25, 27, 28 and 51.)

[110] Nanqing Dong, Michael C. Kampffmeyer, and Irina Voiculescu. Self-supervised multi-task representation learning for sequential medical images. In *ECML/PKDD*, 2021. URL https://api.semanticscholar.org/CorpusID:236207287. (Cited on page 25.)

[111] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy Layer-Wise Training of Deep Networks. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. (Cited on page 24.)

[112] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. Association for Computing Machinery, July 2008. (Cited on page 26.)

[113] Joseph Boyd, Mykola Liashuha, Eric Deutsch, Nikos Paragios, Stergios Christodoulidis, and Maria Vakalopoulou. Self-supervised representation learning using visual field expansion on digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 639–647, 2021. (Cited on pages 26 and 51.)

[114] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934, 2020. (Cited on page 26.)

[115] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020. (Cited on page 26.)

[116] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. (Cited on page 27.)

[117] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y. Ng, and Pranav Rajpurkar. MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, pages 755–769, October 2021. (Cited on page 27.)

[118] Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Shan E Ahmed Raza, Fayyaz Minhas, David Snead, and Nasir Rajpoot. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83:102685, 2023. (Cited on pages 27 and 51.)

[119] Yuhan Zhang, Mingchao Li, Zexuan Ji, Wen Fan, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Twin self-supervision based semi-supervised learning (TS-SSL): Retinal anomaly classification in SD-OCT images. *Neurocomputing*, 462:491–505, 2021. (Cited on page 27.)

[120] Zhaoyang Xu, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. Gan-based virtual re-staining: A promising solution for whole slide image analysis. *ArXiv*, abs/1901.04059, 2019. URL `https://api.semanticscholar.org/CorpusID:58004749`. (Cited on page 27.)

[121] Xu Jin, Teng Huang, Ke Wen, Mengxian Chi, and Hong An. HistoSSL: Self-Supervised Representation Learning for Classifying Histopathology Images. *Mathematics*, 11(1):110, 2022. (Cited on pages 28 and 51.)

[122] Alexandre Tiard, Alex Wong, David Joon Ho, Yangchao Wu, Eliram Nof, Alvin C Goh, Stefano Soatto, and Saad Nadeem. Stain-invariant self supervised learning for histopathology image analysis. *arXiv preprint arXiv:2211.07590*, 2022. (Cited on pages 28 and 51.)

[123] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking Self-Supervised Learning on Diverse Pathology Datasets. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3344–3354, 2023. (Cited on pages 29, 73 and 124.)

[124] Jelica Vasiljević, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Self Adversarial Attack as an Augmentation Method for Immunohistochemical Stainings. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1939–1943. IEEE, 2021. (Cited on page 31.)

[125] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020. (Cited on pages 31 and 110.)

[126] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017. (Cited on pages 35, 79 and 111.)

[127] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes F. Lundström. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25:325–336, 2020. (Cited on pages 35, 36, 111 and 112.)

[128] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018. (Cited on pages 35 and 112.)

[129] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19:47, 2015. URL https://api.semanticscholar.org/CorpusID:7725237. (Cited on pages 36, 112 and 113.)

[130] Spyridon N Papageorgiou. On correlation coefficients and their interpretation. *Journal of orthodontics*, 49(3):359–361, 2022. (Cited on pages 38 and 114.)

[131] Shaojin Cai, Yuyang Xue, Qinquan Gao, Min Du, Gang Chen, Hejun Zhang, and Tong Tong. Stain Style Transfer Using Transitive Adversarial Networks. In *MLMIR@MICCAI*, 2019. URL https://api.semanticscholar.org/CorpusID:204838399. (Cited on pages 39, 40 and 41.)

[132] 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*, pages 770–778, 2016. (Cited on pages 39 and 79.)

[133] Yijie Zhang, Kevin de Haan, Y. Rivenson, Jingxi Li, A. Delis, and A. Ozcan. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light, Science & Applications*, 9, 2020. (Cited on pages 40 and 41.)

[134] Henri Hoyez, Cedric Schockaert, Jason Raphael Rambach, Bruno Mirbach, and Didier Stricker. Unsupervised image-to-image translation: A review. *Sensors (Basel, Switzerland)*, 22, 2022. URL https://api.semanticscholar.org/CorpusID:253412890. (Cited on page 40.)

[135] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21:1500–1512, 2012. URL https://api.semanticscholar.org/CorpusID:14326712. (Cited on page 40.)

[136] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30:79–82, 2005. URL https://api.semanticscholar.org/CorpusID:120556606. (Cited on page 40.)

[137] 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 *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:326772. (Cited on page 40.)

[138] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. (Cited on page 40.)

[139] Stanislav Morozov, Andrey Voynov, and Artem Babenko. On self-supervised image representations for {gan} evaluation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NeRdBeTionN. (Cited on page 40.)

[140] Richard Osuala, Grzegorz Skorupko, Noussair Lazrak, Lidia Garrucho, Eloy García, Smriti Joshi, Socayna Jouide, Michael Rutherford, Fred Prior, Kaisar Kushibar, et al. medigan: a python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging*, 10(6):061403–061403, 2023. (Not cited.)

[141] Lorenzo Tronchin, Rosa Sicilia, Ermanno Cordelli, Sara Ramella, and Paolo Soda. Evaluating gans in medical imaging. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, pages 112–121. Springer, 2021. (Cited on page 40.)

[142] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P. Yung, Ankit B. Patel, and Kristy K. Brock. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend, 2024. (Cited on page 40.)

[143] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *ArXiv*, abs/1607.08022, 2016. URL https://api.semanticscholar.org/CorpusID:16516553. (Cited on pages 40 and 41.)

[144] 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. (Cited on page 41.)

[145] 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. (Cited on page 41.)

[146] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Ling Shao. Structure Preserving Stain Normalization of Histopathology Images Using Self-Supervised Semantic Guidance. In *MICCAI*, 2021. (Cited on page 41.)

[147] Yuxin Wu and Kaiming He. Group Normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. (Cited on page 41.)

[148] Hanwen Liang, Konstantinos Plataniotis, and Xingyu Li. Stain Style Transfer of Histopathology Images via Structure-Preserved Generative Learning. In *MLMIR*, 2020. (Cited on page 41.)

[149] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. In *Advances in NIPS 2016 Deep Learning Symposium*, 2016. (Cited on page 41.)

[150] Ori Nizan and Ayellet Tal. Breaking the cycle – colleagues are all you need. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7857–7866, 2019. URL https://api.semanticscholar.org/CorpusID:208268379. (Cited on pages 43 and 44.)

[151] Xuning Shao and Weidong Zhang. Spatchgan: A statistical feature based discriminator for unsupervised image-to-image translation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6526–6535, 2021. URL https://api.semanticscholar.org/CorpusID:232417108. (Cited on pages 43 and 44.)

[152] Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 44.)

[153] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv preprint arXiv:2012.03583*, 2020. (Cited on page 51.)

[154] Tristan Lazard, Marvin Lerousseau, Etienne Decencière, and Thomas Walter. Giga-SSL: Self-Supervised Learning for Gigapixel Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4304–4313, 2023. (Cited on page 51.)

[155] Hai-Li Ye and Da-Han Wang. Stain-Adaptive Self-Supervised Learning for Histopathology Image Analysis. *arXiv preprint arXiv:2208.04017*, 2022. (Cited on page 51.)

[156] Nicklas Boserup and Raghavendra Selvan. Efficient self-supervision using patch-based contrastive learning for histopathology image segmentation. *Proceedings of the Northern Lights Deep Learning Workshop*, 4, 2023. (Cited on page 51.)

[157] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, June 2021. doi: 10.1109/CVPR46437.2021.01549. Conference Name: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ISBN: 9781665445092 Place: Nashville, TN, USA Publisher: IEEE. (Cited on pages 51 and 58.)

[158] Shuo Li et al. Minent: Minimum entropy for self-supervised representation learning. *Pattern Recognit.*, 138:109364, 2023. (Cited on page 51.)

[159] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, 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, August 2016. ISSN 1558-254X. doi: 10.1109/TMI.2016.2529665. Conference Name: IEEE Transactions on Medical Imaging. (Cited on page 56.)

[160] A. C. Ruifrok and D. A. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299, August 2001. (Cited on page 56.)

[161] Davide Chicco. Siamese Neural Networks: An Overview. In Hugh Cartwright, editor, *Artificial Neural Networks*, Methods in Molecular Biology, pages 73–94. Springer US, New York, NY, 2021. ISBN 978-1-07-160826-5. doi: 10.1007/978-1-0716-0826-5_3. (Cited on page 58.)

[162] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. 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, September 2018. ISSN 1558-254X. doi: 10.1109/TMI.2018.2820199. Conference Name: IEEE Transactions on Medical Imaging. (Cited on pages 58 and 78.)

[163] Pierre H. Richemond, Jean-Bastien Grill, Florent Altch'e, Corentin Tallec, Florian Strub, Andrew Brock, Samuel L. Smith, Soham De, Razvan Pascanu, Bilal Piot, and Michal Valko. BYOL works even without batch statistics. *ArXiv*, October 2020. (Cited on page 60.)

[164] Odyssee Merveille et al. An automatic framework for fusing information from differently stained consecutive digital whole slide images: A case study in renal histology. *Comput. Methods Programs Biomed*, 208:106157, 2021. (Cited on page 68.)

[165] Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive Visual Representations. *Advances in Neural Information Processing Systems*, 34:19538–19552, November 2021. (Cited on page 68.)

[166] Narinder Singh Punn et al. BT-Unet: A self-supervised learning framework for biomedical image segmentation using barlow twins with U-net models. *Machine Learning*, 111(12):4585–4600, 2022. (Cited on page 68.)

[167] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, January 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.106391. (Cited on pages 69, 73 and 124.)

[168] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. (Cited on pages 73 and 123.)

[169] Yiqing Shen and Jing Ke. Staindiff: Transfer stain styles of histology images with denoising diffusion probabilistic models and self-ensemble. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 549–559. Springer, 2023. (Cited on pages 73 and 123.)

[170] Muzaffer Ozbey, Salman UH Dar, Hasan Atakan Bedel, Onat Dalmaz, cSaban Ozturk, Alper Gungor, and Tolga cCukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42:3524–3539, 2022. URL https://api.semanticscholar.org/CorpusID:250627054. (Cited on pages 73 and 123.)

[171] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, June 2022. doi: 10.1109/CVPR52688.2022.01553. ISSN: 2575-7075. (Cited on pages 73 and 124.)

[172] Zhaowen Li, Zhiyang Chen, F. Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. MST: Masked Self-Supervised Transformer for Visual Representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, June 2021. (Cited on pages 73 and 124.)

[173] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from Simulated and Unsupervised Images through Adversarial Training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2242–2251, July 2017. doi: 10.1109/CVPR.2017.241. ISSN: 1063-6919. (Cited on page 78.)

[174] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In *International Conference on Machine Learning*, 2016. URL https://api.semanticscholar.org/CorpusID:8142135. (Cited on page 79.)

[175] Nicolas Papernot, Patrick Mcdaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597, 2015. URL https://api.semanticscholar.org/CorpusID:2672720. (Cited on page 79.)

[176] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Jun Zhu, and Xiaolin Hu. Defense against adversarial attacks using high-level representation guided denoiser. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2017. URL https://api.semanticscholar.org/CorpusID:604742. (Not cited.)

[177] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb. (Not cited.)

[178] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ. (Cited on page 79.)

[179] Brady Zhou and Philipp Krähenbühl. Don't let your discriminator be fooled. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:108308444. (Cited on page 79.)

[180] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, February 2020. ISSN 2078-2489. doi: 10.3390/info11020125. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. (Cited on page 83.)

[181] Dina Bashkirova et al. Adversarial self-defense for cycle-consistent GANs, 2019. (Cited on page 110.)

Résumé

Contexte

L'intégration de l'intelligence artificielle, en particulier du deep learning (DL), avec l'imagerie médicale offre d'énormes promesses et potentiels. Les systèmes de diagnostic assisté par ordinateur (CAD) automatisés, alimentés par le deep learning, sont devenus l'un des domaines de recherche les plus importants dans le domaine de l'imagerie médicale. Dans un tel environnement, l'histopathologie numérique ne fait pas exception. Cependant, un défi majeur dans l'application du deep learning à l'histopathologie réside dans les variations inter- et intra-colorations, voir la Figure 1, résultant de différentes colorations et protocoles. Ces variations conduisent au problème de changement de domaine, ce qui impacte de manière significative la performance des modèles de deep learning à la pointe de la technologie (SOTA) entraînés pour une coloration (également appelée domaine en apprentissage automatique) lorsqu'ils sont appliqués à d'autres colorations (même pour la même tâche). Ce comportement est illustré dans le Tableau 1 (1$^{\text{ère}}$ ligne), où un modèle U-Net entraîné pour la segmentation des glomérules rénaux en utilisant la coloration PAS démontre une baisse notable de performance de segmentation lorsqu'il est appliqué à d'autres colorations. L'acquisition d'étiquettes pour chaque coloration est un processus chronophage et coûteux, principalement en raison de la nécessité de faire appel à des experts médicaux hautement spécialisés pour étiqueter les données.

Pour surmonter ces défis, le transfert de coloration — où l'apparence d'une image est modifiée artificiellement après son acquisition à l'aide d'un cadre de traduction d'image à image non appariée basé sur CycleGAN [47] — a émergé comme la solution à l'état de l'art (SOTA). Ce processus vise à transformer une image colorée

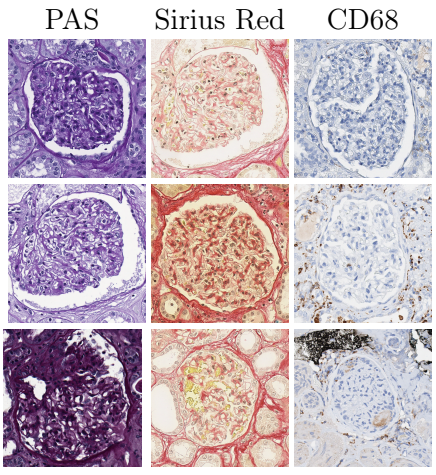


Figure 1: Variabilité de la coloration des structures glomérulaires en histopathologie rénale, les variations inter-colorations étant représentées par les lignes et les variations intra-coloration étant représentées par les colonnes.

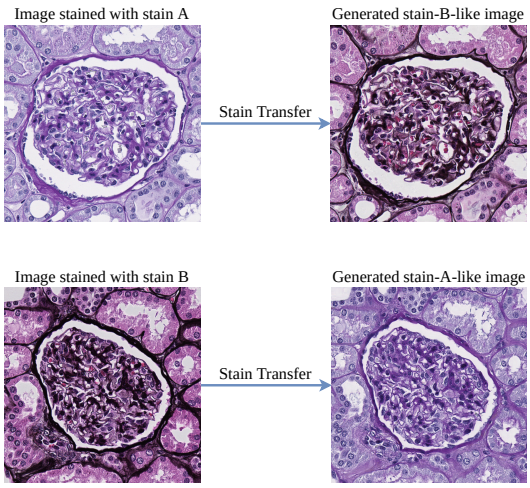


Figure 2: Le principe de base du transfert de coloration consiste à modifier artificiellement l'apparence d'une image. Dans la 1ère ligne, une image avec les caractéristiques de la coloration A est artificiellement transformée pour ressembler à une image avec les caractéristiques de la coloration B. Dans la 2ème ligne, une image avec les caractéristiques de la coloration B est artificiellement transformée pour ressembler à une image avec les caractéristiques de la coloration A.

avec la coloration A pour qu'elle ressemble à une image colorée avec la coloration B, et vice-versa. Un exemple du résultat du transfert de coloration est illustré dans la Figure 2, où une image de type coloration B est artificiellement générée à partir de la coloration A et une image de type coloration A est artificiellement générée à partir de la coloration B. D'un point de vue vision par ordinateur, ces images générées artificiellement peuvent être utilisées pour minimiser les disparités de distribution (c'est-à-dire le décalage de domaine) entre la coloration A et la coloration B, améliorant ainsi la robustesse d'un modèle face à différentes variations de coloration. Par exemple, les modèles d'apprentissage profond entraînés sur des images réelles de la coloration A devraient être capables d'extraire des caractéristiques similaires et de fonctionner efficacement sur des images de type coloration A (traduites à partir de la coloration B), et vice-versa. Cela ouvre la voie au développement de diverses méthodes de segmentation multi-colorations [28, 36, 37, 70–73], utilisant des étiquettes d'une seule (source) coloration. Ces méthodes peuvent être principalement classées en deux stratégies d'entraînement différentes:

Spécifique au colorant : Entraînement d'un modèle de segmentation pour un colorant particulier, appelé le colorant source, pour lequel les étiquettes sont disponibles. Ce modèle spécifique au colorant est ensuite appliqué à divers autres colorants cibles en les traduisant en colorant source pendant le test. La Méthode Multi-Domain Supervised 1 (MDS1) [37] est actuellement la méthode SOTA dans cette stratégie d'entraînement.

Invariant au colorant : Entraînement d'un modèle de segmentation invariant au

Table 1: Le principe de base du transfert de coloration consiste à modifier artificiellement l'apparence d'une image. Dans la 1ère ligne, une image avec les caractéristiques de la coloration A est artificiellement transformée pour ressembler à une image avec les caractéristiques de la coloration B. Dans la 2ème ligne, une image avec les caractéristiques de la coloration B est artificiellement transformée pour ressembler à une image avec les caractéristiques de la coloration A.

Training Strategy	Test Stains				
	HC Stains			IHC Stains	
	PAS	Jones H&E	Sirius Red	CD68	CD34
Baseline PAS	0.894 (0.021)	0.062 (0.011)	0.045 (0.037)	0.044 (0.098)	0.056 (0.090)
MDS1 (Target→PAS)	0.894 (0.021)	0.849 (0.017)	0.870 (0.009)	0.683 (0.043)	0.754 (0.008)
UDAGAN (Stain-Invariant)	0.901 (0.011)	0.856 (0.036)	0.873 (0.025)	0.705 (0.031)	0.799 (0.035)

colorant sur tous les colorants disponibles, en utilisant les étiquettes d'un seul (source) colorant. Ce modèle invariant au colorant peut être directement appliqué à divers autres colorants, y compris les colorants hors distribution, sans besoin de traduction pendant le test. L'Augmentation de Domaine Non Supervisée utilisant des Réseaux Adversariaux Génératifs (UDAGAN) [28] est actuellement la méthode SOTA dans cette stratégie d'entraînement.

Les résultats obtenus avec les modèles de segmentation multi-teinte basés sur le transfert de teinte (y compris MDS1 et UDAGAN) sont également présentés dans le Tableau 1. Étant donné que le transfert de teinte est capable de produire des traductions plausibles[1] et vise à correspondre à la distribution d'une teinte ciblée, comme le montre la Figure 2, il est raisonnable de s'attendre à ce que le modèle de base spécifique à cette teinte soit capable d'extraire un ensemble similaire (ou un sous-ensemble) de caractéristiques dans les teintes traduites. Par exemple, le modèle de base entraîné sur la teinte PAS devrait être capable d'extraire des caractéristiques similaires dans les teintes traduites Target→PAS, ce qui est le cas pour MDS1. D'autre part, UDAGAN utilise des traductions depuis la direction opposée (c'est-à-dire PAS→Target) pour augmenter les données d'entraînement, ce qui facilite l'apprentissage de caractéristiques plus générales (invariantes à la teinte) et robustes. En conséquence, une amélioration significative de la performance de segmentation des modèles de segmentation multi-teinte est observée, comme le montre le Tableau 1 (2ème et 3ème ligne), malgré un entraînement uniquement sur des

[1]Le terme "plausible" fait référence au fait qu'une image histologique, lorsqu'elle est traitée avec d'autres modalités de teinture sans la connaissance des sections de tissu adjacentes et/ou des informations spécifiques au patient (par exemple, la maladie sous-jacente), semble visuellement correcte pour un expert formé en ce qui concerne les caractéristiques de teinture et l'apparence morphologique des composants tissulaires [23].

étiquettes provenant de la teinte PAS.

Bien que efficaces, les méthodes de transfert de taches basées sur CycleGAN sont susceptibles d'être vulnérables aux attaques auto-adversariales [48, 50, 181], où des informations supplémentaires sous la forme de bruit imperceptible sont ajoutées aux traductions de taches générées artificiellement. Ce bruit imperceptible introduit un décalage de domaine dans les taches traduites (images) [125], ce qui peut potentiellement affecter les prédictions finales des méthodes de segmentation multi-taches, comme observé pour les taches ImmunoHistoChimiques (IHC), et suscite donc des inquiétudes quant à l'efficacité et au déploiement des méthodes de segmentation multi-taches dans les applications cliniques. Par conséquent, il est urgent de développer des techniques de transfert de taches plus avancées capables de résoudre efficacement ces limitations, conduisant ainsi à une performance améliorée des approches de segmentation multi-taches.

Bien que ces méthodes éliminent avec succès la nécessité de labels dans les taches cibles, il est crucial de reconnaître que ces méthodes dépendent fortement d'une grande quantité de données étiquetées provenant de la tache source. Cependant, obtenir une quantité suffisante de données étiquetées pour le domaine source reste un défi dans diverses disciplines médicales. Bien que les ensembles de données étiquetées puissent être limités, les avancées dans les technologies de vision par ordinateur et d'imagerie médicale ont considérablement augmenté la disponibilité des données non étiquetées. Dans de telles situations, où les données non étiquetées sont disponibles en grandes quantités, elles peuvent être utilisées dans des scénarios avec peu de labels pour améliorer la performance du modèle grâce à l'apprentissage de représentations non supervisé [87].

Contributions

Cette thèse examine en profondeur et aborde les défis inhérents au transfert de taches, en particulier l'introduction de bruit imperceptible dans les taches traduites, ce qui entraîne le problème de décalage de domaine. Un aspect crucial pour atténuer ce décalage de domaine est la capacité à le détecter. Ainsi, l'une des contributions principales de cette thèse réside dans l'exploration des approches non supervisées pour proposer une métrique permettant de le quantifier. De plus, les résultats révèlent une forte corrélation entre le décalage de domaine et la performance de segmentation des taches traduites. Ces éclaircissements ouvrent donc la voie à l'établissement d'un mécanisme pour déduire la performance moyenne d'un modèle pré-entraîné (entraîné sur un domaine source) lorsqu'il est appliqué à un domaine cible non vu et non étiqueté. Les résultats de cette contribution sont publiés dans [52].

En utilisant cette mesure, nous démontrons la sensibilité du transfert de tâches basé sur CycleGAN aux modifications architecturales subtiles. Bien que ces modifications puissent ne pas nécessairement affecter la qualité visuelle des traductions résultantes, elles ont un impact significatif sur la performance globale des approches de segmentation multi-taches basées sur le transfert de tâches. Cela est vrai à la fois du point de vue diagnostique et appliqué, soulignant ainsi la deuxième contribution

de la thèse. Les résultats de cette contribution sont publiés dans [53].

Nous proposons ensuite une approche novatrice qui minimise l'ajout de bruit (changement de domaine) pendant le transfert de tâches, améliorant ainsi la performance de la segmentation multi-taches, ce qui met en avant la troisième contribution de la thèse. Les résultats de ces travaux sont actuellement en cours de documentation pour être soumis à une conférence ou un journal réputé.

La quatrième et dernière contribution de cette thèse concerne l'intégration des méthodes d'apprentissage de représentation à l'état de l'art, en particulier l'apprentissage auto-supervisé (SSL), afin de réaliser une analyse complète pour réduire le nombre d'étiquettes nécessaires à la segmentation histopathologique. De plus, cette contribution cherche à améliorer les approches de segmentation multi-taches en réduisant leur dépendance aux données étiquetées pour la tâche source—ce qui, à notre connaissance, n'a pas été exploré auparavant—ouvrant la voie à des solutions plus économiques et évolutives pour les algorithmes d'adaptation de domaine. Cette contribution propose également plusieurs modifications pour améliorer l'adaptabilité des méthodes SSL à travers divers protocoles de coloration, en particulier ceux qui sont spécifiques aux tâches et donc limités à un seul type de coloration. Les résultats de cette enquête sont actuellement en cours d'examen [54].

Méthodes et Résultats

Mesurer le Changement de Domaine dans le Transfert de Taches

Étant donné la baisse de performance causée par le bruit imperceptible (changement de domaine) pour les colorations IHC dans le Tableau 1, il est important de gérer ce changement de domaine ou du moins d'estimer quand il est susceptible d'affecter la performance d'un algorithme. Par conséquent, cette méthode est proposée pour détecter et quantifier le changement de domaine lors du transfert de teinture entre une teinture source (PAS) et des teintures traduites Target→PAS. À notre connaissance, aucun travail similaire n'existe pour l'histopathologie numérique, en particulier pour la segmentation des glomérules rénaux.

Deux approches pour mesurer le changement de domaine sont examinées à cet égard : (a) le PixelCNN [126] et (b) la Domain Shift Metric [127]. Les détails méthodologiques pour chaque approche sont les suivants:

PixelCNN [126]: est un modèle génératif conçu pour générer de manière itérative les pixels d'une image. Il apprend la distribution sous-jacente des données de manière non supervisée en quantifiant les pixels d'une image x comme un produit de distributions conditionnelles. En tant que tel, il apprend à prédire la valeur du prochain pixel étant donné (conditionnée par) tous les pixels précédemment générés. Formellement, cela s'exprime comme suit :

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

Ces distributions conditionnelles sont paramétrées par un réseau de neurones con-

volutifs (CNN) et donc partagées à travers tous les pixels de l'image.

Song et al. [128] ont montré qu'un PixelCNN peut être utilisé pour détecter les attaques adversariales dans les images en visualisant les différences dans les distributions de log-vraisemblance des images réelles (propres) et perturbées. Les auteurs ont entraîné un PixelCNN sur un ensemble de données d'images propres pour estimer leur distribution de probabilité sous-jacente. Ce modèle entraîné peut ensuite calculer la log-vraisemblance de toute image donnée, indiquant dans quelle mesure elle s'aligne avec la distribution apprise des images « propres ». À cette fin, les auteurs ont utilisé les bits par dimension (BPD), une mesure normalisée de la log-vraisemblance. Pour une image x avec une résolution de $I \times J$ et K canaux, BPD est défini comme suit :

$$\text{BPD}(x) \triangleq -\log p_{\text{CNN}}(x)/(I \times J \times K \times \log 2), \tag{2}$$

où $p_{\text{CNN}}(x)$ est la probabilité assignée à l'image par le modèle PixelCNN. En utilisant cette formulation, les auteurs ont constaté que les images perturbées affichaient systématiquement des valeurs BPD différentes par rapport aux images propres, entraînant des distributions de log-vraisemblance distinctes.

Nous émettons l'hypothèse qu'une approche similaire peut être utilisée pour détecter le changement de domaine dans les images traduites lors du transfert de teinture. Plus précisément, en utilisant un modèle PixelCNN, nous visons à visualiser les différences dans les distributions de log-vraisemblance entre la teinture PAS réelle et les teintures traduites (Target→PAS). Pour ce faire, un modèle PixelCNN est entraîné sur la teinture PAS réelle pour modéliser sa distribution de données sous-jacente. Une fois entraîné, le PixelCNN peut être appliqué aux teintures traduites Target→PAS pour déterminer si leurs distributions se chevauchent avec celle de la teinture PAS réelle. Nous proposons ensuite d'utiliser la distance de Wasserstein [129] ($\mathcal{W}$) pour quantifier la similarité entre les deux distributions (PAS et Target→PAS). Une valeur $\mathcal{W}$ plus petite indique des distributions plus similaires, fournissant ainsi une mesure fiable du changement de domaine.

Domain Shift Metric: La Domain Shift Metric [127] mesure la différence entre les distributions de deux domaines, appelées ici Scores de Changement de Domaine ou DSS, en utilisant les représentations de caractéristiques d'un modèle pré-entraîné. Considérons un CNN avec des couches $\{l_1, \ldots, l_L\}$. Soit $\Phi(x) = \{\phi_{l1}(x), \ldots, \phi_{lk}(x)\}$ tel que $\Phi_{lk}(x) \in \{\mathbb{R}^{h \times w}\}$ désigne les activations des filtres à la couche l et au filtre k. La valeur moyenne de chaque $\Phi_{lk}(x)$ est calculée comme suit :

$$c_{lk}(x) = \frac{1}{hw} \sum_{i,j}^{h,w} \Phi_{lk}(x)_{i,j}. \tag{3}$$

Soit $p^S_{c_{lk}}(x)$ une distribution de $c_{lk}(x)$ sur la teinture source S et $p^T_{c_{lk}}(x)$ la même sur la teinture traduite (Target→PAS) T, alors la métrique de changement de domaine

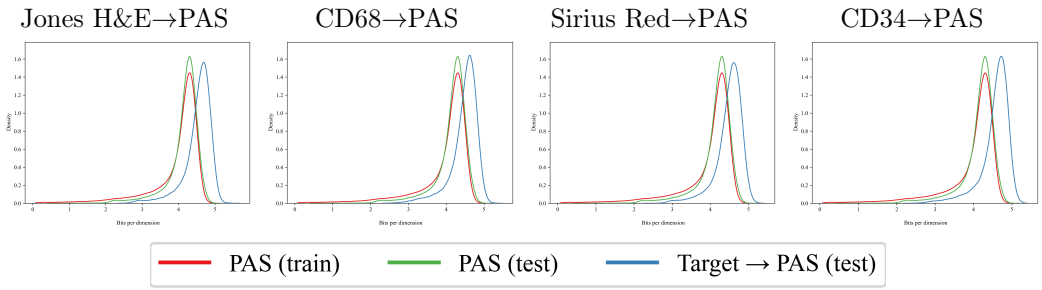


Figure 3: Visualisation basée sur PixelCNN du changement de domaine dans les colorations traduites Target→PAS par rapport aux ensembles de données PAS réels d'entraînement et de test.

(DSM) est définie comme suit :

$$\mathrm{DSM}(p^S, p^T) = \frac{1}{k}\sum_{i=1}^{k} \mathcal{W}\left(p^S_{c_{lk}}, p^T_{c_{lk}}\right), \tag{4}$$

où $\mathcal{W}$ est la distance de Wasserstein [129] entre $p^S_{c_{lk}}(x)$ et $p^T_{c_{lk}}(x)$, qui tend vers zéro lorsque S et T sont similaires.

Résultats

Comme le montre le tableau 1, MDS1 connaît une baisse de performance lorsqu'il est appliqué aux images traduites (Target→PAS) de colorations IHC, telles que CD68 et CD34. En nous basant sur des études récentes [48–50], nous émettons l'hypothèse que cette baisse de performance est causée par un changement de domaine introduit dans les colorations traduites lors du transfert de coloration.

Pour tester cette hypothèse, le modèle PixelCNN entraîné sur PAS est d'abord validé en utilisant les données d'entraînement PAS et un ensemble de test PAS non vu, voir la Figure 3. Il est constaté que leurs distributions de log-vraisemblance suivent le même ordre de grandeur, avec une distance Wasserstein faible de 0.0879 (moyenne sur 5 ensembles de 1000 patches échantillonnés aléatoirement), indiquant un faible changement de domaine. Les distributions de log-vraisemblance des colorations traduites Target→PAS sont également incluses dans cette figure et montrent clairement qu'il y a un changement de domaine par rapport aux distributions PAS d'entraînement/test. Par conséquent, la distance Wasserstein entre PAS d'entraînement et les colorations traduites Target→PAS est observée comme relativement grande, voir le Tableau 2, indiquant un changement de domaine significatif dans les colorations traduites.

En utilisant le modèle de segmentation pré-entraîné pour extraire les représentations de caractéristiques de la coloration source (PAS), le changement de domaine peut également être mesuré en utilisant la métrique de changement de domaine (vue précédemment dans l'Équation (4)). Les DSS respectifs pour toutes les colorations traduites (Target→PAS) sont également inclus dans le Tableau 2.

Table 2: Distance Wasserstein moyenne et scores de changement de domaine de 5 ensembles de 1000 patches échantillonnés aléatoirement pour les colorations traduites Target→PAS ; les écarts types sont entre parenthèses.

Méthodes	Colorations de test				
	PAS	Jones H&E→PAS	Sirius Red→PAS	CD68→PAS	CD34→PAS
Distance Wasserstein ($\mathcal{W}$)	0.087 (0.003)	0.537 (0.012)	0.493 (0.004)	0.481 (0.006)	0.580 (0.005)
Scores de changement de domaine (DSS)	0.032 (0.017)	0.097 (0.008)	0.119 (0.003)	0.248 (0.002)	0.138 (0.002)

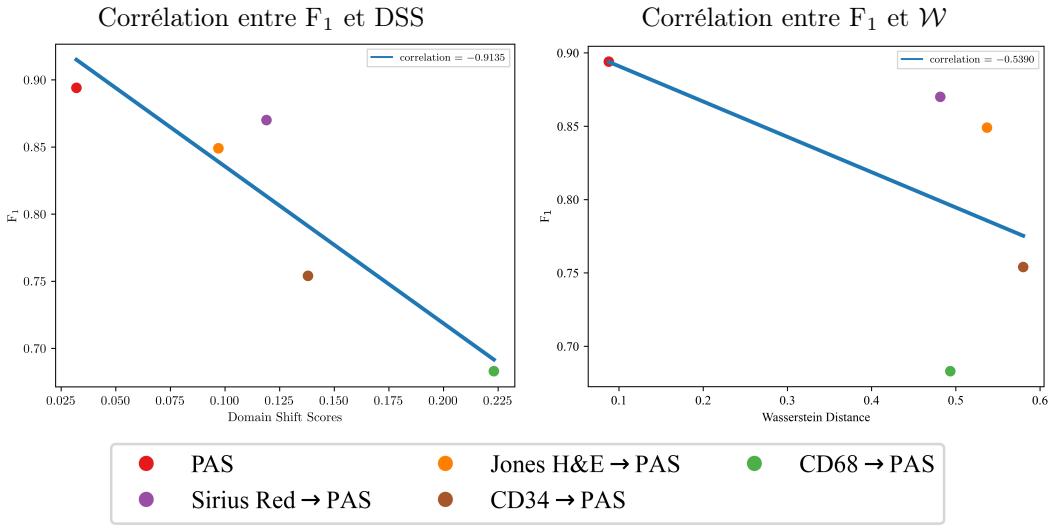


Figure 4: Corrélation entre les scores de segmentation (F_1) des diapositives de test traduites en PAS et le changement de domaine moyen (mesuré en termes de Scores de Changement de Domaine et de Distance Wasserstein) de 5 ensembles de 1000 patches de test échantillonnés aléatoirement.

Maintenant que nous pouvons détecter et mesurer ce qui semble être le changement de domaine, nous enquêtons sur la corrélation avec les scores de segmentation complète (F_1) de MDS1 (fournis dans le Tableau 1: 2^{nd} ligne). La Figure 4 présente les graphiques de dispersion du changement de domaine (mesuré en termes de distance Wasserstein et de DSS) et des scores F_1 basés sur MDS1 pour chaque traduction, révélant une corrélation très forte de -0.9135. Cette forte corrélation négative indique que, à mesure que le changement de domaine (mesuré en termes de DSS) augmente, la performance de segmentation de MDS1 diminue significativement. En revanche, la distance Wasserstein basée sur PixelCNN montre une corrélation modérée (selon les critères spécifiés par [130]) de -0.5390 avec les scores de segmentation complète de MDS1.

Amélioration du Transfert de Taches

Comme indiqué ci-dessus, le transfert de taches basé sur CycleGAN peut entraîner l'ajout de bruit imperceptible dans les taches traduites. Cela conduit à une diminution des performances des méthodes de segmentation multi-taches, telles que MDS1 et UDAGAN, notamment lorsqu'elles sont appliquées à des taches immunohistochimiques traduites. Comme le montre la Figure 4, le DSM est capable de détecter et de mesurer ce bruit de manière efficace. Il reste donc à voir si cette métrique peut être utilisée comme fonction de perte lors de l'entraînement du transfert de taches basé sur CycleGAN. Cette perte, que nous appelons Domain Shift Loss (DSL), peut agir comme une stratégie auto-guidée novatrice pour apprendre des traductions avec un décalage de domaine minimal, améliorant ainsi le transfert de taches et par conséquent, la performance des approches de segmentation multi-taches.

De plus, plusieurs autres développements [49, 49, 77] ont été proposés pour répondre aux limitations du transfert de taches basé sur CycleGAN mentionnées ci-dessus. Bien que ces développements aient montré des améliorations significatives, ils sont principalement conçus pour des applications de vision par ordinateur et n'ont jamais été appliqués aux tâches liées à l'histopathologie (à l'exception de [77]). Par conséquent, dans cette thèse, en plus de notre approche proposée utilisant le DSL, nous proposons d'utiliser les méthodes présentées par Bashkirova et al. [49], Chu et al. [48], et Bouteldja et al. [77]. L'utilisation de ces approches a deux objectifs : d'une part, fournir une analyse comparative avec notre approche proposée utilisant le DSL ; et d'autre part, examiner leur efficacité pour améliorer les approches de segmentation multi-taches, ce qui, à notre connaissance, n'a pas été exploré précédemment.

Les architectures originales, telles que proposées par les auteurs [48, 49, 77], ont été utilisées ici tandis que les détails architecturaux de la Domain Shift Loss sont les suivants :

L'architecture CycleGAN originale est modifiée en intégrant le DSM (défini dans l'Équation (4)) comme perte ($\mathcal{L}_{\text{dsl}}$) pour minimiser l'impact du décalage de domaine dans les images traduites. Le DSM utilise un modèle de segmentation pré-entraîné (entraîné uniquement pour la tache source), donc, dans le cadre bidirectionnel de CycleGAN, il est intégré uniquement dans un sens. En particulier, seules les images de la tache source et les images traduites (target→source) sont fournies au DSM pour calculer la perte ($\mathcal{L}_{\text{dsl}}$). De plus, inspiré par l'idée de Bouteldja et al. [77], $\mathcal{L}_{\text{dsl}}$ est également calculé entre les images de la tache source et leurs reconstructions respectives et le mappage d'identité, tel que

$$\begin{aligned}\mathcal{L}_{\text{dsl}} &= \mathcal{L}_{\text{dsl,translated}} + \mathcal{L}_{\text{dsl,cyc}} + \mathcal{L}_{\text{dsl,id}} \\ &= \mathbb{E}_{s\sim A}\mathbb{E}_{t\sim B}[\text{DSM}(p^s, p^{G_{BA}(t)}) + \text{DSM}(p^s, p^{G_{BA}(G_{AB}(s))}) + \text{DSM}(p^s, p^{G_{BA}(s)})].\end{aligned} \tag{5}$$

Cette modification entraîne la fonction de perte suivante pour CycleGAN :

$$\begin{aligned}\mathcal{L}_{\text{CycleGAN}}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{\text{adv}}(G_{AB}, D_B, G_{BA}, D_A) \\ + w_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) \\ + w_{\text{id}}\mathcal{L}_{\text{id}}(G_{AB}, G_{BA}) \\ + w_{\text{dsl}}\mathcal{L}_{\text{dsl}}(G_{AB}, G_{BA}). \end{aligned} \tag{6}$$

Cette fonction objectif globale guide les images traduites pour qu'elles soient plus proches du domaine source, réduisant ainsi leur décalage de domaine.

Résultats

Dans cette section, les méthodes de transfert de taches basées sur CycleGAN mentionnées précédemment sont évaluées. Ces méthodes sont comparées non seulement à la méthode CycleGAN originale (référence), mais aussi les unes aux autres en utilisant l'approche de segmentation multi-taches MDS1.

Le Tableau 3 présente les résultats pour MDS1 en utilisant chaque méthode de traduction. Les résultats montrent que, pour les taches HC, toutes les méthodes montrent des performances de segmentation similaires ou améliorées par rapport à la méthode de base. Cette amélioration est plus prononcée lors de l'utilisation de 'CycleGAN avec Bruit Gaussien' par rapport aux autres. En revanche, pour les taches IHC, des améliorations de performance sont observées avec toutes les méthodes sauf 'CycleGAN avec Canaux Supplémentaires'. De plus, ces résultats montrent que les gains de performance sont plus substantiels pour les taches IHC, en particulier CD68, par rapport aux taches HC. Cela est dû au fait que la méthode CycleGAN originale éprouve des difficultés avec ces taches car elles sont biologiquement plus distinctes de la tache source (PAS), ce qui introduit plus de bruit dans les images traduites, entraînant une performance de base réduite. Cependant, les méthodes proposées parviennent (dans une certaine mesure) à atténuer cette limitation en réduisant ce bruit des taches traduites. Notamment, la meilleure performance globale (moyenne sur toutes les taches cibles) est obtenue en utilisant à la fois 'CycleGAN avec Bruit Gaussien' et notre 'CycleGAN avec DSL' proposé.

Apprentissage Auto-Supervisé

Les méthodes précédemment proposées ont exploré la limitation inhérente de l'introduction de bruit lors du transfert de coloration et ont introduit différentes stratégies pour minimiser ce bruit, ce qui a abouti à une amélioration des performances des approches de segmentation multi-colorations basées sur le transfert de coloration. Bien que ces méthodes éliminent le besoin de labels dans la coloration cible, il est crucial de reconnaître que ces méthodes dépendent fortement

*Une légère différence est notée dans les résultats MDS1 basés sur CycleGAN présentés dans le Tableau 3 par rapport à ceux présentés dans le Tableau 1. Cela est dû au fait que les expériences du Tableau 3 sont mises en œuvre en utilisant Tensorflow 2, tandis que les expériences du Tableau 1 sont mises en œuvre en utilisant le framework Keras (qui est désormais obsolète et a été intégré dans Tensorflow 2).

Table 3: Scores de segmentation (F_1) basés sur MDS1 pour la segmentation des glomérules à travers diverses taches cibles en utilisant différentes méthodes de transfert de taches basées sur CycleGAN. L'évaluation est réalisée sur un ensemble de données de test indépendant et non vu. Chaque score F_1 est une moyenne de 5 répétitions différentes de UNet, chacune appliquée à 3 répétitions différentes de CycleGAN (15 au total), avec les écarts-types présentés entre parenthèses. Le score (F_1) global le plus élevé (moyenne sur toutes les taches cibles) est indiqué en gras.

Stratégie d'Entraînement	Taches Test				
	Taches HC		Taches IHC		Global
	Jones H&E	Sirius Red	CD68	CD34	
CycleGAN* (référence)	0.844 (0.026)	0.860 (0.023)	0.643 (0.031)	0.747 (0.021)	0.774 (0.025)
avec Bruit Gaussien [49]	0.865 (0.016)	0.878 (0.015)	0.669 (0.026)	0.749 (0.028)	**0.790** (0.021)
avec Auto-supervision [77]	0.840 (0.027)	0.866 (0.021)	0.686 (0.020)	0.753 (0.024)	0.786 (0.021)
avec Canaux Supplémentaires [48, 77]	0.862 (0.019)	0.871 (0.020)	0.634 (0.037)	0.669 (0.041)	0.759 (0.029)
Notre Méthode	0.849 (0.024)	0.862 (0.022)	0.694 (0.021)	0.763 (0.012)	**0.792** (0.020)

d'une grande quantité de données étiquetées provenant de la coloration source. Cependant, l'acquisition d'une quantité suffisante de données étiquetées pour le domaine source reste un défi dans diverses disciplines médicales. Par exemple, en histopathologie, pour certains types de tissus ou de tumeurs, des ensembles de données étiquetées suffisants pour la coloration source peuvent ne pas être facilement disponibles. Néanmoins, les avancées récentes en vision par ordinateur et en imagerie médicale ont conduit à une augmentation significative de la taille des ensembles de données (généralement non étiquetées) de plusieurs ordres de grandeur [42]. Par exemple, en histopathologie, l'avènement des scanners d'imagerie de lames entières (WSI) a facilité la production de vastes quantités de données d'images histopathologiques (non étiquetées). Dans de telles situations, où les données non étiquetées sont disponibles en grande quantité, elles peuvent être utilisées dans des scénarios avec des données étiquetées limitées pour améliorer les performances des modèles grâce à la représentation non supervisée, en particulier l'apprentissage auto-supervisé [87].

L'Apprentissage Auto-Supervisé (AAS) apprend des représentations utiles à partir de données non étiquetées en concevant une tâche prétexte [45], qui peut ensuite être utilisée dans diverses tâches en aval où les données étiquetées sont limitées. En conséquence, cette thèse se concentre sur l'utilisation des méthodes d'apprentissage de représentation auto-supervisées SOTA les plus appropriées, telles que SimCLR [94], BYOL [96], et une nouvelle extension à CS-CO [108]. Ces représentations

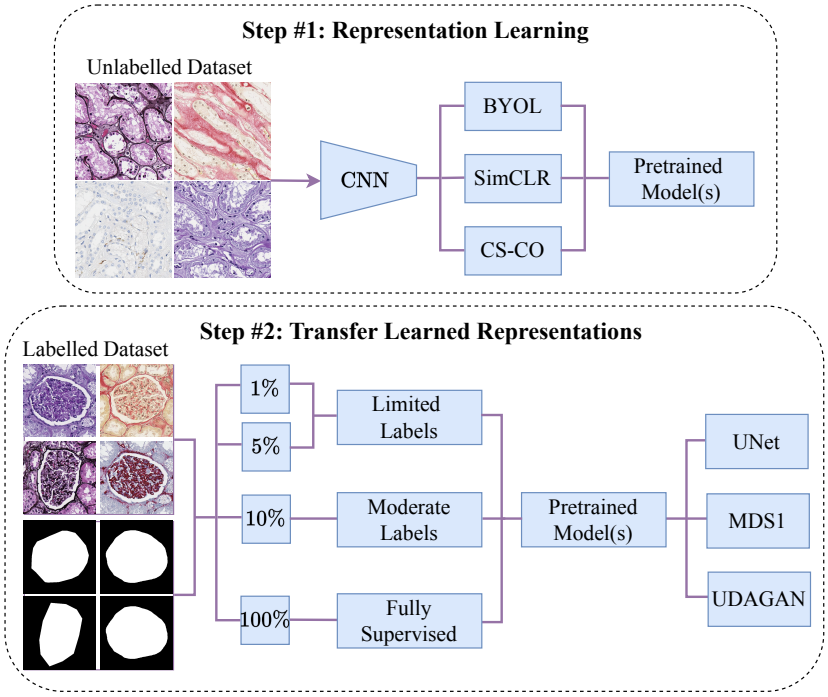


Figure 5: Flux de travail d'apprentissage auto-supervisé en histopathologie. Étape #1 : Différentes méthodes d'apprentissage auto-supervisé sont appliquées pour apprendre des représentations à partir d'un vaste ensemble de données non étiquetées. Étape #2 : Les représentations apprises sont ensuite affinées par un ajustement fin sur plusieurs fractions de données étiquetées pour diverses tâches en aval.

seront ensuite affinées dans plusieurs tâches en aval pour améliorer les performances des approches de segmentation des glomérules rénaux à coloration unique et multi-colorations en présence de données étiquetées limitées, comme illustré dans la Figure 5.

Résultats

Dans cette section, les performances des modèles auto-supervisés pré-entraînés pour l'apprentissage de représentations significatives sont évaluées pour chaque tâche en aval, telles que la segmentation à coloration unique à l'aide de UNet, et la segmentation multi-colorations à l'aide de MDS1 et UDAGAN.

Conformément à [106], plusieurs fractions de l'ensemble de données étiquetées ont été créées, comme le montre la Figure 5. Chaque fraction comprend différents pourcentages de données étiquetées (1%, 5%, 10%, et 100%) provenant des patients d'entraînement de chaque coloration, comme présenté dans le Tableau 4.

Les modèles entièrement supervisés, ou Baseline, ont été entraînés à partir de zéro (c'est-à-dire initialisés aléatoirement) pour établir des baselines avec différentes

Table 4: Données d'entraînement avec différents pourcentages de glomérules étiquetés pour chaque coloration.

% of Labels	Stainings				
	PAS	Jones H&E	CD68	Sirius Red	CD34
1%	6	5	5	6	5
5%	33	31	26	32	28
10%	66	62	52	65	56
100%	662	621	526	651	565

fractions de données, y compris 100% de labels.

En moyenne, dans les cas de marquage limité, correspondant à 5–6 (1%) et 26–33 (5%) glomérules marqués par teinture, les modèles UNet affinés surpassent significativement les modèles UNet de base respectifs (voir la dernière colonne). Cependant, cette supériorité n'est pas uniforme pour toutes les teintures ; notamment, le Sirius Red et le CD34 avec 5% de marquages bénéficient du pré-entraînement, mais pas de manière aussi marquée que pour les autres teintures. Pour certaines teintures, il est possible d'observer que le pré-entraînement avec 100% de marquages peut même surpasser les modèles entièrement supervisés de base, mais ces bénéfices ne sont pas évidents lorsqu'on fait la moyenne sur toutes les teintures. Comme notre objectif est de trouver un niveau de marquage qui minimise l'effort de marquage tout en maximisant les performances, 5% de marquages offrent un bon équilibre entre les deux (10% n'apportant qu'une petite augmentation des performances, tandis que 1% entraîne une baisse considérable). À ce niveau de marquage, une baisse de 11% des performances est observée avec le modèle UNet pré-entraîné avec BYOL par rapport au modèle entièrement supervisé (100%). Cela souligne que le nombre de marquages nécessaires pour l'entraînement peut être réduit de 95%. Si l'apprentissage auto-supervisé n'avait pas été utilisé dans ce cas, une baisse de performance de 26,9% aurait été observée (5e ligne, dernière colonne du tableau 5).

Dans la segmentation multi-teinture MDS1, le même schéma peut être observé. Utiliser 1% et 5% de marquages (mais dans ce cas uniquement à partir de la teinture source, PAS) entraîne une augmentation de performance moyenne considérable par rapport aux modèles de base. En se concentrant sur 5% de marquages, le pré-entraînement SimCLR permet à MDS1 d'atteindre un score F_1 moyen de $0,707$, soit seulement 8,2% de moins que le modèle de référence MDS1 entièrement supervisé $(0,789)$, tout en réduisant l'exigence de marquage de 95%. De plus, cela n'est que 5% inférieur à la meilleure performance moyenne du modèle UNet à teinture unique avec pré-entraînement, qui nécessite des marquages pour toutes les teintures, alors que MDS1 ne les nécessite que pour la teinture source.

Cette tendance se poursuit dans les résultats du modèle UDAGAN invariant aux teintures, où en moyenne, le pré-entraînement et l'affinage avec 1% et 5% de marquages (encore une fois, uniquement pour la teinture source) surpassent considérablement les modèles de base pour toutes les teintures. Le pré-entraînement HR-CS-CO n'est pas évalué car UDAGAN est un modèle unique de segmentation

Table 5: Une comparaison de diverses méthodes d'auto-apprentissage en pré-entraînement et des lignes de base respectives (initialisées aléatoirement sans aucun pré-entraînement) pour les tâches aval de UNet, MDS1 et UDAGAN utilisant différentes divisions de données annotées. Pour UNet, les annotations ont été utilisées pour toutes les colorations, tandis que pour MDS1 et UDAGAN, seules les annotations de la coloration source (PAS) ont été utilisées. L'évaluation est effectuée sur un jeu de données de test indépendant et non vu, en utilisant le score F1. Chaque score F1 est la moyenne de cinq répétitions d'entraînement différentes (les écarts-types sont entre parenthèses). Le score F1 le plus élevé pour chaque coloration, à travers différentes divisions des annotations, est en italique, tandis que le score F1 le plus élevé global, moyenné sur toutes les colorations, est en gras.

Downstream Tasks	Label Splits	Pre-training	Test Stains					Average
			PAS	Jones H&E	CD68	Sirius Red	CD34	
UNet	1%	None (Baseline)	0.015 (0.031)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.253 (0.059)	0.054 (0.018)
		SimCLR	*0.673 (0.021)*	0.519 (0.040)	0.407 (0.015)	0.472 (0.037)	0.652 (0.018)	0.544 (0.026)
		BYOL	0.660 (0.018)	*0.635 (0.055)*	*0.625 (0.042)*	*0.561 (0.044)*	*0.686 (0.030)*	**0.633 (0.038)**
		HR-CS-CO	0.154 (0.044)	0.188 (0.067)	0.048 (0.083)	0.337 (0.082)	0.463 (0.017)	0.238 (0.058)
	5%	None (Baseline)	0.546 (0.084)	0.593 (0.080)	0.370 (0.188)	0.707 (0.055)	0.782 (0.041)	0.600 (0.090)
		SimCLR	*0.852 (0.019)*	*0.760 (0.017)*	0.599 (0.039)	0.618 (0.042)	*0.802 (0.011)*	0.726 (0.026)
		BYOL	0.768 (0.036)	0.746 (0.076)	*0.736 (0.033)*	*0.721 (0.051)*	*0.800 (0.047)*	**0.754 (0.049)**
		HR-CS-CO	0.756 (0.079)	0.628 (0.086)	0.533 (0.067)	0.406 (0.067)	0.707 (0.037)	0.606 (0.067)
	10%	None (Baseline)	0.730 (0.017)	0.792 (0.024)	0.643 (0.053)	*0.788 (0.022)*	0.827 (0.063)	0.756 (0.036)
		SimCLR	*0.867 (0.019)*	0.813 (0.012)	0.690 (0.057)	0.696 (0.060)	*0.838 (0.007)*	**0.781 (0.031)**
		BYOL	0.794 (0.047)	*0.823 (0.054)*	*0.729 (0.052)*	0.722 (0.044)	0.776 (0.057)	0.769 (0.051)
		HR-CS-CO	0.807 (0.058)	0.748 (0.098)	*0.729 (0.040)*	0.711 (0.074)	0.791 (0.026)	0.757 (0.059)
	100%	None (Baseline)	*0.894 (0.021)*	0.840 (0.029)	0.836 (0.031)	0.865 (0.019)	*0.888 (0.015)*	0.865 (0.024)
		SimCLR	0.884 (0.003)	*0.873 (0.007)*	0.840 (0.011)	*0.881 (0.007)*	0.867 (0.027)	**0.869 (0.011)**
		BYOL	0.867 (0.009)	0.842 (0.035)	0.818 (0.036)	0.847 (0.012)	0.874 (0.021)	0.850 (0.022)
		HR-CS-CO	0.843 (0.033)	0.855 (0.015)	*0.872 (0.006)*	0.842 (0.023)	0.870 (0.011)	0.856 (0.018)
MDS1	1%	None (Baseline)	—	0.030 (0.066)	0.024 (0.054)	0.039 (0.086)	0.036 (0.079)	0.032 (0.071)
		SimCLR		*0.615 (0.015)*	*0.403 (0.031)*	*0.594 (0.026)*	*0.614 (0.028)*	**0.556 (0.025)**
		BYOL		0.516 (0.041)	0.363 (0.027)	0.525 (0.047)	0.494 (0.031)	0.474 (0.037)
		HR-CS-CO		0.326 (0.025)	0.224 (0.045)	0.359 (0.050)	0.384 (0.035)	0.323 (0.039)
	5%	None (Baseline)	—	0.711 (0.032)	0.526 (0.041)	0.685 (0.031)	0.613 (0.050)	0.634 (0.038)
		SimCLR		*0.798 (0.005)*	0.534 (0.015)	0.767 (0.008)	*0.729 (0.016)*	**0.707 (0.011)**
		BYOL		0.713 (0.051)	*0.538 (0.047)*	0.733 (0.032)	0.605 (0.061)	0.647 (0.048)
		HR-CS-CO		0.760 (0.028)	0.335 (0.084)	*0.773 (0.015)*	0.607 (0.044)	0.619 (0.043)
	10%	None (Baseline)	—	0.776 (0.017)	*0.575 (0.025)*	0.778 (0.023)	0.656 (0.030)	0.696 (0.024)
		SimCLR		*0.784 (0.026)*	0.541 (0.029)	0.752 (0.040)	*0.722 (0.016)*	**0.700 (0.028)**
		BYOL		0.706 (0.063)	0.541 (0.060)	0.731 (0.084)	0.650 (0.043)	0.657 (0.062)
		HR-CS-CO		0.771 (0.037)	0.433 (0.059)	*0.804 (0.041)*	0.633 (0.033)	0.660 (0.042)
	100%	None (Baseline)	—	0.849 (0.017)	*0.683 (0.043)*	0.870 (0.009)	*0.754 (0.008)*	**0.789 (0.032)**
		SimCLR		0.826 (0.033)	0.638 (0.056)	0.836 (0.034)	0.712 (0.030)	0.753 (0.038)
		BYOL		0.833 (0.032)	0.632 (0.042)	0.864 (0.028)	0.652 (0.066)	0.745 (0.042)
		HR-CS-CO		*0.863 (0.017)*	0.614 (0.067)	*0.878 (0.018)*	0.730 (0.040)	0.771 (0.036)
UDAGAN	1%	None (Baseline)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SimCLR	0.477 (0.015)	0.403 (0.025)	0.261 (0.053)	0.408 (0.010)	0.518 (0.016)	0.413 (0.024)
		BYOL	*0.647 (0.062)*	*0.504 (0.083)*	*0.401 (0.099)*	*0.513 (0.088)*	*0.598 (0.064)*	**0.533 (0.079)**
	5%	None (Baseline)	0.669 (0.038)	0.498 (0.056)	0.352 (0.056)	0.618 (0.072)	0.692 (0.024)	0.566 (0.049)
		SimCLR	0.719 (0.018)	0.616 (0.020)	0.524 (0.014)	0.632 (0.015)	0.716 (0.015)	0.641 (0.016)
		BYOL	*0.815 (0.027)*	*0.730 (0.071)*	*0.603 (0.028)*	*0.732 (0.028)*	*0.726 (0.055)*	**0.721 (0.042)**
	10%	None (Baseline)	0.816 (0.031)	0.687 (0.014)	0.614 (0.019)	0.750 (0.069)	0.770 (0.022)	0.727 (0.031)
		SimCLR	0.781 (0.013)	0.712 (0.013)	0.606 (0.015)	0.706 (0.026)	0.768 (0.012)	0.715 (0.016)
		BYOL	*0.834 (0.035)*	*0.767 (0.051)*	*0.654 (0.040)*	*0.742 (0.090)*	*0.781 (0.037)*	**0.755 (0.051)**
	100%	None (Baseline)	*0.901 (0.011)*	0.856 (0.036)	0.705 (0.031)	0.873 (0.025)	0.799 (0.035)	0.827 (0.027)
		SimCLR	0.892 (0.008)	*0.866 (0.018)*	*0.777 (0.013)*	*0.888 (0.015)*	*0.844 (0.003)*	**0.853 (0.011)**
		BYOL	0.883 (0.019)	0.854 (0.039)	0.722 (0.051)	0.818 (0.068)	0.792 (0.036)	0.814 (0.042)

Table 6: Performance des tâches en aval avec 5% d'étiquettes d'entraînement, sans ensemble de validation. UNet, 5% d'étiquettes sont utilisées pour toutes les teintures, MDS1 et UDAGAN, 5% d'étiquettes sont utilisées uniquement pour la source, PAS, teinture. L'évaluation est réalisée sur l'ensemble de test. Chaque score F1 est la moyenne de cinq répétitions d'entraînement différentes (écarts types entre parenthèses). Le score F1 le plus élevé pour chaque teinture est en italique, et le score F_1 le plus élevé en général, moyenné sur toutes les teintures, est en gras.

Downstream Tasks	Pre-training	Test Stains					Average
		PAS	Jones H&E	CD68	Sirius Red	CD34	
UNet	SimCLR	*0.812 (0.019)*	0.795 (0.034)	0.575 (0.146)	0.612 (0.066)	0.810 (0.020)	0.720 (0.057)
	BYOL	0.786 (0.020)	*0.839 (0.025)*	*0.771 (0.027)*	*0.788 (0.021)*	*0.870 (0.003)*	**0.810 (0.019)**
	HR-CS-CO	0.777 (0.032)	0.695 (0.092)	0.428 (0.086)	0.425 (0.094)	0.700 (0.060)	0.605 (0.072)
MDS1	SimCLR	—	0.787 (0.016)	0.608 (0.015)	0.770 (0.021)	*0.704 (0.022)*	0.717 (0.018)
	BYOL		*0.813 (0.037)*	*0.646 (0.038)*	*0.823 (0.037)*	0.695 (0.038)	**0.744 (0.037)**
	HR-CS-CO		0.776 (0.013)	0.251 (0.051)	0.812 (0.007)	0.599 (0.026)	0.609 (0.024)
UDAGAN	SimCLR	0.402 (0.193)	0.389 (0.078)	0.000 (0.000)	0.072 (0.120)	0.359 (0.260)	0.244 (0.130)
	BYOL	*0.850 (0.008)*	*0.822 (0.021)*	*0.650 (0.029)*	*0.815 (0.026)*	*0.771 (0.011)*	**0.765 (0.022)**

multi-teinture invariant aux teintures, tandis que HR-CS-CO est entraîné séparément pour chaque teinture. Dans ce cas, nous observons une baisse de 10,6% des performances lors de l'affinage avec 5% de marquages (et un pré-entraînement avec BYOL) par rapport aux modèles entièrement supervisés (100%). Si le modèle avait été entraîné de manière entièrement supervisée avec cette quantité de marquages, une baisse de 26,1% aurait été observée, ce qui montre que l'affinage permet de minimiser l'impact du manque de marquages.

Omission des données de validation

Comme montré ci-dessus, un équilibre entre la minimisation des étiquettes et la maximisation des performances est atteint en utilisant 5% d'étiquettes. Néanmoins, lors de l'entraînement des modèles finaux, les résultats ont été obtenus en utilisant un ensemble de validation entièrement étiqueté. Par conséquent, le tableau 6 évalue si l'ensemble de validation est nécessaire ou si cette exigence d'étiquetage peut également être réduite. Il est démontré que, dans de nombreux cas, les performances sans ensemble de validation surpassent celles obtenues en utilisant un ensemble de validation étiqueté. Cela s'explique par le fait que, dans le jeu de données utilisé, il y a un moindre décalage de domaine (mesuré selon [52]) entre les distributions des ensembles d'entraînement et de test, qui est de 0.0655 (moyenne pour toutes les colorations), comparé aux distributions des ensembles d'entraînement et de validation, qui est de 0.1857. Cela permet aux modèles entraînés sans données de validation de surpasser (sur les données de test) ceux sélectionnés en utilisant la perte de validation. Bien que ce comportement soit spécifique aux ensembles de données avec cette caractéristique mentionnée, il n'affecte que la différence de performance entre les deux paramètres expérimentaux et non les conclusions elles-mêmes. Imaginons qu'il y ait eu un décalage de domaine plus faible entre les ensembles de validation et d'entraînement, dans ce cas, la suppression de l'ensemble de validation n'aurait

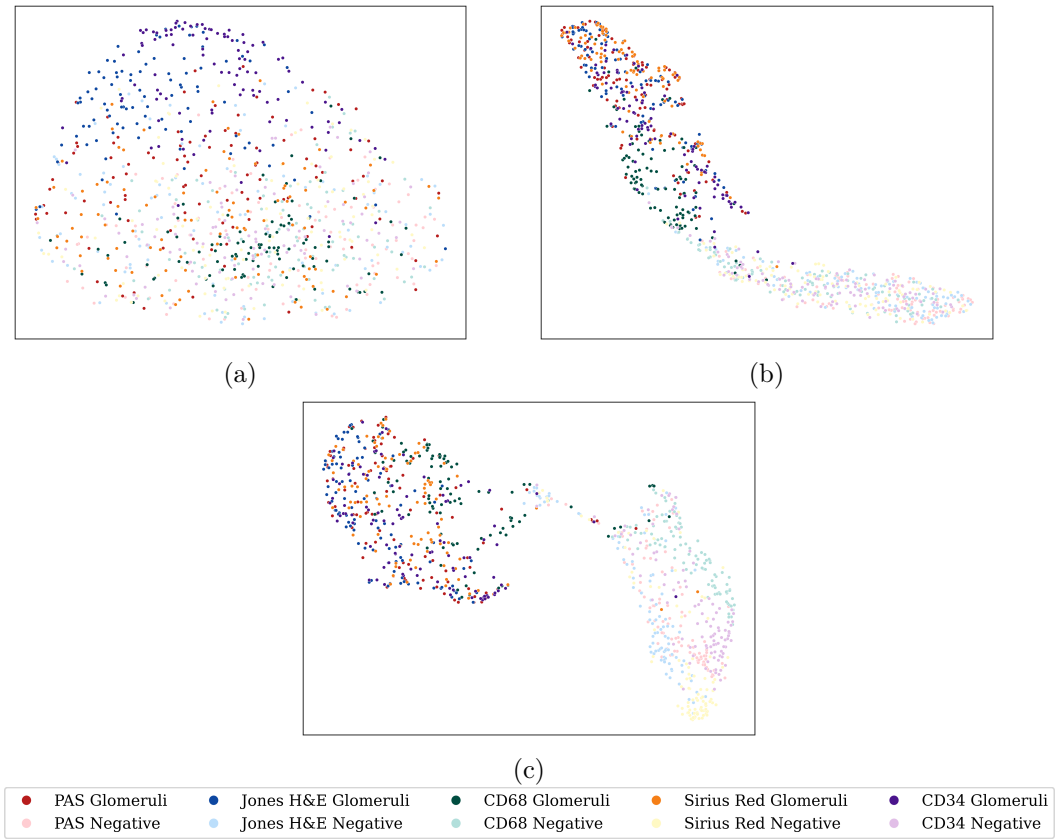


Figure 6: Emeddings UMAP en deux dimensions des représentations apprises par : (a) les modèles SimCLR et (b) BYOL basés sur UDAGAN, entraînés sans ensemble de validation, et (c) SimCLR UDAGAN avec un ensemble de validation étiqueté à 5%. Modèles choisis au hasard, représentations échantillonnées à partir de la couche convolutionnelle avant-dernière, 100 patchs par tache par classe à partir de l'ensemble de test non vu. Chaque point est un patch de la classe et de la coloration respectives

fait que supprimer l'augmentation de performance observée ici. Cela n'invalide donc pas les conclusions présentées ici, selon lesquelles l'ensemble de validation peut être supprimé pour réduire encore les besoins en étiquetage.

Cependant, avec SimCLR et UDAGAN, une chute considérable des performances est observée. Cela est probablement dû à un surapprentissage en l'absence de données de validation. Le modèle est entraîné en deux étapes : (1) pré-entraînement à l'aide de SimCLR sur des morceaux d'images d'origine ; (2) traduction (utilisant des modèles CycleGAN) de PAS vers toutes les autres teintures lors de l'ajustement fin. Au cours de la deuxième étape, un bruit imperceptible causé par le transfert CycleGAN [52] est introduit dans les morceaux d'entraînement. Cela provoque un décalage de domaine entre les données d'entraînement et les images de test, réduisant ainsi les performances. Cela est exacerbé par l'absence d'un ensemble de validation, qui empêcherait normalement le surapprentissage de ces données d'entraînement

“bruyantes”. En revanche, BYOL n’est pas affecté car il utilise la normalisation par lots, ce qui aide à stabiliser le processus d’entraînement et à prévenir le surapprentissage des entrées bruyantes. Cela peut être visualisé dans la Fig. 6, où il y a un manque notable de frontières de classe entre les glomérules de test et les morceaux négatifs lors de l’entraînement de SimCLR-UDAGAN sans ensemble de validation, voir Fig. 6(a). Une telle frontière existe dans la représentation BYOL-UDAGAN sans données de validation, voir Fig. 6(b), et dans un SimCLR-UDAGAN entraîné avec 5% d’étiquettes de validation, voir Fig. 6(c) (pour comparaison, ce modèle obtient un score F_1 moyen de 0.686, contre 0.244 sans l’ensemble de validation).

Perspectives

Le travail présenté dans cette thèse ouvre plusieurs axes de recherche possibles, dont certains concernent l’amélioration directe des méthodes proposées, tandis que d’autres impliquent l’application de la méthodologie développée à des domaines différents mais connexes.

La thèse a démontré qu’il est possible de segmenter les glomérules à travers plusieurs colorations en utilisant seulement quelques étiquettes d’une seule coloration. Pour étayer davantage ces résultats, des explorations supplémentaires devraient être menées pour confirmer que les mêmes approches peuvent être utilisées pour classifier, détecter ou segmenter d’autres structures diagnostiquement pertinentes en histopathologie numérique (par exemple, les tubules, etc.), indépendamment de la modalité de coloration. Une application réussie exige que les structures cibles maintiennent une morphologie cohérente à travers différentes colorations, même si les informations texturales et de couleur varient. Cette approche pourrait également s’étendre à d’autres modalités d’imagerie médicale, telles que l’IRM ou les scanners CT, où les structures anatomiques d’intérêt conservent leur apparence générale malgré les différentes techniques d’imagerie, ainsi qu’à des problèmes plus larges de vision par ordinateur où les objets gardent leur apparence générale malgré les changements de contexte, de texture ou de couleur.

Les avancées proposées dans l’architecture CycleGAN ont amélioré sa robustesse et ses capacités de généralisation lorsqu’elle est utilisée pour des modèles de segmentation multi-coloration basés sur le transfert de coloration. Cela ouvre la voie à l’exploration de plusieurs autres méthodes d’apprentissage profond à la pointe de la technologie qui reposent sur des principes similaires. Récemment, des modèles d’apprentissage profond basés sur la diffusion ont été introduits pour générer des images de haute qualité [168] et attirent une attention significative dans les tâches d’histopathologie numérique, notamment pour le transfert de coloration [169, 170]. En conséquence, l’un de nos axes de recherche futurs est d’explorer les modèles de diffusion pour améliorer encore l’efficacité des méthodes de segmentation multi-coloration.

Malgré les progrès réalisés par les méthodes d’apprentissage auto-supervisé (SSL), notre travail reconnaît certains défis. Notamment, lors de l’entraînement de SimCLR sur des ensembles de données avec une diversité de classes limitée, il y a un risque accru de générer de faux négatifs. De plus, les augmentations utilisées dans les méth-

odes SSL basées sur la contrastivité sont spécifiquement conçues pour les images naturelles et les images médicales sont très sensibles à ces augmentations [123, 167]. Pour surmonter ces défis, nos travaux futurs se concentreront sur la modélisation d'images masquées basée sur des transformateurs (MIM) [171, 172]. Cette approche représente une approche plus robuste de l'apprentissage auto-supervisé car le MIM vise à apprendre des représentations en générant les parties manquantes d'une image, obligeant ainsi le modèle à apprendre les relations entre les différents éléments d'une image. Notamment, ces méthodes n'ont pas besoin de étapes d'augmentation supplémentaires et ont démontré une grande évolutivité et robustesse.