Deep Learning for Blood Cell Image Analysis

Agnieszka Tomczak

TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Computation, Information and Technology

Deep learning for blood cell image analysis

Agnieszka Maria Tomczak

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz::		Prof. Dr. Julien Gagneur
Prüfer*innen der Dissertation:		
	1.	Prof. Dr. Nassir Navab
	2.	Prof. Dr. Nasir Rajpoot
	3.	Prof. Dr. Shadi Albarqouni

Die Dissertation wurde am 22.02.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 17.07.2023 angenommen.

Abstract

The chemical staining of samples in haematology is one of the crucial components of blood analysis. The dye components, such as methylene azure and eosin, allow the colouring of the white blood cell, enhancing structures otherwise invisible to the human eye. After staining, the cytoplasm has a pink to violet shade, the granules visibly differing in size and shape. The nucleus is suddenly pronounced, and the nucleoli are visible inside it. All these features allow the haematologist to classify the white blood cell, leading to diagnosing infections and diseases as severe as leukaemia or lymphoma. However, what would happen if we could replace the chemical process with a digital one? Instead of a time- and chemicals-consuming process, an unstained sample could be captured and artificially stained. It would reduce the effort and waste put into the process of chemical staining on daily bases. This thesis investigates if such change is possible.

The images of unstained white blood cells were captured with a Differential Inference Contrast microscope and stained with multiple deep learning-based techniques. After the same samples were chemically stained and captured with a traditional Bright-field microscope, we could evaluate the algorithms and the staining they produced against the respective chemically stained cells. This work documents the creation of multiple algorithms for artificial staining of both paired and unpaired data sets, the tools for estimating uncertainty connected with such process, and at the end, clinical validation and feasibility of eliminating the chemical staining of blood cells from laboratory pipeline.

We discover the dependency between image generation, classification and segmentation tasks, showing how the segmentation influences structure preservation and classification enhances class-relevant features during the image translation in unpaired datasets. We also show that having even partially paired images in the presence of multiple domains significantly improves the image quality of the generated images and classification results. Furthermore, we investigate the style-structure disentanglement in haematological images based on pseudo-segmentation masks and show its effectiveness. We use the disentangled structure and style representations to present the possible ways of calculating the confidence of the generated image by exploring its latent representation. We show that such a confidence score correlates with the quality of the generated image. Additionally, we analyse the confidence score in the context of the downstream segmentation task on the target domain, showing that when the segmentation networks perform well, the segmentation results correlate with the quality of latent representation in the generative model. Lastly, we present a study with haematology experts, showing that paired dataset with additional class information is a promising approach to introducing artificial staining in the clinical set-up.

This thesis investigates whether we can use deep learning to discover new possibilities in analysing the physical world. Is chemical staining necessary to correctly identify the blood cells? Or could the automated networks analyse the information seemingly invisible to the human eye and discover overlooked patterns?

Zusammenfassung

Die chemische Färbung von Proben in der Hämatologie ist eine der wichtigsten Komponenten der Blutanalyse. Die Farbstoffkomponenten, wie Methylenazer und Eosin, ermöglichen die Färbung der weißen Blutkörperchen und heben Strukturen hervor, die für das menschliche Auge sonst unsichtbar sind. Nach der Färbung hat das Zytoplasma einen rosa bis violetten Farbton, die Körnchen unterscheiden sich sichtbar in Größe und Form. Der Zellkern ist plötzlich ausgeprägt, und die Nukleoli sind in seinem Inneren sichtbar. All diese Merkmale ermöglichen es dem Hämatologen, die weißen Blutkörperchen zu klassifizieren, was zu einer Diagnose von Infektionen und so schweren Krankheiten wie Leukämie oder Lymphomen führt. Doch was wäre, wenn wir den chemischen Prozess durch einen digitalen ersetzen könnten? Anstelle eines zeit- und chemieaufwändigen Prozesses würde eine ungefärbte Probe erfasst und künstlich angefärbt. Dies würde den Aufwand und die Verschwendung verringern, die mit der täglichen chemischen Färbung verbunden sind. In dieser Arbeit wird untersucht, ob eine solche Änderung möglich ist.

Die Bilder von ungefärbten weißen Blutkörperchen wurden mit einem Differentialinferenzkontrastmikroskop aufgenommen und dann mit mehreren auf Deep Learning basierenden Verfahren gefärbt. Nachdem dieselben Proben mit einem herkömmlichen Hellfeldmikroskop gefärbt und aufgenommen wurden, konnten wir die Algorithmen und die von ihnen erzeugten Färbungen anhand der jeweiligen chemisch gefärbten Zellen bewerten. Diese Arbeit dokumentiert die Entwicklung mehrerer Algorithmen für die künstliche Färbung sowohl gepaarter als auch ungepaarter Datensätze, die Werkzeuge zur Abschätzung der mit diesem Prozess verbundenen Unsicherheit und schließlich die klinische Validierung und Machbarkeit der Eliminierung der chemischen Färbung von Blutzellen aus der Laborpipeline.

Wir entdecken die Abhängigkeit zwischen Bilderzeugung, Klassifizierung und Segmentierungsaufgaben und zeigen, wie die Segmentierung die Strukturerhaltung beeinflusst und die Klassifizierung die klassenrelevanten Merkmale während des Bildübersetzungsprozesses in ungepaarten Datensätzen verbessert. Wir zeigen auch, dass selbst teilweise gepaarte Bilder bei Vorhandensein mehrerer Domänen die Bildqualität der generierten Bilder und die Klassifizierungsergebnisse deutlich verbessern. Darüber hinaus haben wir die Stil-Struktur-Entflechtung in hämatologischen Bildern auf der Grundlage von Pseudo-Segmentierungsmasken untersucht und ihre Effektivität gezeigt. Wir verwenden die entwirrten Struktur- und Stilrepräsentationen, um die möglichen Wege zur Berechnung des Vertrauens des generierten Bildes durch die Erforschung seiner latenten Repräsentation aufzuzeigen. Darüber hinaus analysieren wir den Konfidenzwert im Kontext der nachgelagerten Segmentierungsaufgabe auf der Zieldomäne und zeigen, dass bei guter Leistung der Segmentierungsnetzwerke die Segmentierungsergebnisse mit der Qualität der latenten Repräsentation im generativen Modell korrelieren.

Schließlich stellen wir eine Studie mit Hämatologieexperten vor, die zeigt, dass gepaarte Datensätze mit zusätzlichen Klasseninformationen ein vielversprechender Ansatz für die Einführung künstlicher Färbungen in der klinischen Praxis sind.

In dieser Arbeit wird untersucht, ob wir mithilfe von Deep Learning neue Möglichkeiten für die Analyse der physischen Welt entdecken können. Ist eine chemische Färbung notwendig, um die Blutzellen richtig zu identifizieren? Oder könnten die automatisierten Netzwerke die Informationen analysieren, die für das menschliche Auge unsichtbar sind, und übersehene Muster entdecken?

Acknowledgments

First, I would like to thank my PhD supervisor Prof. Dr. Nassir Navab, for giving me the opportunity to pursue the PhD and including me in the vibrant CAMP chair community. I'd also like to thank Prof. Dr. Shadi Albarqouni for guiding and inspiring me through all these years, not letting my excitement about research fade and routinely motivating me with discussions and questions. Thank you very much for introducing me to deep learning for medical applications and supporting me through the process of becoming a scientist. I would also like to wholeheartedly thank my mentor and boss, PD Dr. Slobodan Ilic, for his advice and patience, always helping me when I needed support and, especially, for all the runs and bike rides. Thank you for making this time so much fun.

Big thank you to Dr. Claudio Laloni for supporting this PhD and for the opportunity to work at Siemens AG. I would also like to thank everyone on the Blood cells project: Dr. Gaby Marquart, Dr. Thomas Engel, Dr. Frank Forster, Dr. Laura Boldu, and Jens Brock for their cooperation and support, and all the Siemens colleagues: Peter, Andres, Lucas, Claudia, Eik, Adrian and Carsten, for funny chats, lunches and help with the formalities and infrastructure. Also, thank you to those who used to be senior PhD students when I joined and now are already successful doctors: Sergey, Haowen and Mai, for advice and support at the beginning of this adventure. Huge thank you to my office mates: Fabian, Ivan and Roman, for all the coffee breaks, the food and drinks and for being you.

I especially want to thank Anne-Marie Rickmann for the countless discussions, morning walks during the lockdown, giggles, outings, and, most importantly, teaching me how to bravely try new ideas and endlessly inspiring me with unlimited curiosity and perseverance. It was amazing to share this weird journey with you.

I want to thank my friends and flatmates Izabela 'you can do it!' and Paula 'go girl!' for keeping me together through quarantines, rejection letters and acceptance letters, for enjoying many hikes, walks, cooking experiments and lots of cake. I'd also like to thank Nanda for being an inspiration, for her advice, support and the delicious brownies. Thank you to Dino for the support and motivation. And thanks to Tobi for the final kick.

And to my parents, thank you for all the love, support, advice and acceptance, to my brother Adam for helping me to gain perspective, and to my brother Marcin for motivating and grounding me at the same time, and most importantly, for telling me to go and study computer science in the first place.
Love,

– Agnieszka

Contents

List of Figures

List of Tables

Part I

Introduction and Background

1 Introduction

A blood smear examination is one of the most common procedures during a medical check-up. It is fundamental in diagnosing a plethora of disorders and diseases. It consists of counting and identifying red and white blood cells in the peripheral blood smear. The number of different kinds of white blood cells is a vital determining factor indicating our condition and allowing for a potential diagnosis.

One of the first steps to complete the blood cell examination is the chemical staining of the blood smear. Haematological laboratories stain blood samples with chemical agents to enhance specific white blood cell structures not easily visible by the human eye. Replacing this procedure with a digital one would substantially facilitate the diagnostic process.

Since differential inference contrast microscopy can capture images of unstained blood cells, it opens the possibility of replacing the chemical staining process with a digital one. Such change would be a potential breakthrough for haematology laboratories, as chemical staining is expensive and time-consuming. Additionally, once a sufficiently extensive common database was created, it would help to eliminate the staining artefacts and to unify staining protocols between laboratories. The current developments in the field of artificial intelligence offer multiple tools that can be employed to automatically generate an image of stained blood cells.

This thesis presents a building stone in this direction, investigating whether a deep-learning-based model can approximate a complicated chemical process well enough to be used in a clinical setup.

1.1 Objectives

Given the motivation outlined, my objective was to digitally stain a leukocyte (white blood cell). In other words, to automatically transform an image of an unstained blood cell into its stained version. As simple as it sounds, there are certain challenges connected to this objective: the class characteristics must be preserved and generated during the staining process, and for the comfort of haematologists working with the images, the artificially stained cells must be indistinguishable from chemically stained ones.

Considering the requirements, my first task is to develop a method for generating a stained image conditioned on its unstained counterpart and to investigate the applicability of Generative Adversarial Networks [39] both for paired and unpaired datasets in this domain. Additionally, I explore using the class information during the generation process and employing the segmentation masks to preserve the blood cell structure, as cycle-consistency loss does not constrain it explicitly.

In the dissertation's second part, we focus on estimating the network's confidence during the generation process. In clinical practice, the usage of neural networks is often corrupted by the lack of confidence evaluation, which would allow reason about the trustworthiness of the prediction. To address this concern, we assume that a more stable latent representation leads to a better quality of a generated image. Additionally, we investigate the relationship between such confidence measure and the downstream segmentation task.

The final objective was to present the study of the applicability of Generative Adversarial Networks for this problem. To this end, together with the clinical partners, we conducted a validation study on artificially stained images that indicated the possibility of using such a pipeline in clinical practice.

To sum up, the objectives of this work are the following:

- digitally stain a leukocyte (white blood cell), preserving its structure and key characteristics;
- estimate confidence of the network generating a given image;
- validate the applicability of the new staining process during a study with clinical partners.

1.2 Contributions

To fulfil the objectives listed in the previous section, we develop several methods that have the following contributions:

- **Multi-task, multi-domain framework for digital staining with structure preservation for unpaired data.** I propose a novel method to simultaneously generate the staining on unstained images captured with a differential inference contrast microscope and classify the images of white blood cells according to the types defined by haematologists. The model builds on an intuition that features necessary for staining and classifying can benefit from each other once the gap between the domains of the input data is closed. We analyse how auxiliary tasks such as segmentation and pair-wise reconstruction influence the quality of generated images in terms of fine-grained details and well-defined structures. Furthermore, we construct domain-agnostic latent space so that the features extracted in an unsupervised manner can be used for the downstream classification task.

- **A style-structure disentanglement algorithm for unpaired data.** I present a disentanglement technique for the style and structure of haematological cell images by having multiple encoders constrained differently during the training phase. We take advantage of the fact that the structure can be easily defined in the images of white blood cells by considering the shape of the nucleus and cytoplasm. Therefore, we can use pseudo-segmentation masks of the cells and their intracellular structures to guide the network to generate the desired structures. Moreover, we can do it independently of generating the style. We use these separate representations to analyse the generated image components.

- **Confidence estimation method for disentangled representations and full images.** We estimate the confidence of the latent representation to recognize poorly generated samples, both in terms of style and structure, as in terms of the whole image. We present a noise injection technique that allows generating multiple outputs Quantifying the differences between these outputs provides a confidence score that can be used to determine the uncertain parts of the generated image, the quality of the generated sample, and some extent, result on a downstream task. We investigate this technique, corrupting the disentangled latent representation of structure and style and the whole image's latent representation.

- **A validation study with medical experts confirming the possibility of practical use of GANs in a clinical setup.** We develop and evaluate a GAN-based model to generate realistic-looking stained images by preserving morphological cell features. This model allows the recognition of normal blood cells and the identification of severe haematological diseases. We conduct a detailed qualitative study of the samples with clinical experts. We show that humans and convolutional neural networks can correctly classify most artificially stained samples.

1.3 Outline

This section provides a brief overview of each of the subsequent chapters. Most of the methods and materials of this thesis are published or are under submission for a major conference or journal. Therefore, we provide the work related to each chapter and encourage the interested reader to consult the online material for presented methods.

Chapter 2 We first provide the theoretical foundation upon which this thesis is built. In particular, we outline the theoretical deep learning background focusing on image-to-image translation with Generative Adversarial Networks. In addition, we describe the imaging modalities used to collect data for this work.

Chapter 3 This chapter presents the process of chemical staining that this thesis aims to substitute. After summarising the procedure, we list the considered leukocytes together with their characteristics, challenges and corresponding graphical representations. After this overview, I comment on the clinical relevance of the white blood cell classification.

Chapter 4 In this chapter, we detail the problem statement and its challenges. We comment on the related work that aimed to approach similar problems and introduce the construction process of the dataset for digital staining and evaluation metrics used in this work.

Chapter 5 Here we build upon the concepts presented in Chapter 4 and employ them for multi-domain multi-task learning. We propose a method for image-to-image translation developed on three domains that simultaneously generates an image from the target domain, classifies it and uses the segmentation module to preserve cells structure. The related work is:

- Tomczak, A., Ilic, S., Marquardt, G., Engel, T., Forster, F., Navab, N., Albarqouni, S.: Multi-task multi-domain learning for digital staining and classification of leukocytes. IEEE Transactions on Medical Imaging (2021). doi: 10.1109/TMI.2020.3046334

Chapter 6 We now switch focus to the more challenging problem of uncertainty estimation for the generative process. In this chapter, we propose an approach to disentangle the generator's latent space and corrupt such latent representations with noise to estimate the confidence of the generation process. The related work is:

- Tomczak, A., Ilic, S., Marquardt, G., Engel, T., Navab, N., Albarqouni, S.: Digital staining of white blood cells with confidence estimation

Chapter 7 Next, we extend the latent space corruption approach to the whole image, showing that the method generalizes well using a different dataset. The related publication is:

- Tomczak, A., Gupta, A., Ilic, S., Navab, N., Albarqouni, S.: What can we learn about a generated image corrupting its latent representation? MICCAI (2022)

Chapter 8 While all methods described so far were focused on the advancement of algorithms for artificial staining, in this chapter, we report a study conducted with clinical experts to assess the feasibility of artificial staining in the clinical pipeline. The related publication is:

- Tomczak, A., Boldu, L., Brock, J.P., Merino, A., Engel, T., Marquardt, G.: Diagnosing artificially stained images of peripheral blood cells. International Symposium on Technical Innovations in Laboratory Hematology (2022)

Chapter 9 Finally, we summarize our findings and lay out directions for future research in a concluding chapter.

2 Fundamentals

This chapter gives an overview of the required technical fundamentals, including the neural networks, specifically convolutional neural networks and how to apply them to problems such as classification and segmentation. Next, we explain the concept of Generative Adversarial Networks and their conditional version. Finally, we summarize the characteristics of imaging modalities upon which this thesis is based.

2.1 Deep Learning

Deep learning is a subset of machine learning dating back to the 1960s [96, 54], when the first components of today's systems were introduced, such as the multilayer perceptron and ways to compute gradients in such models. However, the idea fully formed in the 1980s in three parallel works [68, 85, 97]. Although it was computationally not feasible to take full advantage of the theories developed at that time, the fact is that the core ideas behind modern networks have not changed substantially since then.

2.1.1 Neural Networks

The quintessential part of deep learning is research focused on designing and developing neural networks. Neural networks [11] are a tool used to model functional nonlinear dependence between given input and output. They are defined by a set of consecutive transformations: linear layer transformations and nonlinear functions called activations [11, 38]. They are optimized for a defined, differentiable loss function through forward- and back-propagation [97]. The optimization process is not much different than any other machine learning gradient descent-based optimization. In principle, the biggest difference is the introduction of the nonlinearity layer, which can cause the optimization process to become nonconvex, as opposed to models such as linear regression or Support Vector Machines (SVM). There were many improvements and refinements introduced to improve the optimization process of neural networks, such as momentum [109], gradient clipping [77], normalized weight initialization [36], Adam optimizer [62], or batch normalization [52]. In this thesis, we used most of these refinements.

2.1.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) [69] are a type to neural networks designed to process grid-like data, such as images that can be interpreted as a 2D grid of pixels. CNNs are largely responsible for the success of deep learning for the computer vision applications. The main idea is to use convolution, a specialized type of linear transformation, instead of matrix multiplication. This is a big change, because instead of optimizing all the parameters of a fully connected layer that represents the matrix multiplication, convolutional neural network optimizes primarily the parameters of the convolutional kernels. Other layer that had been proposed in CNNs is a pooling layer that changes dimensions of the input data grid. The most popular is the use of max-polling proposed in [137]. Convolutional Neural Networks are one of the key models in deep learning. They are, to a great extent, responsible for its popularisation, as they have significantly broadened the range of applications. The main interest was sparked in 2012 when a competition on object recognition was won by [65]. Since then, CNNs have been the bases for many more specialized models, built to accommodate the growing datasets and challenges such as object recognition, semantic segmentation, object pose estimation, registration, and many more.

2.1.3 Classification with supervised deep learning

Neural networks for classification try to find the most optimal mapping $f(x) = y$ from input x to a category y. They are trained by minimizing a certain loss function with respect to a set of parameters θ to find an approximation $f(x;\theta) = y$. The most common loss for classification is optimizing the Cross-Entropy (CE) function:

$$CE = -\sum_{c=1}^{N} y_{o,c} log(p_{o,c}) \tag{2.1}$$

where N is the number of classes in our problem, y is an indicator (0 or 1) if the class c is correct for observation o, and p is the predicted probability of class c. In a supervised setting, we have access to labelled training examples.

2.1.4 Semantic segmentation with supervised deep learning

Supervised segmentation can be defined as a pixel-based classification. This means that to each pixel from an image x, we want to assign a category y. The task can be optimized by minimization of a combination of CE (see Eq. 2.1) and a DICE coefficient [78]. The DICE coefficient is defined as:

$$D = \frac{2\sum_i^N x_i y_i}{\sum_i^N x_i^2 + \sum_i^N y_i^2} \tag{2.2}$$

where x_i and y_i represent corresponding pixels in prediction mask and the ground truth mask. The denominator considers the total number of pixels present in both sets, and the numerator considers only the overlap between them. It combines the global and local information about the prediction in context of the ground truth.

2.2 Image-to-image translation

The task of transforming an image from one domain such that it has the style and characteristics of images from another domain is commonly called image-to-image translation. That could mean changing the structure's appearance in the image (style transfer) or changing its attributes (changing a serious face to a smiling face). In a biological setting, it could mean making an image captured with one modalit look like it was captured with a different modality, or in context of haematology, staining normalization or staining generation.

2.2.1 Generative Adversarial Networks

The direct Image-to-image translation problem was addressed with conditional Generative Adversarial Networks [39] (GANs). Before introducing them, let's focus on traditional GAN. The most basic GANs are models designed for image synthesis, where two separate networks, Generator and Discriminator, compete against each other. In practice, GANs for image generation consist of two convolutional neural networks; a generator that maps a vector sampled from a normal distribution to an image and a discriminator which learns to recognize if a given image is real or fake. Both networks are trained competitively in a minimax fashion. The training is based on the adversarial loss:

$$L_{adv} = log(x) + log(1 - D(G(z)) \tag{2.3}$$

where G is the generator network, D is the discriminator network, x is an image from real data distribution and z is a noise vector. The generator G is optimized to minimise this loss, and the discriminator D maximises it. During the optimization process, the network weights are updated in turns.

2.2.2 Conditional Generative Adversarial Networks

In principle, the output of a generative adversarial network can be conditioned on, amongst others, class, text, image or features. Instead of taking a noise vector as input, a conditional GAN uses the provided information (in the form of, for example, an image). Conditional image generation aims to solve the image-to-image translation problem. One of the first proposed approaches to conditioning the generation process on an input image was Pix2Pix [53], where the network transformed edges into photos and day images into night images. It takes an image from one domain, for example, an image of an unstained cell, and transforms it into the target domain to look like a stained cell. The optimization process is, in this case, based on pixel-wise loss -

most frequently, L_1- or L_2-distance, sometimes a combination thereof. The drawback here is the strong requirement for precisely matched image pairs. In reality, such datasets are rare due to costly preparation, which usually includes constraints on image acquisition or very precise image registration, to obtain pixel-wise correspondences. Therefore, there are not many works addressing this problem.
The recent works concentrate on improving the quality of the style of generated objects, making them to a great extent unrecognizable as fake, and enhancing the resolution, which leads to beautiful images with high definition. For a simple image-to-image translation, Pix2PixHD [123] was proposed to define a new standard with a multi-scale generator and multiple discriminators that allow translation of high-resolution images.

However, what should we do when image pairs are unavailable? The solution to this problem was proposed by J. Zhu et al. in [138]. In this work, they introduced cycle consistency. The idea here is to translate the image first from the source to the target domain and then back to the source. So given an image $x \in X_A$, image $y \in X_B$, generator G_A translating from domain A to domain B and generator G_B translating from domain B to domain A, the reconstruction cycle loss term is:

$$L_{cycle} = L_1(x, G_B(G_A(x)) + L_1(y, G_A(G_B(y)) \tag{2.4}$$

and it penalizes the differences between the original image and its translation to the target domain and back to the original domain. This proved to provide sufficient constraint to train a network successfully changing the appearance of an image. However, it should be mentioned here that this loss comes with a limitation. It provides virtually no control over what features define the appearance of a given domain. It provides both changes in the style and in the structure of a given image unless constrained further. The idea of unpaired image-to-image translation was further extended to multi-domain image translation by StyleGAN [56] and StyleGAN2[57] (alternative generator architecture), SPADE [84] (introducing new spatially adaptive normalization) model and StarGANv2 [23] (including style encoding network next to generator and discriminator) which I elaborate on further in Section 4.3.

The models that work on unpaired data are used mostly to generate different attributes and are not directly concerned with the faithfulness of the structure transfer. However, this is crucially important for medical applications. As shown by Cohen et al. in [24], the distribution matching losses (so cycle-consistency loss) can hallucinate features. In medical imaging, that can also mean generating tumours that have not been present in the original image or failing to generate ones that exist. Of course, from a clinical point of view, the second scenario can have catastrophic consequences, hindering such losses in a medical setting.

On the other hand, the models that learn direct mapping require many paired data points. The paired and unpaired data can be mixed in real-life scenarios, as just part of the data was processed to calculate the pixel-wise alignment.

2.2.3 Adventages and disadvantages of GANs

Generally speaking, GANs can generate high-quality samples and learn the underlying distribution. They are also very flexible when it comes to architecture. However, there are additional limitations to using GANs, the most significant being the training instability that is a direct consequence of the minimax training fashion. One of the results is the so-called mode collapse, where *G* degenerates and only synthesises images from one mode of distribution. There were many works addressing this limitations: WassersteinGAN [4], WassersteinGAN with gradient penalty [42], Hinge loss [15, 132], Progressive training strategy [55]. In this thesis, we use the Wasserstein gradient penalty [42] for all the optimization tasks based on the cycle loss.

2.3 Imaging modalities

There are different ways to capture images of peripheral blood, depending on whether we are interested in stained or unstained samples. Here, we will briefly introduce the two types of microscopy we used to collect data for this work.

2.3.1 Bright-field microscopy

Bright-field microscopy (BF) is considered one of the most straightforward optical microscopy illumination techniques. Sample illumination is transmitted (i.e., illuminated from below and observed from above). The attenuation of the transmitted light in dense areas of the sample causes white light and contrast. Bright-field microscopy uses a condenser lens that focuses light from the light source onto the sample. It receives light from the light source and concentrates the light rays on the object. However, bright-field microscopy typically has low contrast on most biological samples because only a few absorb light significantly. Therefore, staining is often required to increase the contrast [14].

CellaVision System One of the examples of commercial systems based on BF microscopy is the CellaVision system [83]. As an input, the system requires a fixed glass with a stained blood sample; as an output, it produces 360x360 image crops containing blood cells, together with their corresponding labels. However, in the case of pathological specimens, these labels need to be double-checked by a haematologist.

2.3.2 Differential Inference Contrast microscopy

There is also another type of microscopy that would allow the taking of images of low-contrast structures. Differential Inference Contrast (DIC) microscopy produces a distinctive, shadow-cast appearance to capture images of unstained blood cells. The DIC microscope uses dual-beam interference optics, transforming local gradients in the optical path length of an object into regions of contrast in the object image. The transformation is based on polarized light and two crystalline beam-splitting devices. In DIC microscopy, the specimen, which is in our case blood, is sampled by pairs of closely spaced rays generated by a beam splitter. Suppose the members of a ray pair traverse a phase object in a region with a gradient in the refractive index or thickness. In that case, there will be an optical path difference between the two rays upon emergence from the object, which is translated into a change in amplitude in the image. Since an optical path difference corresponds to a relative phase difference between the two rays, phase gradients are present in a DIC image [80].

In the following thesis, I will use the discussed techniques on the data collected with described modalities, aiming to mitigate their drawbacks via the new methods that we propose.

Part II

Staining in Hematology

3 Chemical Staining

The current laboratory pipeline relies on the process of chemical staining that enhances features allowing interpretation of the samples. This chapter includes a detailed description of the workflow, specifically the staining part of it, highlights the features relevant to each blood cell class, outlines challenges connected to classifying each of them and finally, comments on the clinical relevance of the correct white blood cell classification.

3.1 Blood film preparation and examination

Some requirements must be fulfilled before obtaining the blood sample and performing the blood cell count. The steps should be performed carefully so that no artefacts are introduced during these procedures. After diligently checking the subject's identity, a qualified person performs a venepuncture (usually in the antecubital fossa) using a needle and draws a blood sample into a syringe or an evacuated tube. Next, the blood specimen is mixed either in a mechanical rotating mixer or by manual inversion. The next step is preparing the blood film - it is to be prepared and examined on only a small part of the specimen. A drop of blood is placed near one end of the glass slide, which was thoroughly cleaned before. Then it can be spread either manually (with a smooth, steady motion) or automatically with a mechanical spreader integrated into a staining machine or an automated blood counter. Spread blood film is subsequently air-dried and fixed in absolute methanol for 10-20 minutes. After the sample is fixed, the actual staining procedure can start.

Staining. The staining process begins with stain mixture preparation. The laboratories differ in the exact stain composition used to prepare a blood film for microscopic examination [5]. Most of the institutions in Europe, including the laboratory processing samples used in this thesis, use May-Grünwald-Giemsa (MGG) stain, which is a mixture of:

- a methylene azure which conveys a blue-violet to nucleic acids and nucleoprotein, to the granules of basophils and slightly less, to the granules of neutrophils,

- eosin which brings a red or orange colour to haemoglobin and the eosinophil granules and contributes to the colour of the stained nucleus.

The MGG staining allows humans to see features such as granulation, cytoplasm and nucleus which are indispensable to correctly identify the cell type, as shown in Fig. 3.1.

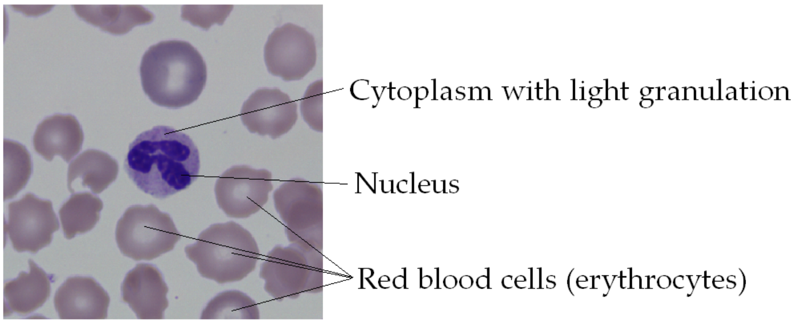
Cytoplasm with light granulation
Nucleus
Red blood cells (erythrocytes)

Figure 3.1: Leukocyte example.

However, there can be a variance in the outcome of the staining, depending on the protocol. The staining protocols differ across hospitals and countries. It is the root cause of the lack of standardisation between samples stained in different places. It introduces a lot of challenges for automatizing the procedures. Namely, models trained on data stained with one protocol will underperform on samples stained according to a different protocol. Moreover, even often within laboratories, the staining can differ as it is not immune to human impreciseness. All these changes make it impossible to guarantee the results of the algorithms trained in different centres. This significantly complicates the problem of collecting data and creating quality datasets that could be used to train deep learning models.

3.2 Types of white blood cells

In this work, we were interested in distinguishing up to 14 classes of leukocytes, artefacts and nucleated red blood cells. We list them here together with the reference images, characteristic features enhanced by the chemical staining that allow their identification [5] and the main challenges one faces during the classification process:

- Segmented neutrophil (see Fig. 3.2) Their size is around 1015 microns, the nucleus has coarse, clumped chromatin and 35 segments. The cytoplasm is pink to colourless with moderate to plentiful, very fine violet granulation. The SNEs are one of the most common classes - they comprise around 60%-70% of all white blood cells in the healthy blood sample. The main challenge is distinguishing them from their predecessor - Band

Neutrophils (Fig. 3.3). Although difficult, some haematology experts agree that this differentiation has no clinical significance.

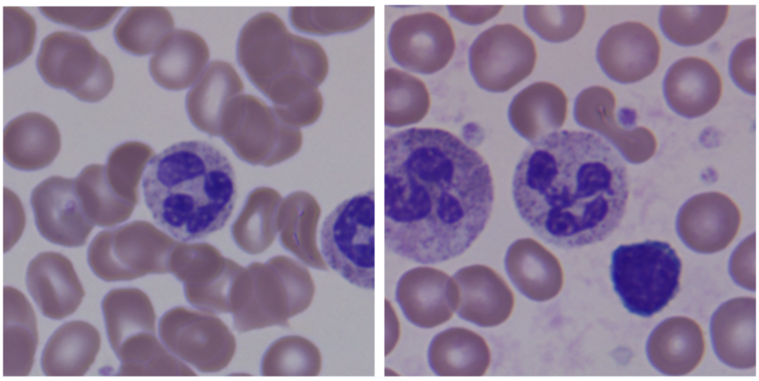

Figure 3.2: Segmented neutrophils (SNE).

- Band neutrophil (see Fig. 3.3) Their size is around 1015 microns, the nucleus is band-shaped, has clumped chromatin and no nucleoles. The cytoplasm is pink to colourless with plentiful, very fine violet granulation. The main challenge is to distinguish them from their successor, segmented neutrophils. The differentiating quality is the number of segments of the nucleus - although already banded, it remains one segment.

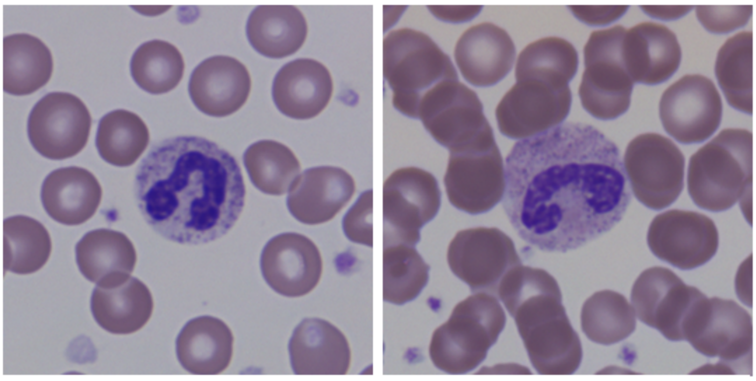

Figure 3.3: Band neutrophils (BNE).

- Basophil (see Fig. 3.4) Their size is around 1015 microns, the nucleus is diffused, has clumped chromatin and 34 poorly defined segments. Cytoplasm is plentiful, pale pink to colourless, with plentiful, unevenly sized, round, bluish-black granulation. The most important feature is characteristic granulation.

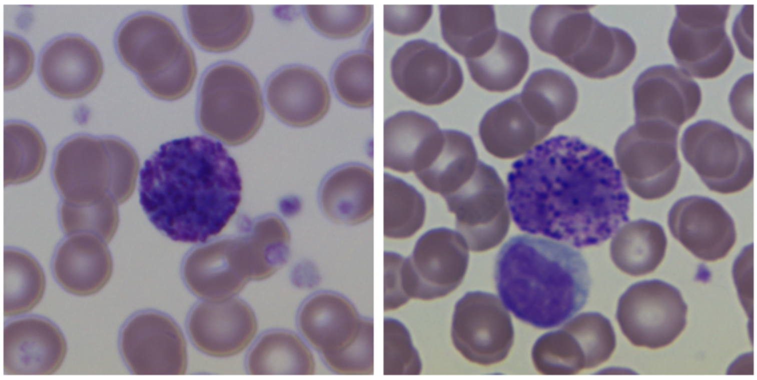

Figure 3.4: Basophils (BA).

- Eosinophil (see Fig. 3.5) Their size is around 1015 microns, the nucleus has coarse, clumped chromatin and 23 segments. Cytoplasm is nearly invisible, pale, pink to colourless with plentiful, large, round, brick-red granulation. The most important feature is characteristic granulation.

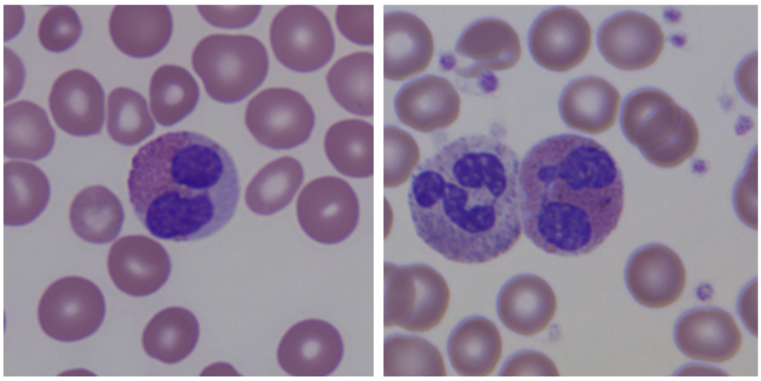

Figure 3.5: Eosinophils (EO).

- Monocyte (see Fig. 3.6) Their size is around 1224 microns, and the nucleus is kidney-shaped, bulging or segmented. Cytoplasm has vacuoles, is plentiful, dirty grey, with powder-fine, violet granulation. The most significant features for the classification are the nucleus' shape and the vacuoles. In more complex cases - to distinguish them from certain types of blasts - the colour and texture of the cytoplasm are crucial.
- Lymphocyte (see Fig. 3.7) Their size is around 78 microns, and the nucleus is round to oval, with condensed chromatin and no nucleoles. Cytoplasm is a light to dark blue band with few bright violet granulations. The

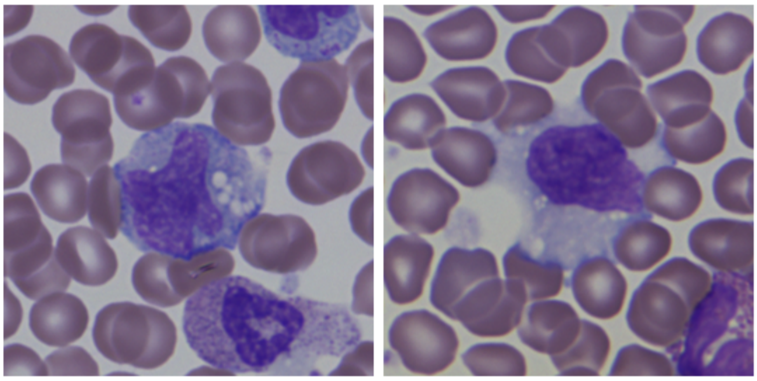

Figure 3.6: Monocytes (MO).

identifying features are the shape and colour of the nucleus and the colour of the little cytoplasm that is present. It can be difficult to distinguish them from their pathological versions - reactive and atypical lymphocytes.

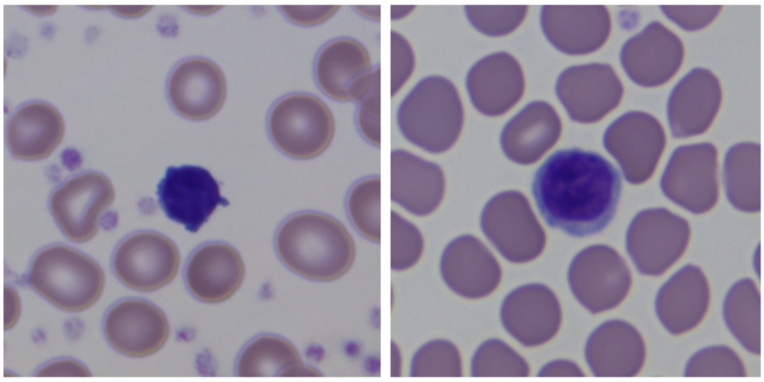

Figure 3.7: Lymphocytes (LY).

- Smudge (see Fig. 3.8) Smudges are cells destroyed during the smearing or staining procedures. The destruction may be caused by mechanical factors or diseases such as leukaemia that weaken the cells. The main challenge in classifying smudges is to differentiate them from the cell classes they used to be before, which may be difficult depending on the level of impairment. They are clinically significant if they result from the pathological weakening of the cells. This can be generally defined based on their number in the smear.

- Promyelocyte (see Fig. 3.9) Their size is around 1224 microns, the nucleus

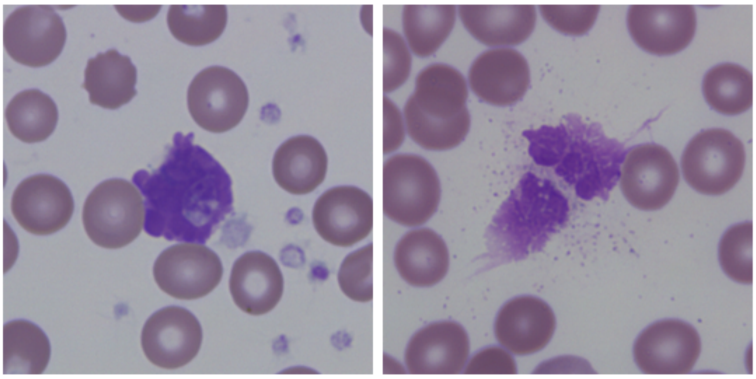

Figure 3.8: Smudges (SMU).

is oval, possibly bulging, eccentric with coarse chromatin and 34 nucleoles. Cytoplasm is plentiful and pale blue with moderate to plentiful, unevenly sized, violet granulation. The main challenge is to distinguish them from their successors - myelocytes.

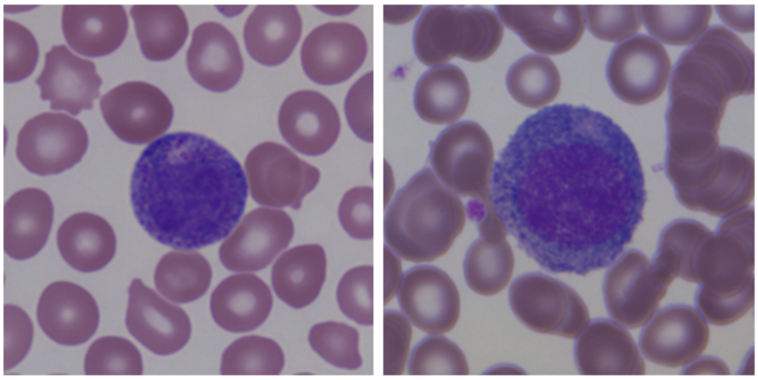

Figure 3.9: Promyelocytes (PMY).

- Myelocyte (see Fig. 3.10) Their size is around 1018 microns, and the nucleus is round to oval, eccentric with coarse, condensed chromatin. Cytoplasm is plentiful, pink to colourless, sometimes bluish with moderate to dense, very fine, violet, pink or black granulation. The challenge is distinguishing them from their predecessors - promyelocytes and successors - metamyelocytes.
- Metamyelocyte (see Fig. 3.11) Their size is around 1016 microns, and the nucleus is indented with coarse, clumped chromatin and no nucleoles. Cytoplasm is pale pink to colourless with plentiful violet, pink or

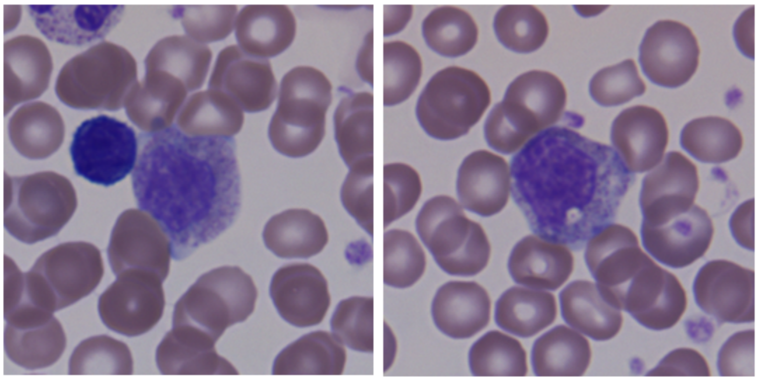

Figure 3.10: Myelocytes (MY).

black granulation. The main challenge is to distinguish them from their predecessors - myelocytes.

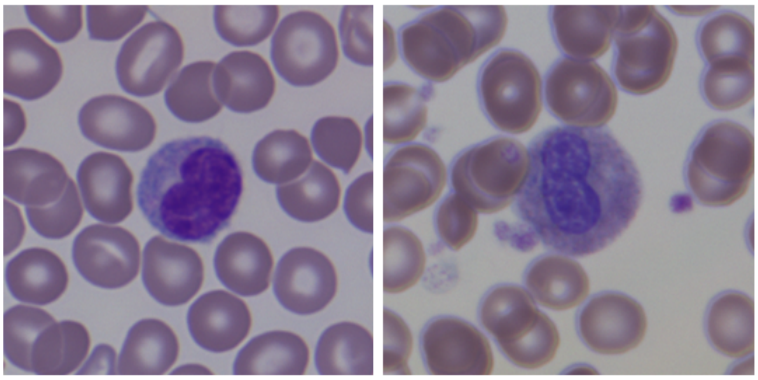

Figure 3.11: Metamyelocytes (MMY).

- Blast (see Fig. 3.12) Their size is around 1020 microns, the nucleus is large, round to oval with evenly stained chromatin and 15 nucleoles. There is a small amount of blue cytoplasm with a clear zone next to the nucleus and no granulation. Blasts are pathological cells that are normally found in the bone marrow. Their presence in the blood smear is indicative of a disease. Blasts are predecessors of most other classes, and the main challenge is not to confuse them with their successors - promyelocytes, monocytes and lymphocytes.
- Atypical lymphocyte (see Fig. 3.13) Their size is around 915 microns, the nucleus is oval or bulging with looser chromatin and no nucleoles. Cytoplasm is plentiful, colourless to light blue, with few bright violet granules.

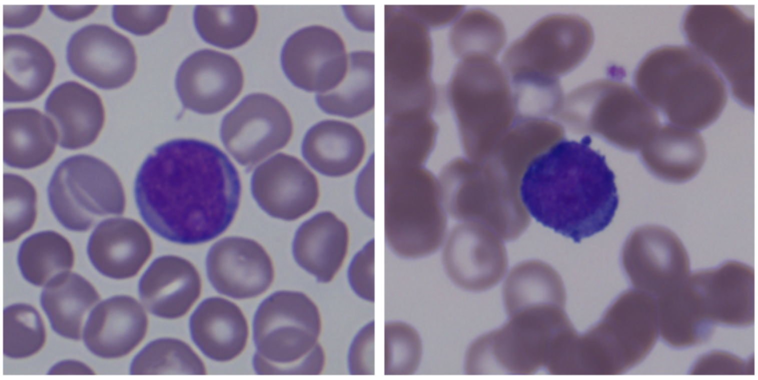

Figure 3.12: Blasts (BL).

The main challenge is not to confuse them with normal lymphocytes.

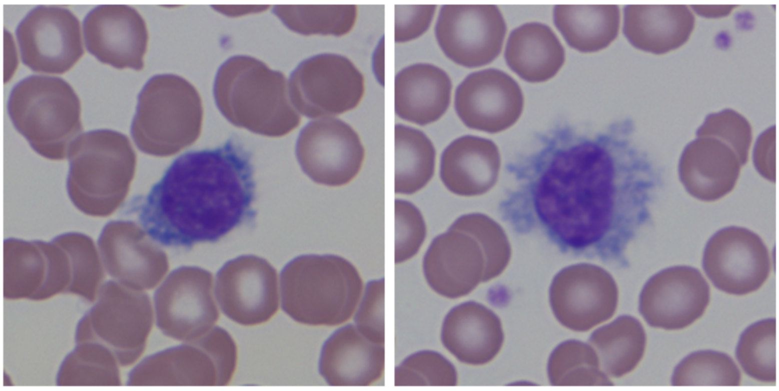

Figure 3.13: Atypical lymphocytes (ALC).

- Reactive lymphocyte (see Fig. 3.14) Their size is around 1530 microns, and the nucleus is round, oval, folded or lobulated with no to multiple nucleoles. Cytoplasm is plentiful, grey to blue, darker in the periphery and lighter near the nucleus with no granulation. The main challenge is not to confuse them with normal lymphocytes and blasts.

- Plasma cell (see Fig. 3.15) Their size is around 715 microns, and the nucleus is round, eccentrically with compact chromatin. Cytoplasm is dark blue with occasional vacuoles and no granulation. Their most characteristic qualities are an abundance of cytoplasm and basophilic granulation. The main challenge results from the fact that they are very rare, which in the context of deep learning is unequivocal to underrepresented in the data set.

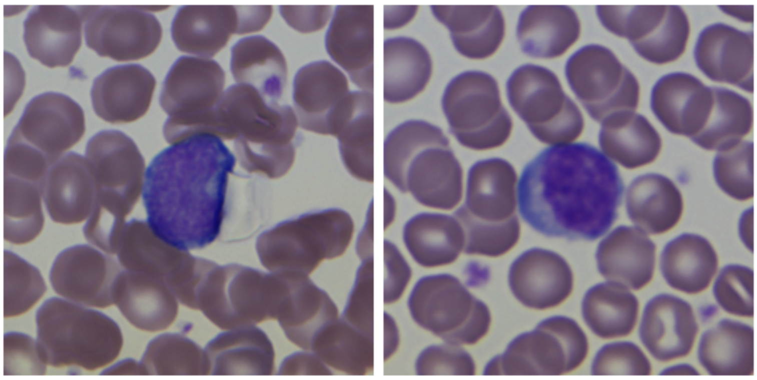

Figure 3.14: Reactive lymphocytes (RL).

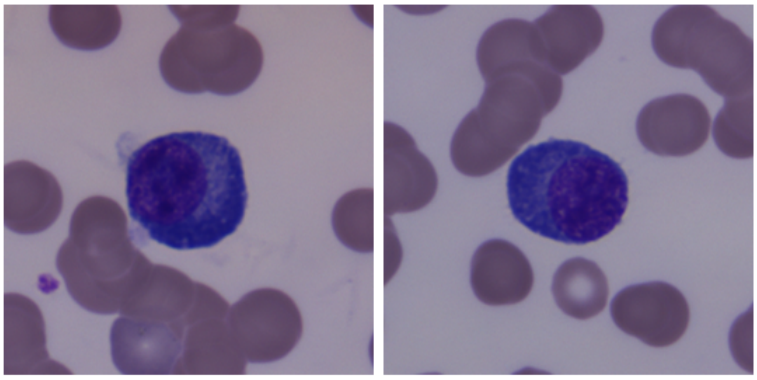

Figure 3.15: Plasma cells (PC).

- Nucleated red blood cell (see Fig. 3.16) Their size is around 810 microns, and the nucleus is small, round to oval with coarse chromatin. The cytoplasm is pink with no granulation. Although technically, they are not leukocytes; their appearance often causes them to be detected as ones. In contrast to normal red blood cells, they possess a nucleus. [!h]

- Artifact (see Fig. 3.17) This class does not represent a blood cell type but a broad category that includes all types of artefacts created during the staining procedure. Although clinically insignificant, must be differentiated from white blood cells, as common detection models often erroneously include them as cells.

As presented in the Fig. 3.2-3.17, the most significant features are the size of the cell, the colour of the cytoplasm, size and colour of the granulation, shape and colour of the nucleus and visible nucleolus. These features allow for the correct classification of a sample. A successful staining preserves and

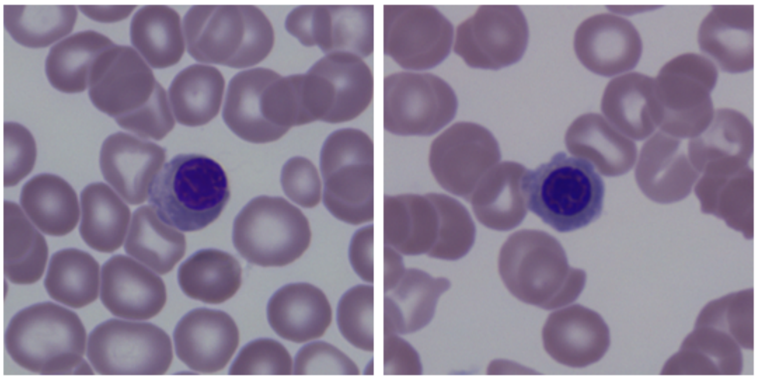

Figure 3.16: Nucleated red blood cells (NRBCs).

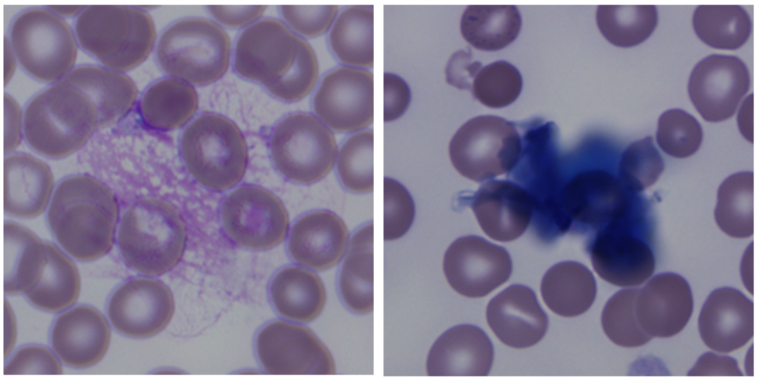

Figure 3.17: Artifacts (ART).

enhances all of these features so that the white blood cell can be identified after the processing. However, the aforementioned changes in the staining protocol or procedure lead to variations in the appearance of the cells and can contribute to misclassifications. The main challenge of the classification is that all the white blood cells come from a single root. As the cells mature, they change class, for example, the promyelocyte matures into a myelocyte and later into a metamyelocyte. The boundaries between them are not completely strict. We present the development stages of different blood cells in Fig. 3.18. This complexity of the classification problem also outlines one of the main challenges for artificial staining: identification and preservation of the relevant features.

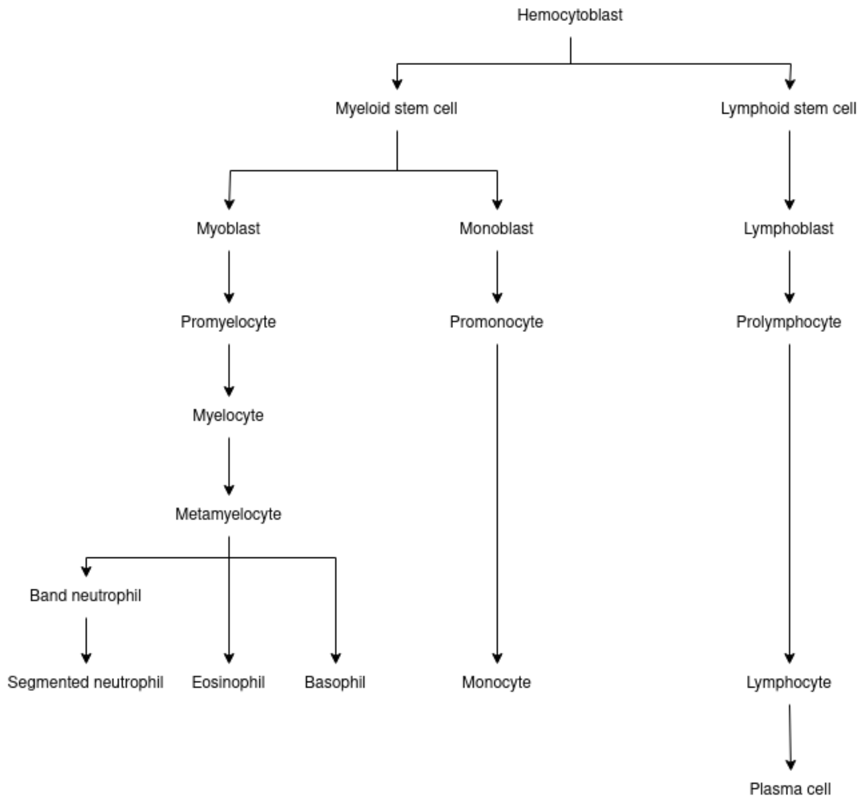

Figure 3.18: Development stages of white blood cells. The tree-like structure illustrates the interdependence between cells and underlines how challenging correct classification is.

3.3 Clinical significance of correct identification of white blood cells

It is critical in clinical practice to correctly identify types of the present white blood cells. In general, they can be divided into two categories: non-pathological and pathological. The non-pathological classes include segmented neutrophils, band neutrophils, basophils, eosinophils, monocytes and lymphocytes. Although just the presence of such classes does not indicate a disorder, an unbalance in the proportions considered normal is indicative of a pathological condition. For example, an increased number of neutrophils hints at bacterial or fungal infection; a rise in eosinophils means allergy, parasite, autoimmune disease or, in certain cases, spleen or central nervous system disease; an increase in basophils is a response to allergens and invaders; more monocytes than usual can signify infection or, in more severe cases, leukaemia. On the other hand, the presence of classes such as promyelocyte, myelocyte, metamyelocyte, blast, atypical lymphocyte, reactive lymphocyte, and plasma cell is a sign of a severe disorder. Promyelocytes signify acute promyelocytic leukaemia; myelocytes - chronic myeloid leukaemia; metamyelocytes and blasts - cancers such as myelodysplastic syndrome and myelogenous leukaemia. Further, atypical and reactive lymphocytes are signs of diseases such as mononucleosis, hepatitis A, B, and C, toxoplasmosis and acute lymphocytic leukaemia. Plasma cells in the blood smear indicate multiple myeloma. The severity of these diseases points to the severity of misclassifications, in particular, the false negative predictions that would result in the misdiagnosis of diseases where time is of the essence.

To summarize, the correct identification of white blood cells allows for the diagnosis of a wide range of disorders, from infections to cancer types. Mistakes in the predictions carry severe clinical significance.

4 Digital Staining

In this chapter, we start with the definition of the digital staining problem and the challenges it poses. Next, we present the research work related to the problem: the deep learning algorithms already used in haematology, the disentanglement techniques for neural network's latent representation, the methods to preserve structure on a generated image, the multitask learning set-up and the uncertainty estimation for image generation. Finally, we present the process of constructing the datasets used in this thesis and the evaluation metrics we used to analyse our results.

4.1 Problem statement

The problem I wanted to solve in this work is transforming an image of an unstained white blood cell into its stained version while preserving all the features relevant for classifying the blood cell. Namely, to take an image of an unstained cell captured with a DIC microscope and transform it to look like a stained cell captured with a BF microscope. To put it in the context of computer vision: it is a problem of image-to-image translation, where our source domain is an unstained set of cells, and our target domain is a stained set of cells. Our downstream task is image classification, so certain features, for example, size and shape, have to be preserved. Additionally, the colours need to be adequately generated.
We can separate such three parts of the problem:

- Building the most suitable data set and designing models that use the available information to the maximum.
- Development of methodologies preserving the features during the translation process and considering the probabilistic nature of the output.
- Answering the question if there exists a real possibility for haematologists to work with digitally stained images instead of chemically stained ones.

All these parts have different complexity and pose distinct challenges.

4.2 Challenges

We can divide the main challenges of this problem into three groups: data-driven, problem-driven and nature-based.

- **Data-driven:** the first part is mainly connected to the technical side of the problem: how to create the best quality dataset? The distribution of the classes prevents the full control of the class balance in the dataset. Additionally, to the fact that some cell types are extremely rare, there exists variability within a given cell class, as different diseases can influence the same classes. There is also the challenge of creating a fully pixel-wise aligned dataset - it would provide better results because of approximating the direct mapping between cells, but with a price of elaborate preprocessing. These challenges need to be addressed as deep learning is a data-driven approach, and dataset construction is crucial to the final result.

- **Problem-driven:** how to preserve all the classification relevant features mentioned in Section 3.2, namely the granulation, cell and nucleus shape and the correct colour scheme? Even slight variations in any of these qualities may lead to a wrong prediction in terms of classification. This challenge needs to be addressed with a specific methodology. Additionally, knowing the possibility of the features not being present on the unstained images, we also need to answer a question: how confident the network was while generating the images?

- **Nature-based:** challenges can also come from the fact that assigning the right class to the most demanding examples is difficult even for haematologists. We can attribute it to the fact, that different classes come from a single stem cell as pictured in Fig. 3.18. It makes the classification problem naturally difficult.

All these challenges are part of this thesis and we address them with a mix of algorithmic and methodological solutions.

4.3 Related work

In this thesis, I tackle multiple problems connected to image classification and image-to-image translation. I will now comment on using deep learning for haematology, multi-domain and multi-task learning, latent space disentanglement, structure preservation and uncertainty estimation. These subjects are relevant to the methodologies presented later in this work.

4.3.1 Deep Learning for haematology

Many deep learning based-algorithms have been explored in the domain of haematology in terms of the classification of blood cells, staining generation and staining normalization.

Blood cell classification Early works on WBC classification used classical machine learning methods like the k-Nearest Neighbor (KNN) algorithm [127], Bayesian classifiers [35] or Support Vector Machines (SVM) [26][105]. Those works are based on small datasets using handcrafted colour, intensity, shape and texture features. Depending on the data used, the classification accuracies oscillate in a range of 83-99%. Recently, deep learning-based approaches have been successfully applied to leukocyte classification [75, 107, 113, 122, 89, 71, 79, 44] [82]. In [51][108], ANNs and CNNs outperform SVMs on WBC classification when applied to the same data. Yildirim et al. [129] used a dataset with four classes (excluding Basophils) and compared different neural network architectures: AlexNet [64], DenseNet201 [49], GoogLeNet [110] and ResNet50 [46], achieving the highest accuracy of 83.44% with a DenseNet201. Almezhghwi et al. [3] compared a VGG, ResNet and DenseNet on a five-class dataset achieving accuracy up to 98.8%. Their findings showed that models with pre-trained weights on the CIFAR-100 dataset perform better than models with random weight initialization. Other works used ResNets for WBC classification [45, 66, 121]. They achieved test accuracies between 88.29 % and 99.84%. Additional works used VGG-16, InceptionNet-v3 and ResNeXt or other CNN based architectures obtaining similar results [2][76][102]. On the other hand, Liang et al. proposed a Recurrent Neural Network (RNN) architecture combined with a CNN to exploit long-term dependence relationships in WBC images [72]. They outperformed their CNN baseline by around 2% to achieve a test accuracy of 90.79% using four classes in the dataset. They stacked the RNN and CNN network in parallel, merging the features with an attention mechanism for a final linear classification layer. Most of these approaches rely on the blood cells of healthy donors, which lack the difficulty present in the dataset we used. All of these methods are applied to private in-house datasets, all with a different number of classes included and posing different challenges. They all work with stained white blood cells.

Staining normalization GANs for image-to-image translation are also present in histopathology. Previous works have demonstrated great effectiveness of GANs for the problem of intra-/inter-scanner variability and stain normalization [131, 30, 101, 93, 9, 98]. Although making the algorithms more efficient on a different kind of staining, as well as on data coming from multiple institutions, it does not eliminate the staining process from the laboratory pipeline, remaining more of a bandage than a remedy. Interestingly, the authors in [30] noticed the problems with structure preservation, and later D. Mahapatra et al. in [74] approached it with self-supervised learning. Recent works focus on the correct structure preservation while using cycle-consistency loss [74].

Staining generation Recently, some works [94, 90] have successfully applied adversarial generative models to digital staining in histology to generate the staining for unstained samples. Two works, which are closely relevant to ours, transfer the knowledge of staining to unstained sample, Rivenson et al. [94] to quantitative phase images using Pix2Pix [53] and matching pairs of images. Rana et al. [90] to unstained whole slide images using CycleGAN [138].

However, neither of these works tackles the problem of a mixed dataset containing both paired and unpaired data or performing any additional task, like classification.

4.3.2 Multidomain learning

Multi-domain classification problem has been addressed by domain adaptation methods, such as DANN [32], DADA [87], MADA [86], MuLANN [100], MDAN [136] and SIFA [18]. Most of these methods rely on Generative Adversarial Networks (GANs). Those image translation methods produce realistically-looking images that preserve the target domains' image style without explicitly preserving the structure. In contrast, our application requires the structure to be preserved because haematologists are trained to identify distinctive landmarks, such as shape, size, and granules in the cytoplasm.

4.3.3 Multitask learning

Multiple works tackle the problem of multi-task learning, where the data representation is built using autoencoders or GANs[48, 92, 34, 99]. Ghifary et al. propose in [34] an autoencoder-based neural network, with the additional use of a middle feature map for cross-domain object recognition, whereas Sankaranarayanan et al. in [99] uses GAN-based network to construct a hidden feature map that can be later used for classification. Remi et al. [92] uses GAN conditioned on segmentation map for cross-view image synthesis. In the context of histology data, Bashir et al. [6] proposed a model for simultaneous detection and classification of cells on a smear, later Graham et al. [40] presented a solution for simultaneous classification and segmentation of a histology image. Song et al. [106] investigated simultaneous detection and classification.

4.3.4 Disentanglement

One of our goals through this work is to disentangle the style and structure of an image. It would allow us to preserve the structure and only change the style while translating an image. The style of an image can be defined in multiple ways: from the art style of a painting, through the type of the animal, up to the lightning of the image. Most works consider attributes such as hair colour or facial expression style characteristics. The content of the images is even more complicated: from the strict definition provided by semantic segmentation to the angle at which a person is facing the camera. However, independently of the definition, there is joint agreement that the unsupervised image translation may contain significant entanglement of content and style, causing changes in the constitution of a given object that appear during the translation process. An example of this is hallucinating cancer tumours in MRI images, as mentioned [24]. Therefore, many works concentrate on disentanglement of the content and style, which could potentially adapt the style without losing the crucial structural characteristics of the objects. Both DRIT++[70] and [139] investigate using distinctive style and content codes for image generation, using an image of a different domain as a reference instead of explicitly learning the content code. [59] proposed SC-GAN using features from a pre-trained VGG network to disentangle content and style. [63] explored the use of fixpoint disentanglement loss for content and style in the context of artistic style transfer. [37] proposed a model enforcing representation disentanglement through several losses and a Gradient Reversal Layer. The problem of content-style disentanglement was tackled not only in GANs setting but also for the VAEs [134, 29, 124]. These works study unsupervised disentanglement of structure and style rather than using a semi-supervised approach with a pre-trained segmentation network as we do, possibly due to the more complex structures of the considered objects.

4.3.5 Structure preservation

Another set of works concentrates directly on the structure preservation for the end task (segmentation or style transfer). The structure of an image is understood as its semantic segmentation mask or extracted edges. Not all works consider it necessary to disentangle the latent space to preserve the structure and treat the structure preservation simply as one of the desirable characteristics of a model. [17] propose a Domain Invariant Structure Extraction framework with multiple encoders and a discriminator with a goal of successful semantic segmentation across domains. Both [112] and [21] use an additional Edge Generator Network or Edge Detector, respectively, to preserve the semantic information in a generated image. On the other hand, [125] proposes a model composed of two autoencoders and two transformers to preserve the geometrical information in an image. Another area of research concerned with preserving structural information is semantic image synthesis [84, 25, 73, 111].

4.3.6 Uncertainty estimation

The methodologies for estimating uncertainty and confidence in deep neural networks mainly focus on classification problems [12, 31, 67, 60, 88]. These methods are, unfortunately, not applicable for quantifying generator confidence in GANs. Over time, many metrics have been proposed to evaluate GANs [8, 41, 13, 104], some being image quality measures and others focused on how well the learned distribution reflects the real one. However, few works have explored the confidence related to the generated images. Indeed, such measures would be particularly interesting in medical applications as suggested by [7]. In one of the recent works, U. Upadhyay et al. propose modelling per-pixel heteroscedasticity as generalized Gaussian distribution [119] and using it to guide a progressive GAN [120] for PET to CT translation.

4.4 Dataset

For this project, we collected the data and built our in-house dataset. Here I will describe all the steps taken to construct it.

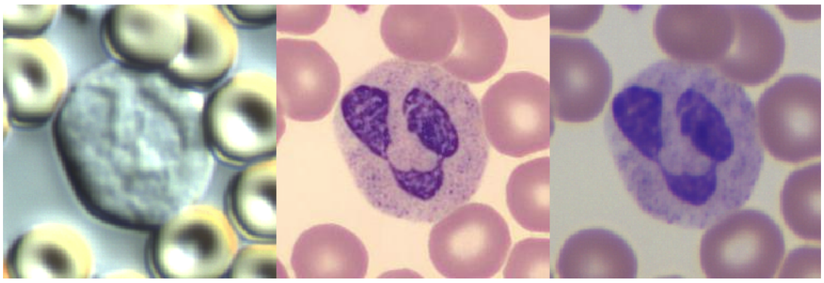

Figure 4.1: Example of a cropped white blood cell in three different domains (from the left): DIC, BF, CV.

Blood samples collected in EDTA were semi-automatically prepared using the slide maker HemaPrep (CellaVision [1]). Digital images of blood cells were acquired with the automated LeicaDMi8 microscope in three different imaging modes: (described in Section 2.3)

- Differential inference contrast (DIC) technology, a high contrast strain-free imaging technique to capture unstained images of blood cells
- Brightfield (BF) microscopy to capture stained images of blood cells. Smears were stained with May Grünwald-Giemsa.
- Commercial CellaVision system (CV).

We show examples of cell images belonging to all three domains in Fig. 4.1. To obtain the cell crops, big field images (2,064x1,544 pixels, see Fig. 4.2) were acquired with DIC and BF microscopy and processed further.

4.4.1 Dataset structure

Our dataset consists of leukocyte images acquired by three scanners from 24 healthy donors. As reported in Table 4.1, our training set contains white blood cell images from three different domains coming from 20 healthy patients. DIC domain has images of unstained cells captured with high contrast stain-free imaging technique. BF domain includes images of the same blood cells, but after the staining process, captured with a brightfield microscope. CV domain contains images of blood cells from the same patients, captured and automatically labelled with the CellaVision system and later validated by experts. This is the only almost completely labelled domain (6951 labelled images out of 7017 available images). Since the data comes from the same patients, a complete overlap exists between domains DIC and BF, and a partial overlap exists between DIC and CV. Because of the overlap, we were able to transfer part of the labels from CV to both DIC and BF. We classify the white blood cells into the seven most common classes for healthy donors: Basophil, Eosinophil, Monocyte, Neutrophil, Lymphocyte, Artifact, and Smudge. The classes are heavily unbalanced. We account for it, oversampling the underrepresented classes. Our testing set consists of leukocyte images from DIC coming from four different patients. For evaluation of the digital staining, we have 545

Table 4.1: Structure of our dataset. It composes of three domains: $\mathcal{X}_{DIC}$ (unstained images), $\mathcal{X}_{CV}$ (stained images captured with CellaVision system) and $\mathcal{X}_{BF}$ (stained images captured with brightfield microscopy) with a different number of samples and labels.

	Training set - X_{CV}	Training set - X_{DIC}, X_{BF}	Testing set - classification	Testing set - staining
ART	609	2	50	1
BA	40	15	21	7
EO	136	24	40	9
LY	2051	691	680	14
MO	310	94	150	28
SNE	3239	59	1611	339
SMU	566	1098	257	147
No. of labels	6951	1983	2809	545
No. of cells	7017	11227	2809	545
No. of patients	20		4	

matching images from DIC, CV, and BF. For evaluating the domain agnostic classifier, we utilize 2809 unstained leukocyte samples from DIC. The labels were obtained by transferring from paired images in BF provided by an expert.

4.4.2 Data preparation for digital staining

We detected the white blood cells using the YOLO detector [91] on original histological brightfield microscopy images (example shown in Fig. 4.2). Next, we aligned the brightfield images, and unstained DIC images using Enhanced Correlation Coefficient (ECC) maximization [28] and transferred the detected bounding boxes to the unstained images. Based on the bounding box coordinates, we cropped 256x256 patches from stained and unstained images.

4.4.3 Data preparation for classification

The labels for classification were automatically generated by the CellaVision system and only validated by a haematologist. However, since the overlap between the images in the domain CV and DIC is not large, we could transfer just a fraction of the labels from CV to DIC (1983 out of available 6951 labels). Since we want to classify unstained images for which a human annotator cannot provide the labels, the alternative would be to have someone manually label the stained brightfield images. Compared to this option, we significantly reduced the labeller workload. The lack of labels in the domains DIC and BF and the tedious process of labelling by experts is the primary motivation for us to build domain agnostic feature representation capable of performing good classification regardless of the domain from which the data are coming from.

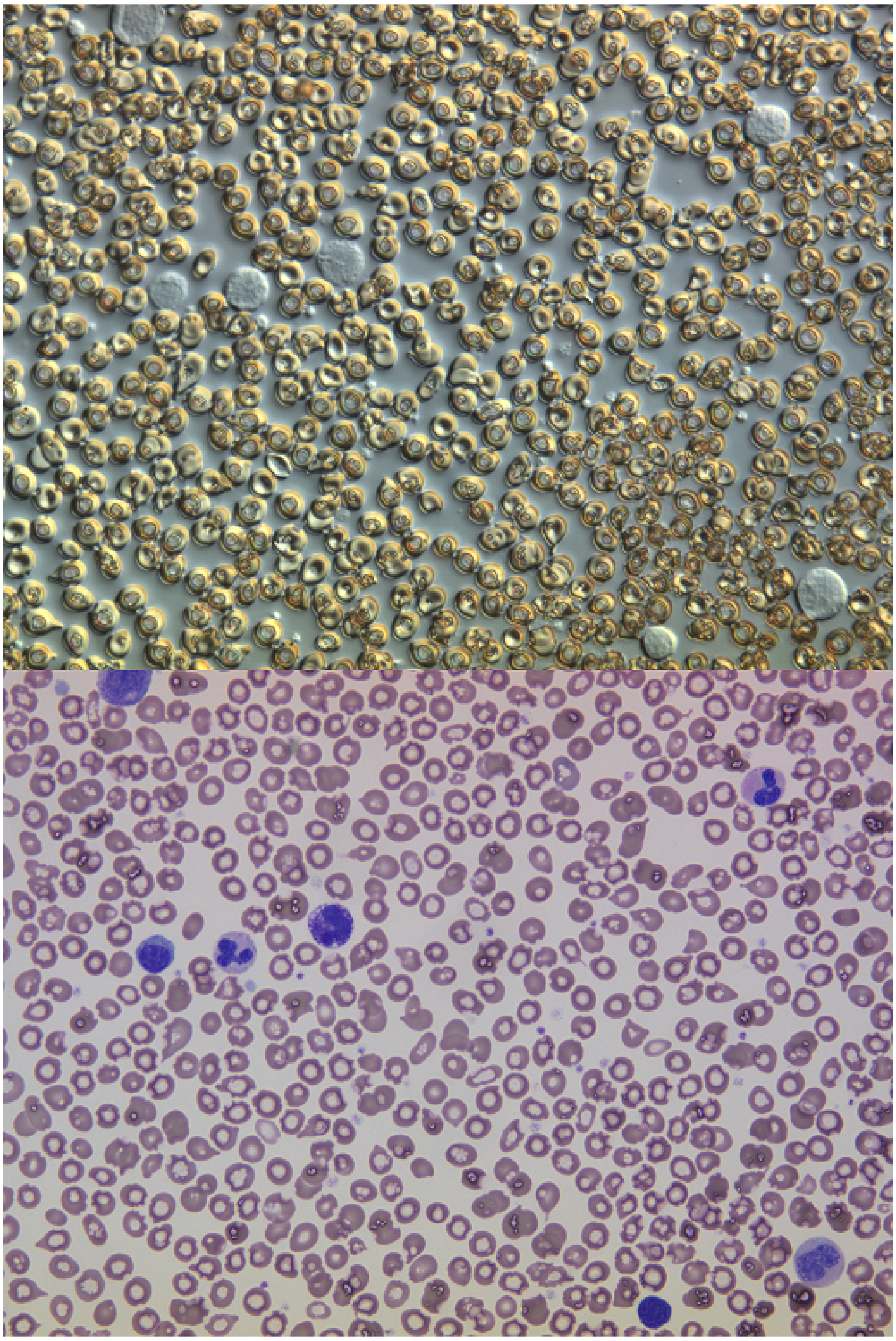

Figure 4.2: Example of an original unstained (top) and stained (bottom) whole-slide image.

4.4.4 Data preparation for segmentation

We need pseudo-labels for the segmentation task to help with structure disentanglement, so we trained a U-net segmentation network on a different set of 1239 manually labelled images from the domain with three classes: Background, Nucleus, and Cytoplasm. The fourth class - red blood cell - was obtained with thresholding. These images were not part of our training or testing set.

4.5 Evaluation

We approach the evaluation of digital staining from the side of image quality and in context of classification. Classification can be evaluated both directly and indirectly. In this section we describe all of the metrics used in this thesis and comment on their limitations.

4.5.1 Direct evaluation

Image quality. To evaluate the quality of digitally stained images, we use well-established metrics: Mean Squared Error (MSE), Structural Similarity Index (SSIM), Fréchet Inception Distance (FID) [47] and Learned Perceptual Image Patch Similarity (LPIPS) [133]. MSE and SSIM are traditional metrics quantifying similarity between images in terms of exact pixel values (MSE) and mean, variance and covariance (SSIM). As shown by [133], they do not always manage to capture a perceptual similarity. Therefore, we additionally use two "deep learning" metrics based on the differences in the features extracted by networks pre-trained on the ImageNet dataset. FID score measures the distance between the distributions of two sets of images. LPIPS is another perceptual metric that uses the L1 norm to calculate the differences between the features extracted by all the network layers.

Classification and segmentation. To evaluate the classification performance, we use accuracy and confusion matrices. We use the defined earlier (please see Eq. 2.2) DICE coefficient to evaluate the segmentation performance.

4.5.2 Comments on image quality evaluation metrics

The image quality metrics described above have theirs limitations. The traditional metrics such as MSE and SSIM do not fully capture the complexity of perceptual image similarity. On the other hand, while using the perceptual metrics such as FID and LPIPS, we have to consider them with a grain of salt as well, since the pretrained models used to extract the compared features were trained of natural images dataset.

4.5.3 Indirect evaluation

As a means of indirect evaluation, we have chosen automatic classification of artificially stained images - namely, applying models trained on stained data to the artificially stained samples. Of course, this comes with its limitations - there is no way of specifying whether the model misclassified a cell because significant features were not preserved or because the overall quality of an image was too poor.

4.5.4 Human evaluation

Lastly, we considered human evaluation, when haematologists themselves classify the artificially stained images or determine an image to be of quality too poor to be identified. This evaluation is especially important for clinical applications and ultimately validates the quality of the generated staining.

Part III

Multidomain Multitask Learning

5 Simultaneous image translation and classification

In this chapter, we explore the possibilities of combining multi-task multi-domain learning for haematology, as every used modality provides data that differ significantly. We take advantage of labels automatically provided for one of the domains and use segmentation pseudo-mask to improve the image generation process.

Our model is confronted with data belonging to multiple domains, as described in Section 4.4. We propose a model with two objectives: (1) automatically classify unstained images and (2) digitally stain them so that they can be verified by the haematologists, who are used to working with stained images. In our setting, most classification labels come from the CellaVision (CV) system and are verified by experts. Therefore, the domain CV is completely labelled. The second domain consists of stained images captured with Bright Field Microscope (BF). And finally, the third domain includes unstained DIC images. Images of all three domains, DIC, CV and BF, come from the same patients. Due to a partial overlap between CV and BF smears, part of the labels can be transferred from CV to BF data. Next, the BF labels are transferred to corresponding DIC images relying on the image alignment. With such a data setup, we developed a method that fulfils two objectives: multi-domain classification and digital staining.

5.1 Motivation for multidomain multitask learning

Image-to-image translation methods, in principle, are not designed to produce domain-invariant features suitable for multidomain image classification. Both StarGAN [22] and StarGANv2 [23] address the problem of image translation for multiple domains, but they ignore structure preservation and focus on style transfer and domain diversification. We are less interested in the variety of generated outputs and more in structure preservation. Since the StarGANv2 [23] model is excellent at generating diverse images and not concerned with structure preservation, we build our method on the original StarGAN [22] architecture. We enforce structure preservation crucial for the digital staining and keep the latent space feature domain agnostic. We add an auxiliary segmentation task and, optionally, reconstruction for a limited number of matched images to achieve this. Segmentation enforces that the network learns to generate the correct nucleus and cytoplasm shape, and reconstruction enforces reliable translation between the images of the matched domains. The segmentation masks used for training were outputs of a pre-trained network. Even though they are weak labels, they contributed to a significant improvement.

The domain-agnostic latent space is essential for us. Auxiliary classification task helps accurately classify the samples regardless of the scanning protocol. Compared to StarGAN [22], where the target domain labels are fed together with the input, we provide them directly to the generator. The auxiliary classification task forces the encoder to learn how to extract features from all three domains independently of the target staining. It results in a robust and domain-invariant representation. Despite having fewer convolutional layers aware of the target domain, such change does not significantly influence image reconstruction quality. Furthermore, in this way, the encoder is compelled to produce a uniform representation for all the domains in the bottleneck. After investigating the latent space, it was evident that the representation built by the original StarGAN was dependent on the target domain label only. With domain-agnostic latent space, the class information can be efficiently transferred from stained to unstained images, enabling us to train a model on annotated and unannotated data.

The proposed model builds on an intuition that the features necessary for staining and classification can benefit from each other once the gap between the input data domains is closed. The contributions of this chapter are as follows:

- Novel combination of image generation with auxiliary tasks such as segmentation and pair-wise reconstruction. Demonstration of their influence on the quality of generated images in fine-grained details and well-defined structures.
- Auxiliary classification and construction of a domain agnostic latent space to use the features extracted in an unsupervised manner for the classification task.
- Quantitative and qualitative evaluation comparing our method to the state-of-the-art methods, an exhaustive ablation study and discussion of the approach.

5.2 Methodology

Given a dataset $\mathcal{X} = \{X_{DIC}, X_{CV}, X_{BF}\}$, which consists of three domains, each having a variable number of labeled (L) and unlabeled (U) images, $\mathcal{X}_{DIC} = \{(x_1, y_1), \ldots, (x_L, y_L), x_{L+1}, \ldots, x_{L+U}\}$ with $x \in \mathbb{R}^{H \times W \times 3}$ being an RGB image, and $y \in \mathbb{R}^C$ is the corresponding class label, our objective is to train a model

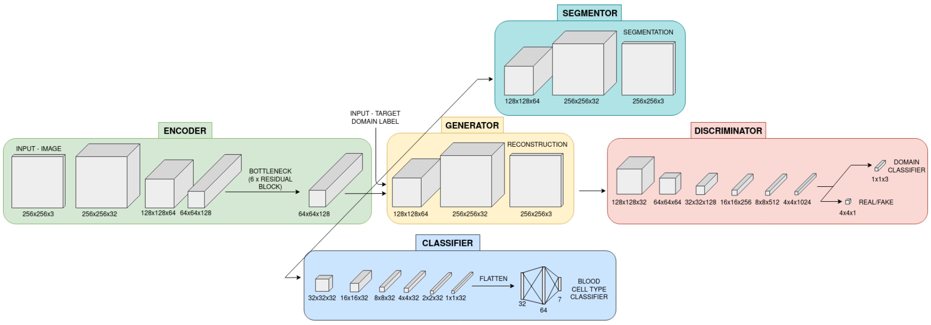

Figure 5.1: Overview of our framework: The body consists of the domain-invariant encoder and generator followed by a multi-head discriminator. The left and right wings are segmentation and classification networks, respectively.

$f(\cdot)$ which maps an input x_S from a source domain $\mathcal{X}_S$, to a domain-agnostic hidden representation h_S, before mapping it to the target domain $\hat{x}_T \in \mathcal{X}_T$ - unstained to stained image. Additionally, the model learns the mapping of the hidden representation h_S to the class label y independently on the input image domain x_S.

We propose a framework consisting of two main task modules, as depicted in Fig. 5.1. First, the digital staining task transfers the unstained input images to virtually stained ones for clinical examination. Second, a domain-agnostic classification task transfers the knowledge acquired on stained images to the unstained ones. Two auxiliary tasks are introduced to improve the quality of the virtually stained images and hence their hidden representations, namely pair-wise image reconstruction and segmentation tasks.

5.2.1 Digital staining

As shown in Fig. 5.1, the main body of our network is a module responsible for the digital staining, which consists of three main parts; encoder E, generator (decoder) G, and discriminator D. Our architecture is similar to StarGAN [22]; however, in contrast to their approach, the label of the target domain is fed into the bottleneck rather than along with the input image. We hypothesize that feeding it in the bottleneck yields a domain-agnostic feature representation. Next, we briefly explain the task of each component and the associated objective functions.

The encoder E embeds the high-dimensional input image x_S to a lower dimensional feature vector as $h_S = E(x_S; \theta_E)$, while the generator G maps the given hidden representation h_S together with the given label of the target domain c_T to the image space as $\hat{x}_T = G(h_S, c_T; \theta_G)$. We broadcast the target domain labels c_T to the middle feature map dimensions and concatenate in the channel axis with the generator's input (see Fig. 5.1). The generator G is trained to fool i) the discriminator D_{src} parametrized by θ_{C_R} by generating images that are indistinguishable from real images as

$$\begin{aligned}\mathcal{L}_{adv}(\theta_E, \theta_G) = \mathbf{E}_{x_S}[\log D_{src}(x_S; \theta_{C_R})] \\ +\mathbf{E}_{x_S, c_T}[\log(1 - D_{src}(G(h_S, c_T; \theta_G)); \theta_{C_R})],\end{aligned} \tag{5.1}$$

and ii) the discriminator $D_{cls}(\cdot)$ parametrized by θ_{C_L} which is a simple domain classifier trained to distinguish between different target domains,

$$\mathcal{L}^{f}_{cls}(\theta_E, \theta_G) = \mathbf{E}_{x_S, c_T}[-\log D_{cls}(c_T | G(h_S, c_T; \theta_G); \theta_{C_L})]. \tag{5.2}$$

Similar to [22, 138] approaches, a cycle-consistency loss is introduced to minimize the discrepancy between the input image and its recovered version as

$$\mathcal{L}_{cyc}(\theta_E, \theta_G) = \mathbf{E}_{x_S, c_T, c_S}[||x_S - G(G(h_S, c_T; \theta_G), c_S; \theta_G)||_1]. \tag{5.3}$$

The discriminators are trained to i) distinguish between real and fake images by maximizing the negative, adversarial loss of Eq. (5.1) with respect to the parameter θ_{C_R}, and ii) distinguish between different target domains as

$$\mathcal{L}^{r}_{cls}(\theta_{C_L}) = \mathbf{E}_{x_S, c_S}[-\log D_{cls}(c_S | x_S; \theta_{C_L})]. \tag{5.4}$$

5.2.2 Domain agnostic classifier

The critical component of our model is the domain-agnostic latent space which allows for more effective transfer learning. The aim is to classify unstained images of leukocytes, while the labels are provided for stained samples. Instead of directly enforcing the representation to be independent of the domain, as it could be done with a discriminator or Gradient Reverse Layer [32], we pass the target domain labels right after the bottleneck of the network, which allows the encoder to extract features relevant for the reconstruction task that are independent of the input domain (see Fig. 5.1). These features are later used for effective classification and segmentation. We broadcast the target domain labels to the middle feature map dimensions and concatenate them in the channel axis with the domain-agnostic classifier's input (DAC). To build a DAC, we feed the classifier with the latent representation h_S and minimize the cross-entropy loss,

$$\mathcal{L}_{DAC}(\theta_E, \psi) = -\mathbf{E}_{x_S, y_S}[y_S \log DAC(h_S; \psi)]. \tag{5.5}$$

5.2.3 Auxiliary tasks

We designed two auxiliary tasks to improve the quality of the translated images. First, we employ a direct pair-wise reconstruction loss for the domains where paired images are available as

$$\mathcal{L}_{rec}(\theta_E, \theta_G) = \mathbf{E}_{x_T, x_S, c_T}[||x_T - G(h_S, c_T; \theta_G)||_1]. \tag{5.6}$$

Second, we employ a segmentation task enforcing the network to focus and encode fine-grained details in the white blood cells, for example, granulation in the cytoplasm or shape of the nucleus, which is crucial for both digital staining and classification tasks. To realize this, weakly annotated images are utilized to minimize the following cross-entropy loss,

$$\mathcal{L}_{seg}(\theta_E, \phi) = -\mathbf{E}_{x_S,Y_S}[Y_S \log S(h_S; \phi)], \tag{5.7}$$

where $S(\cdot)$ is the segmentor, and Y_S is the pixel-wise annotation of the input image x_S.

5.3 Experiments

To validate the importance of auxiliary tasks and domain-agnostic latent space, we conduct experiments measuring generated image quality and classification accuracy of our model.

5.3.1 Implementation details

Our model was implemented in PyTorch 1.3. The size of the input and output image is 256x256 pixels. We present the overview of the framework in Fig. 5.1. For all the experiments, we followed the StarGAN [22] in replacing Eq. (5.1) with Wasserstein GAN [4] objective with the gradient penalty.

5.3.2 Training procedure

The training procedure of this model is relatively complex. The model is first trained as an original StarGAN, with a low learning rate of 10^{-6} for 50k iterations, to avoid mode collapse. After this unsupervised pre-training phase, the learning rate increases to 10^{-5} for the subsequent 150k iterations until the model converges and the losses saturate. After that, we add the reconstruction loss Eq. (5.6) and turn to supervised learning by adding the classification and segmentation parts of the network for the final 100k iterations.

As in standard GAN training, the generator and discriminator networks are trained competitively in a minimax fashion. During training, images from all three domains are input into the network. A single training iteration consists of four steps as described in Algorithm 1:

1. Five passes through the discriminator network (Step 3 of Algorithm 1) alternating with one pass through the encoder and generator (Step 8 of Algorithm 1), with input data consisting of both real and generated data. The discriminator outputs for each data point a vector specifying if it is a real or fake image and to which domain it belongs. The discriminator is optimized to maximize the probability of distinguishing between the domains and between real and fake samples. The following steps are done after each fifth training epoch and after the first 200 training epochs.

2. A single pass through the encoder and the domain agnostic classifier (Step 11 of Algorithm 1). The input data is passed to the domain-agnostic classifier as feature maps extracted by the encoder. The network outputs

Algorithm 1 Training procedure

```
Require: $N$: number of training iterations, $k$: batch size
for for $t$ in $1 : N$ do
    Sample $k$ images randomly from domains $\mathcal{X}_{DIC}, \mathcal{X}_{CV}, \mathcal{X}_{BF}$
    Pass through the real/fake image discriminator $D_{src}$ and domain discriminator $D_{cls}$ with input data consisting of both real and generated images
    Update image discriminator weights $\theta_{C_L}$ and domain discriminator weights $\theta_{C_R}$ according to Eq. (4) and negative, adversarial loss of Eq. (1)
    if $t\%5 = 0$ then
        if $t < 200$ then
            Pass $k$ images with their permuted labels as target domains through the $G$
            Update encoder $\theta_{En}$ and generator $\theta_G$ weights according to Eq. (1)(2)(3)
        else
            Pass through the domain agnostic classifier $DAC$
            Update encoder $\theta_{En}$ and agnostic classifier $\psi$ weights according to Eq. (5)
            Pass through the segmentation module $S$
            Update encoder $\theta_{En}$ and segmentation module $\theta_S$ weights according to Eq. (7)
            Pass $k$ images with their permuted labels as target domains through the generator $G$
            Update $\theta_{En}, \theta_G$ according to Eq. (1)(2)(3)(6)
        end if
    end if
end for
```

Table 5.1: The results of the experiments validating the order of passes in a training iteration.

		Discriminator	Discriminator	**Ours**: Discriminator
	Target	→ Classifier	→ Segmentor	→ Classifier
	domain	→ Generator	→ Generator	→ Segmentor
		→ Segmentor	→ Classifier	→ Generator
FID ↓	X_{CV}	64.504 ±1.208	67.344 ±0.874	**57.394** ±0.736
	X_{BF}	63.610 ±0.998	66.019 ±1.020	**53.439** ±0.836
RMSE ↓	X_{CV}	0.180 ±0.003	0.181 ±0.003	**0.180** ±0.004
	X_{BF}	0.129 ±0.003	0.135 ±0.003	**0.121** ±0.003
SSIM ↑	X_{CV}	0.590 ±0.006	0.589 ±0.005	**0.616** ±0.006
	X_{BF}	0.607 ±0.005	0.586 ±0.006	**0.653** ±0.008
Accuracy	-	0.901 ±0.003	**0.917** ±0.012	0.912 ±0.012
F1-score	-	0.902 ±0.015	**0.904** ±0.012	0.903 ±0.012
AUC	-	0.988 ±0.003	**0.989** ±0.004	**0.989** ±0.002

a prediction to which class the given cell belongs. The domain agnostic classifier is trained with cross-entropy loss.

3. A single pass through the segmentation module optimized with a cross-entropy loss (Step 12 of Algorithm 1). The output is a segmentation map.

4. A single pass through the generator (Step 14 of Algorithm 1). The input data consists of the images and labels of target domains (the ones to which we want to transfer our images). The output is images of the same size as the input but in the style of the target domain. The generator is optimized to minimize the probability that the discriminator can differentiate between real and fake images and minimize the distance between input and output images transferred back to the original domain. Suppose an exact match between an input and target domain is available. In that case, the generator has an additional loss term that minimizes the distance between the transferred input and its direct match in the target domain.

We validated this order of passes through the network during the optimization process with an experiment. Table 5.1 includes its results. Changing the order of forward passes can potentially disturb the training process; it happens if the pass-through classifier comes after the pass-through generator. Placing the segmentation optimization step before the generator improves the results the most. This order was motivated by the fact that we wanted to direct the attention of the generator to the segmented structures, not the other way around. Besides, the classifier should help the generator recognize the correct class and generate associated features.

The presented approach results in training a network that can digitally stain and classify white blood cells based on a ground truth from multiple domains. During inference, we use images from only one domain (namely, the unstained

Table 5.2: Comparison with state-of-the-art in image-to-image translation and an ablation study on our model, in X_{CV} domain. After disabling different network parts, we conclude that all the additional modules are necessary for the good quality of the reconstructed image and for closing the classification domain gap.

Target domain X_{CV}	RMSE ↓	SSIM ↑	FID ↓
StarGANv2 [23]	1.330±0.012	0.418±0.008	**25.821**±0.347
Pix2PixHD [123]	1.393±0.013	0.606±0.005	31.084±0.179
StarGAN [22]	0.194±0.006	0.593±0.007	78.027±0.857
Ours	**0.180**±0.006	**0.616**±0.006	57.394±0.736
StarGAN+Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV},\mathcal{X}_{BF}}$ +DAC+S	0.182±0.004	0.599±0.007	60.349±0.925
Ours -Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV},\mathcal{X}_{BF}}$ -DAC-S	0.194±0.004	0.592±0.007	79.983±0.853
Ours-Rec$_{\mathcal{X}_{CV},\mathcal{X}_{BF}}$ -Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{BF}}$-DAC-S	0.192±0.004	0.581±0.007	58.342±0.863
Ours-Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV}}$ -Rec$_{\mathcal{X}_{CV},\mathcal{X}_{BF}}$-DAC-S	0.211±0.004	0.600±0.006	71.684±0.639
Ours-Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV}}$ -Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{BF}}$-DAC-S	0.199±0.004	0.590±0.007	73.118±0.883
Ours-DAC-S	0.191±0.004	0.598±0.008	58.935±0.684
Ours-S	0.193±0.004	0.580±0.007	81.729±0.938
Ours-DAC	0.189±0.005	0.591±0.008	76.157±0.854

white blood cells) to take a step closer to the elimination of the blood samples' chemical staining and, at the same time, retain the ability to classify them and give the medical experts a possibility for to inspect the blood cells visually.

5.3.3 Comparison with state-of-the-art

In our application, there are two main objectives in mind: (1) digital staining and (2) domain adaptation. We address digital staining as an image translation problem. In our medical application of leukocyte classification, it is crucial to transfer the structure of the cells used by the pathologist as landmarks, like the nucleus, cytoplasm, and granularity, for the classification task. In this regard, the variety of generated outputs is less appreciated than structure preservation. To compare our method to the related approaches, we perform quantitative evaluation and compare the image-to-image translation results of different methods. The second objective is domain adaptation, e.g. producing domain invariant feature representation. For example, unstained images get correctly classified given the weak labels from other domains, e.g. stained images. We quantify this by comparing our approach to other domain adaptation methods regarding classification accuracy.

Image to image translation In terms of image reconstruction, we compare our model against StarGAN [22], StarGANv2 [23], and Pix2PixHD [123]. As

Table 5.3: Comparison with state-of-the-art in image-to-image translation and an ablation study on our model, in X_{BF} domain. We disable different network parts to validate their significance. The trends observed in Table 5.2 also hold here.

Target domain X_{BF}	RMSE ↓	SSIM ↑	FID ↓
StarGANv2 [23]	1.738±0.024	0.443±0.007	**16.867**±0.276
Pix2PixHD [123]	0.782±0.067	0.639±0.006	22.986±0.218
StarGAN [22]	0.158±0.005	0.569±0.008	73.725±0.895
Ours	**0.121**±0.003	**0.653**±0.008	53.439±0.836
StarGAN+Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV},\mathcal{X}_{BF}}$ +DAC+S	0.146±0.004	0.612±0.008	58.334±0.994
Ours -Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV},\mathcal{X}_{BF}}$ -DAC-S	0.160±0.004	0.566±0.006	76.333±0.947
Ours-Rec$_{\mathcal{X}_{CV},\mathcal{X}_{BF}}$ -Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{BF}}$-DAC-S	0.1343±0.005	0.619±0.008	55.745±0.739
Ours-Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV}}$ -Rec$_{\mathcal{X}_{CV},\mathcal{X}_{BF}}$ - DAC-S	0.142±0.004	0.572±0.007	74.270±0.717
Ours-Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV}}$ -Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{BF}}$-DAC-S	0.148±0.004	0.578±0.007	70.894±0.894
Ours-DAC-S	0.131±0.004	0.633±0.008	55.439±0.590
Ours-S	0.154±0.004	0.561±0.007	80.762±0.962
Ours-DAC	0.148±0.005	0.570±0.009	75.887±0.892

shown in Tables 5.2 (X_{CV} domain) and 5.3 (X_{BF} domain), our method has the lowest reconstruction error measured with RMSI and the highest SSIM of all methods. It shows that our method performed very well on the reconstruction task regarding structure preservation thanks to the auxiliary tasks of segmentation and pair-wise reconstruction (available for a limited number of matched images). However, the method falls short regarding the FID score, which reflects the perceived quality of images. It outlines the problem of choosing the right metrics when employing GANs in medical applications as discussed in [13, 104, 41]. Of course, having realistic crisp and sharp images is of high interest. However, as mentioned before, it can have severe consequences in a biomedical setting if the structures are not preserved, leading to wrong classification and diagnosis. Having this in mind, we prioritize maintaining the shapes of generated objects. We believe domain adaptation in the image space for medical applications needs an additional metric that would consider current limitations. As Cohen et al. [24] outlined, networks can hallucinate features on the generated images. Without additional metrics accounting for this specific problem of structure preservation, we cannot straightforwardly employ the models cannot be in a medical setting.

The qualitative results comparing our method and different related methods are shown in Fig. 5.2 (image translation form the domain of unstained images $\mathcal{X}_{DIC}$ to CellaVision images of the domain $\mathcal{X}_{CV}$) and Fig. 5.3 (image translation form the domain of unstained images $\mathcal{X}_{DIC}$ to Bright Field Microscopy images of the domain $\mathcal{X}_{BF}$). It is also evident that the methods are

Table 5.4: Distances between the latent space representations of either two samples from different domains with the same target domain or the same sample with different target domains.

	Source domains	Target domain	Cosine similarity ↑
StarGAN [22]	$\mathcal{X}_{DIC}$, $\mathcal{X}_{BF}$	$\mathcal{X}_{BF}$	0.349
	$\mathcal{X}_{DIC}$, $\mathcal{X}_{BF}$	$\mathcal{X}_{DIC}$	0.493
	$\mathcal{X}_{BF}$	$\mathcal{X}_{DIC}$, $\mathcal{X}_{BF}$	0.672
	$\mathcal{X}_{DIC}$	$\mathcal{X}_{DIC}$, $\mathcal{X}_{BF}$	0.619
Ours	$\mathcal{X}_{DIC}$, $\mathcal{X}_{BF}$	-	**0.782**

excellent in style preservation, like StarGANv2. These good-looking output images are impossible to distinguish from the real images, but sometimes they are completely wrong regarding the preserved cellular structure. It means they are not usable in our application, where correct classification, either by experts or an automated system, is the priority.

Domain Adaptation We compare the classification results against three models: DANN [32], DADA [87], and StarGAN [22] augmented with all the improvements we proposed, so the additional reconstruction loss, the segmentation network, and the domain agnostic classifier. Table 5.5 includes the results. Our model outperforms previous approaches in the classification of unstained white blood cells coming from domain $\mathcal{X}_{DIC}$, most efficiently closing the gap between the annotated domain $\mathcal{X}_{BF}$ and the $\mathcal{X}_{DIC}$ and $\mathcal{X}_{CV}$.

5.3.4 Ablation studies

For ablation study, we follow such notation: DAC - domain agnostic classifier, S - Segmentor and $\text{Rec}_{\mathcal{X}_i,\mathcal{X}_j}$ - additional reconstruction loss (Eq. 5.6) on domains $\mathcal{X}_i$ and $\mathcal{X}_j$.

Domain Agnostic Latent Space. To facilitate the domain adaptation process, we construct domain-agnostic latent space by feeding target domain labels after the bottleneck of the encoder, so directly to the generator. It differs from the original StarGAN approach, where the labels are provided together with the input. Feeding the target domain labels to the generator forces the Encoder to learn how to extract features from all three domains independently of the target staining, resulting in a robust and domain-invariant representation. We validate this idea by measuring the cosine similarity between feature vectors from the latent space. As shown in Table 5.4, Cosine Similarity for our approach is the highest. The Cosine Similarity in the latent space of StarGAN between two images of the same instance, but coming from different domains and having the same target domain, is significantly lower than the one measured between the same images with different target domains. It indicates the high dependence of StarGAN's representations on the target domain.

Table 5.5: Comparison with state-of-the-art in classification problem in terms of accuracy (Acc.), F1-score (F1) and AUC. The additional reconstruction loss together with changing the way the target domain label is passed to the model lead to more effective closure of the domain gap.

	StarGAN [22] + S + Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV},\mathcal{X}_{BF}}$ + DAC	Ours - Rec$_{\mathcal{X}_{DIC},\mathcal{X}_{CV},\mathcal{X}_{BF}}$	Ours
Acc. $\mathcal{X}_{DIC}$	0.825 ±0.011	0.839 ±0.008	**0.912** ±0.012
Acc. $\mathcal{X}_{CV}$	0.967 ±0.010	**0.973** ±0.008	0.970 ±0.010
Acc. $\mathcal{X}_{BF}$	0.890 ±0.011	**0.938** ±0.009	0.937 ±0.010
F1 $\mathcal{X}_{DIC}$	0.815 ±0.009	0.820 ±0.009	**0.903** ±0.012
F1 $\mathcal{X}_{CV}$	0.956 ±0.009	0.962 ±0.009	**0.964** ±0.010
F1 $\mathcal{X}_{BF}$	0.878 ±0.010	0.924 ±0.010	**0.935** ±0.010
AUC $\mathcal{X}_{DIC}$	0.958 ±0.004	0.961 ±0.003	**0.989** ±0.002
AUC $\mathcal{X}_{CV}$	0.997 ±0.004	0.997 ±0.003	**0.997** ±0.002
AUC $\mathcal{X}_{BF}$	0.990 ±0.006	**0.996** ±0.003	0.996 ±0.003

Pair-wise Reconstruction Loss. Since we have available matching pairs for the samples from $\mathcal{X}_{DIC}$, $\mathcal{X}_{BF}$, and partially $\mathcal{X}_{CV}$, in addition to the cycle-consistency loss, we use direct reconstruction loss (Eq. 5.6). Tables 5.2 (X_{CV} domain) and 5.3 (X_{BF} domain) show the effect it has on the reconstruction quality. We observe that adding direct matches between two domains ($\mathcal{X}_{DIC}$ and $\mathcal{X}_{BF}$) improves the quality of image reconstruction of the third domain ($\mathcal{X}_{CV}$). Adding the auxiliary reconstruction loss term improves the latent space of all three domains, even though it is optimized only on two of them.

Classification We performed several experiments considering the classification part of the model and included the results in Table 5.5. First, we compare against state-of-the-art: DANN [32], and DADA [93]. Next, we divide the models by manner of introducing the target domain label; for example, directly with an image or afterwards, only to the generator part of the network. Looking at the difference in accuracy score between the testing set composed of unstained images $\mathcal{X}_{DIC}$ and the same stained images $\mathcal{X}_{BF}$, we observe a consistent improvement in $\mathcal{X}_{DIC}$, together with closing the gap in between the domains. We also noticed that jointly training all the parts is the best training strategy, and the classification and image reconstruction greatly benefit from using the segmentation module.

Segmentation We evaluate the segmentation task on four classes (Background, Red Blood Cell, Cytoplasm, and Nucleus) against the ground truth labels on 545 unstained images. We obtained an IoU of 0.9139, 0.8281, 0.5902, and 0.7109, respectively. We show in Fig. 5.6, Fig. 5.7 and in Table 5.6 how it positively influences the quality of generated images by helping to preserve such features as the shape of the nucleus.

Table 5.6: Comparison of RMSE average values from domains X_B and X_C for models with and without segmentation module. We notice the biggest improvement in the artefact (ART), monocyte (MO) and smudge (SMU) classes, so the ones where the size of the cell is the greatest.

	Ours - S	Ours
ART	0.237 ±0.004	**0.196** ±0.003
BA	0.153 ±0.004	**0.141** ±0.004
EO	0.154 ±0.003	**0.131** ±0.003
LY	0.142 ±0.004	**0.120** ±0.004
MO	0.178 ±0.003	**0.131** ±0.003
SMU	0.200 ±0.004	**0.155** ±0.003
SNE	0.150 ±0.003	**0.125** ±0.003

Table 5.7: Comparison of simplified models. We show that truncation of specific tasks leads to deteriorated performance. Combining segmentation and image generation overloads the decoder which results in worse quality of images in terms of RMSE, SSIM and FID. Combining classifer and discriminator, as proposed in [81], results in worse classification rate and quality of digital staining.

	Target domain	Truncated D and DAC	Truncated G and S	Ours
FID ↓	X_{CV}	75.376 ±0.863	78.382 ±1.392	**57.394** ±0.736
	X_{BF}	70.196 ± 0.714	74.028 ±0.992	**53.439** ±0.836
RMSE ↓	X_{CV}	0.195 ±0.003	0.205 ±0.004	**0.180** ±0.004
	X_{BF}	0.129 ±0.003	0.130 ±0.003	**0.121** ±0.003
SSIM ↑	X_{CV}	0.574 ±0.006	0.567 ±0.007	**0.616** ±0.006
	X_{BF}	0.630 ±0.008	0.617 ±0.008	**0.653** ±0.008
Accuracy	-	0.897 ±0.055	0.888 ±0.128	**0.912** ±0.012
F1-score	-	0.891 ±0.013	0.884 ±0.013	**0.903** ±0.012
AUC	-	0.986 ±0.003	0.985 ±0.003	**0.989** ±0.002

Tasks truncation We performed an additional ablation study to check if we could simplify the architecture of the proposed model. We tested two other architectures: one combining the discriminator D and classifier C (as proposed in [81]), and one combining the segmentation module S and generator G. Table 5.7 shows the results and demonstrates that the truncation of a specific task leads to deteriorated performance. Combining segmentation and image generation overloads the decoder, resulting in worse image quality regarding RMSE, SSIM, and FID. Combining classifier and discriminator, as proposed in [40], leads to a worse classification rate and deteriorated quality of digital staining. For these reasons, we use several parallel tasks. The features needed to perform them complement each other without deteriorating the performance. Namely, the information about the leukocyte class will improve image generation. Suppose the extracted features differ for each class (which the classifier requires to output correct prediction). In that case, it will be easier for the generator to avoid changing the cell type during the staining process, which is crucial for diagnostic purposes. Likewise, if the extracted features are used later for segmentation, it concentrates the attention on specific structures (nucleus, cytoplasm).

5.4 Discussion

In this section, I will first discuss the annotation efficiency, then the agnostic latent space, the choice of evaluation metrics, the failure cases, the choice of the backbone network architecture and finally, the scalability of our model.

Annotation efficiency. Image translation for medical imaging is a complex problem that exceeds usual domain adaptation requirements. The exact and truthful transfer of structures is crucial for a correct outcome that could be utilized later in a medical setting. The variety of generated outputs is less appreciated than structure preservation. The auxiliary task of segmentation is beneficial in this regard. It enforces that the network learns to generate the correct nucleus shape. Even though the segmentation masks used can be considered weak, as they were automatically generated, they introduced significant improvement.

Agnostic Latent Space is one of the critical elements of the proposed model. It allows for efficient classification of the samples regardless of the scanning protocol. It was not enforced directly but built by changing where the generator obtains information to which domain the target image should belong. Despite having fewer convolutional layers aware of the target domain, such change does not significantly influence the image reconstruction quality. Furthermore, in this way, the encoder is compelled to produce a uniform representation for all the domains in the bottleneck. Without a powerful domain-agnostic representation, the generator could not reconstruct the image with only three convolutional layers. After investigating the latent space, it was apparent that the representation built by the original StarGAN was dependent on the target domain label when fed together with the input (see Table 5.4). This way, the class information can be efficiently transferred from stained to unstained images, enabling us to train a model on annotated and unannotated data.

Failure cases. For some cases, all of the existing models fail (see Fig. 5.4 and 5.5). The granulation of the cytoplasm seems to be the most problematic element to generate in a stained image correctly. It can have two possible causes: first of all, in the segmentation mask, the cytoplasm is labelled as a uniform class, discouraging any particular variations. It could be solved by adding a granulation class for the segmentation task. Secondly, heavy granulation is, to a different extent, only present in Basophil and Eosinophil classes. It may be underrepresented in the dataset. On the other hand, any physical artefacts result in the wrongly generated image as well (see the last row of Fig. 5.4 and Fig. 5.5). The factors of influence include camera focus and the illumination of the samples. Variation in one of these parameters can confuse the model, poor reconstruction quality, and incorrect classification.

Network backbone. We decided to base our model on StarGAN [22]. The StarGANv2 [23] model was designed to generate diverse images. For our application, it is only partially applicable since it is not our goal to generate diverse images but to have a faithful reconstruction. It underlines the problem of lack of structure preservation in the medical domain while using only cycle loss (as presented in [24]). We show that even a few matching pairs improve the reconstruction quality and help preserve semantic consistency. The other aspect is the style generated. We augmented the original StarGANv2 model with additional modules and trained it with and without diversification loss terms. Unfortunately, we cannot control what is encoded in the ’style code’ injected in the network’s bottleneck. It dominates the final result of the generated image. So even though the segmentation network converges, the structure is not preserved. In this case, the direct reconstruction loss alone is not enough to enforce the correct shape of the cell, although the SSIM improves with respect to the original StarGANv2. The results will remain unsatisfactory without ensuring that the ’style network’ is, in reality, only encoding the style of the given domain and not the object’s structure.

Scalability. The last point to consider is the technical side of these models. For instance, StarGANv2 [23] requires 1.5x more parameters than our model because of additional style encoding networks. On the other hand, our model requires a more sophisticated training procedure with an engineered learning rate schedule. Pix2PixHD [123] approach would not be scalable to multiple domains since it assumes training two networks per domain pair, and the number of required models would grow exponentially whenever a new domain (for example additional staining scheme) is introduced. Moreover, Pix2PixHD uses pairs of images, which would be impossible to map a different kind of staining since each sample can be stained only once. Our approach extends easily to multiple domains, which can be useful while having data with different staining protocols.

5.5 Conclusions

This chapter presents a novel method for handling a multi-domain database to digitally stain unstained microscopic images of white blood cells and build a domain-agnostic classifier. Our additional auxiliary tasks have demonstrated their effectiveness: incorporating the segmentation task enhances the digital staining process by enforcing attention on the structure transfer. Combining the cycle loss with direct reconstruction for the images where matches are available improves the image-to-image translation. Furthermore, by not providing the target label with the input image but feeding them directly to the generator and adding an auxiliary target domain classifier, our encoder learns to output domain-agnostic feature space, significantly improving the multi-domain classification results. We performed an exhaustive ablation study that supports our choice of architecture and the validity of the proposed training procedure. We compared our results to state-of-the-art methods in image-to-image translation and multi-domain classification.

For further research, combining the benefits of enforced structure transfer by incorporating the additional segmentation information with better perceptual style generation would be interesting. Since there were still some failure cases, it would be interesting to investigate if additional information in the segmentation masks, like an additional class for granulation, could help overcome this problem. To support the clinical applicability, we should consider a quality assessment of digitally stained images by medical experts. It would include their impression of the image quality and the possibility of working directly with digitally stained images of blood cells to classify them. The final improvement would include uncertainty estimation of generated images, allowing for automatic quality assessment and minimising the probability of erroneous blood cell classification. We tackle this problem in the next chapter.

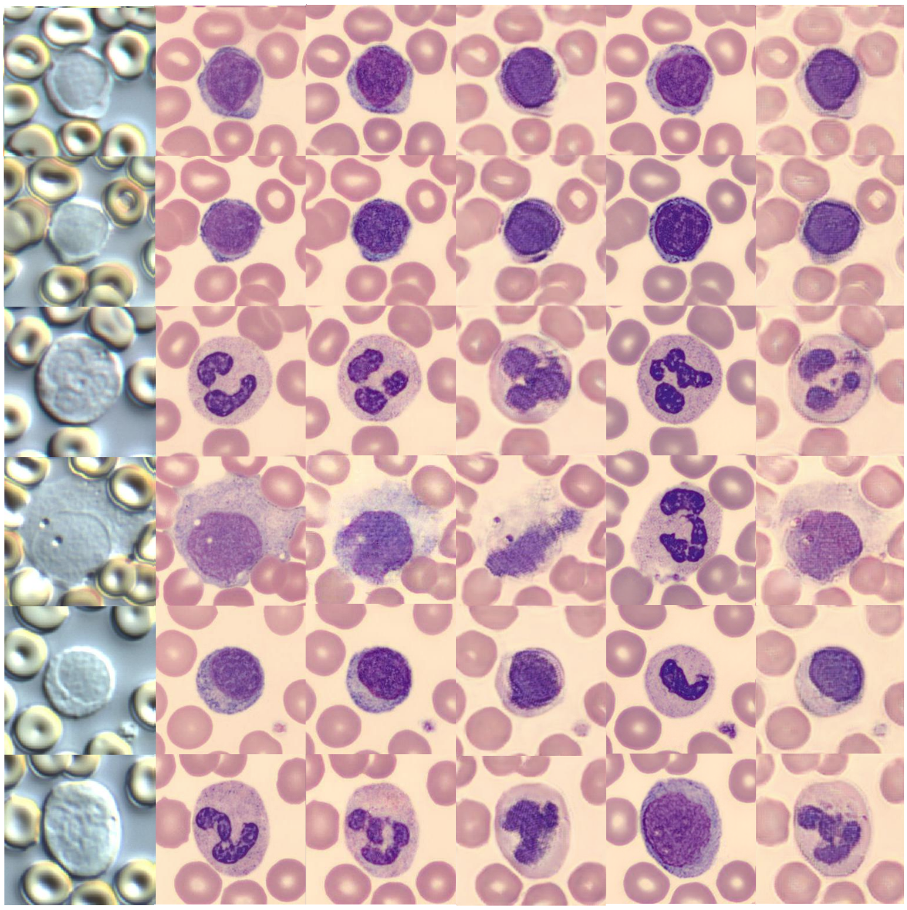

Figure 5.2: Comparison of four methods for image translation from the unstained image domain X_{DIC} to the CellaVision image of the domain X_{CV}. From the left: input, ground truth, Pix2PixHD [123], StarGAN [22] StarGANv2 [23] and Ours. Even though the images generated by StarGANv2 are incredibly crisp, it fails to preserve the cell's structure, generating the nucleus's wrong shape.

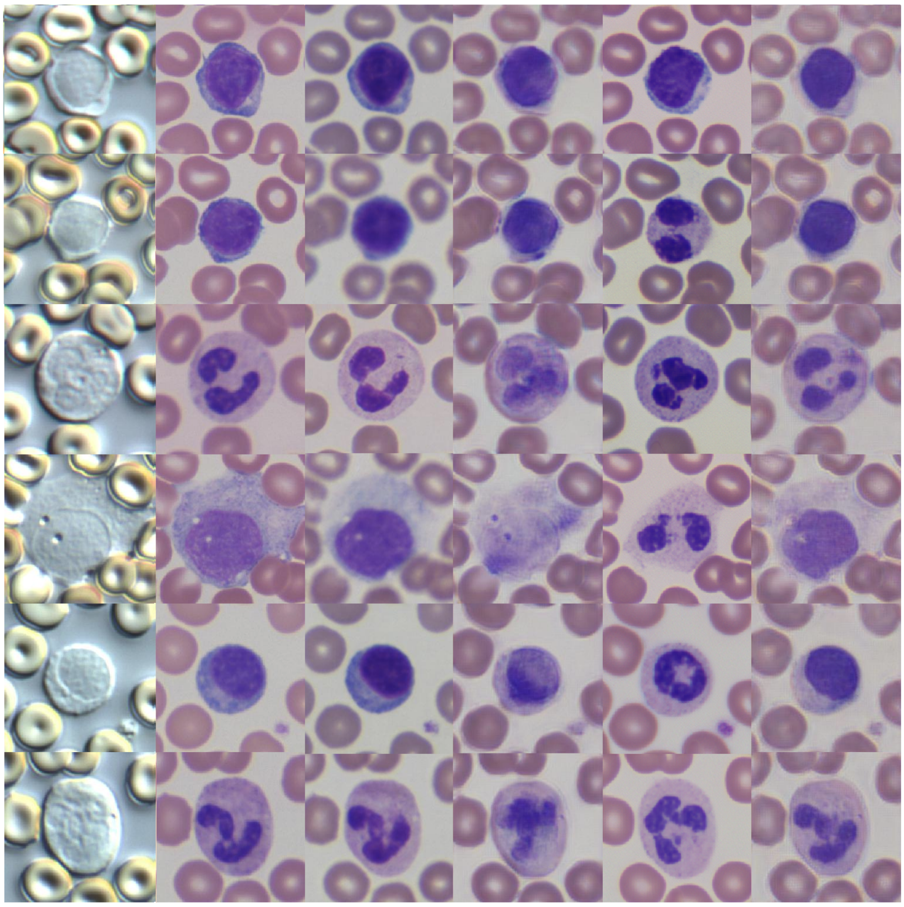

Figure 5.3: Comparison of four models for image translation from the unstained image domain X_{DIC} to the Bright Field Microscopic image domain X_{BF}. From the left: input, ground truth, Pix2PixHD [123], StarGAN [22], StarGANv2 [23] and Ours. Here as well, StarGANv2 generated crisp images with the nucleus's wrong shape. The images generated by Pix2PixHD preserve the structure, but the network could not learn the correct colour scheme.

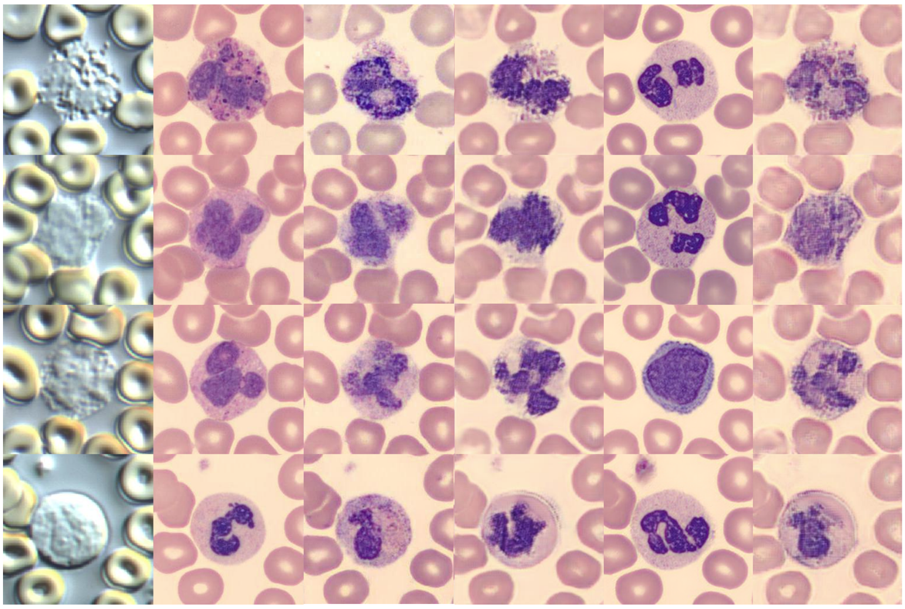

Figure 5.4: Comparison of four models on failed cases for domain X_{CV}. From the left: input, ground truth, Pix2PixHD [123], StarGAN [22], StarGANv2 [23] and Ours. For some of the cells, all four models failed.

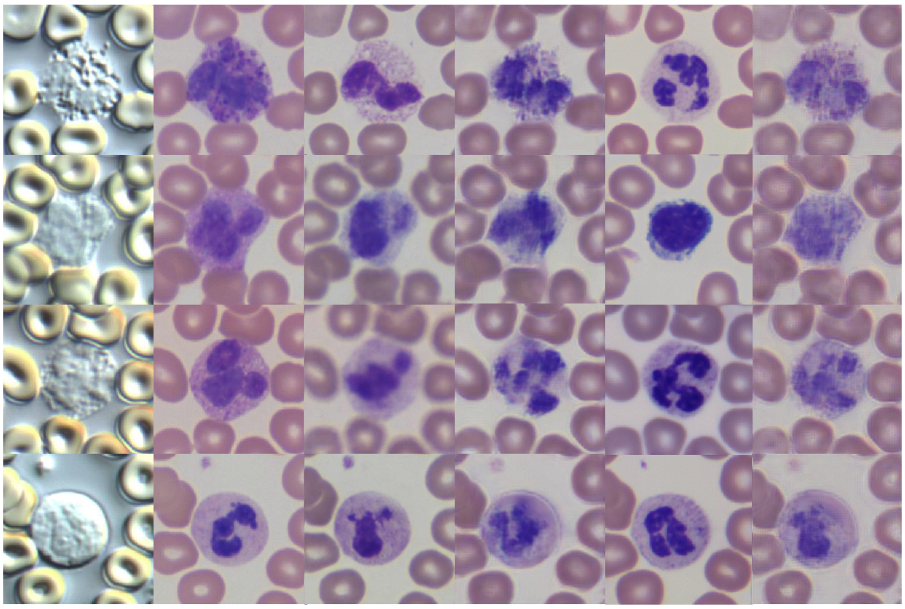

Figure 5.5: Comparison of four models on failed cases for domain X_{BF}. From the left: input, ground truth, Pix2PixHD [123], StarGAN [22], StarGANv2 [23] and Ours. For some of the cells, all four models failed.

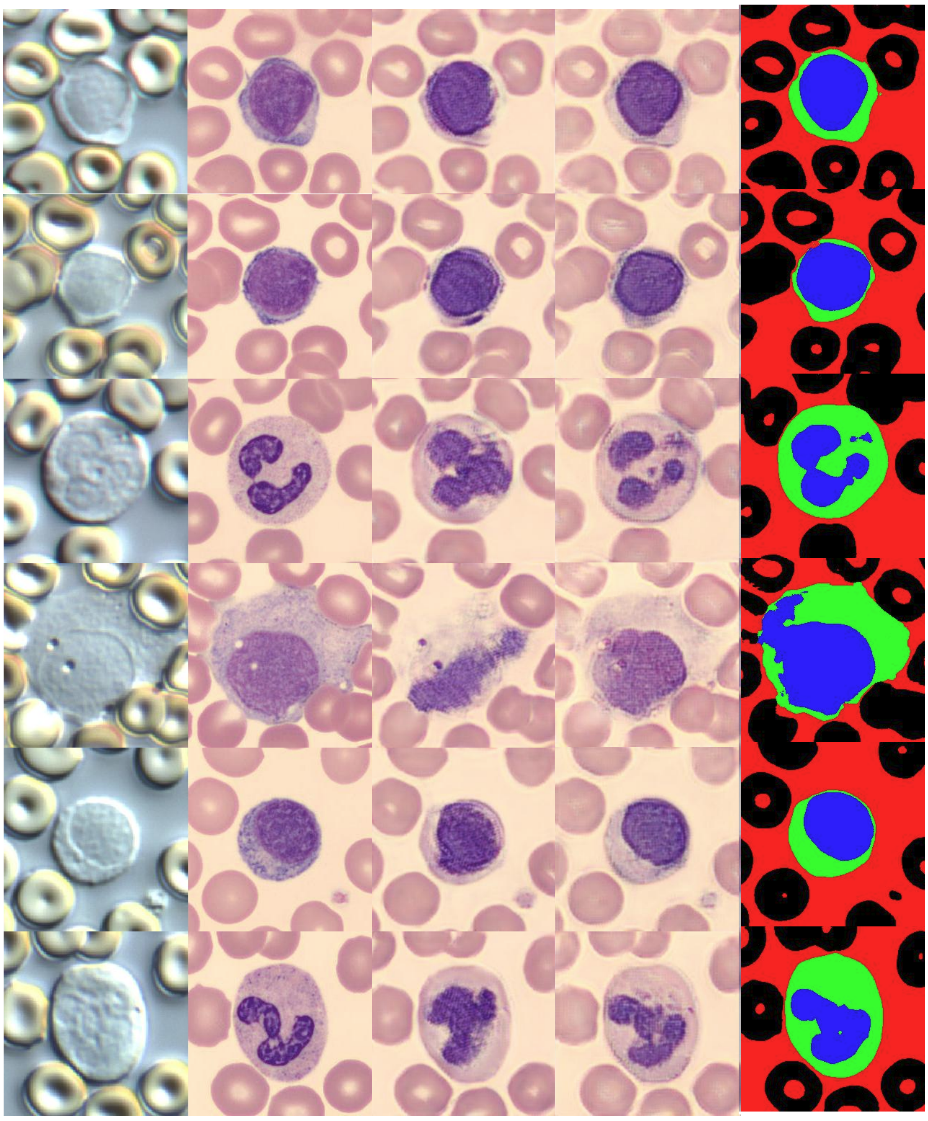

Figure 5.6: Comparison of the models with and without segmentation module in X_{CV} domain. From the left: input, ground truth, Ours without segmentation module, Ours, Segmentation output. It illustrates how the auxiliary segmentation task helps to preserve the structures on generated images. Especially the shape of the nucleus and cytoplasm is better maintained.

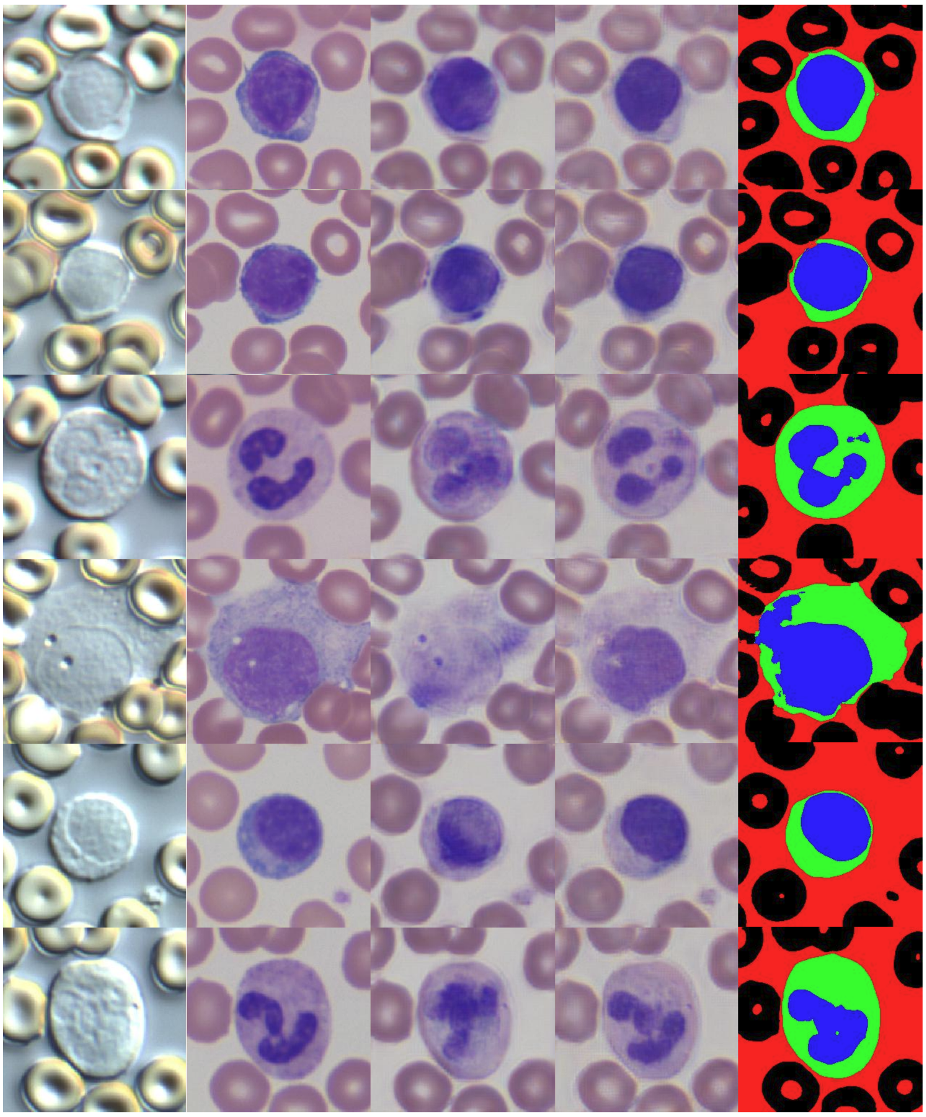

Figure 5.7: Comparison of the models with and without segmentation module in X_{BF} domain. From the left: input, ground truth, Ours without segmentation module, Ours, Segmentation output. It illustrates how the auxiliary segmentation task helps to preserve the structures on generated images. Here also, the shape of the nucleus and cytoplasm is better maintained.

Part IV

Uncertainty

6 Uncertainty estimation on disentangled latent representation

Most current state-of-the-art image-to-image translation and style transfer methods concentrate on creating increasingly realistic-looking images or diversifying the generator's output [53, 123, 22, 23]. However, in medical applications, there is another factor to consider. Changes in structures and texture between the input and generated images could alter the class label. For example, during the digital staining, an image containing a blood cell of a healthy patient could be converted to a pathological sample or, more critically, a pathological sample converted to a healthy one. In practice, a lack of preservation of certain features may lead to misclassifications and erroneous diagnoses, as shown by Cohen et al. [24], where the network hallucinates brain tumours, or as we have seen in the Chapter 5, where the model relying on standard cycle-consistency loss changes the classes of white blood cells. We want to estimate how confident the generator is during the digital staining to prevent such cases.

To successfully digitally stain a white blood cell, we need to correctly map the colour scheme of chemical staining and faithfully preserve its critical structural components, such as the nucleus's shape and the cytoplasm's colour, so that the haematologist or automated system can classify it. We developed a model consisting of two interdependent steps to fulfil these requirements. In the first step, we disentangle the style and structure of the generated images.

The disentanglement is necessary to easily trace the effect of changes in a given latent representation. Otherwise, what influences different parts of the finally generated image is unclear. In the second step, we estimate the confidence of the latent representation to recognize poorly generated samples in terms of style and structure. We disentangle the style and structure by constraining encoders differently during training. We estimate the confidence of the generated structure and style based on the robustness of the structure and style representations. We corrupt the latent representations with noise and measure the similarity between the generated samples to estimate the confidence.

We take advantage of the fact that we can define the structure in the images of white blood cells straightforwardly by considering the shape of the nucleus and cytoplasm. Therefore, we can use pseudo-segmentation masks of the cells and their intracellular structures to guide the network to generate the desired structures. Moreover, we can do it independently of generating the style. In practice, I implemented it using three different modules as depicted in Fig. 5.1. The structure module is responsible for learning the representation of white blood cells, optimized so that the segmented nucleus and cytoplasm have the same shape as those in the input image. The background module is responsible for the background generation, optimized again with segmentation masks, but this time using segmentation masks of red blood cells and the background. Finally, the style module learns the style representation. Unlike previous approaches, our method proposes a semi-supervised way of disentangling the image components [70, 139, 59, 63, 37]. Unlike Pix2Pix [53], and Pix2PixHD [123], which require matching image pairs from the source and target domains, our method does not require it, as long as the pseudo-segmentation masks for both domains are provided. The main reason for keeping separate foreground and background encoders is that we mainly need the representation for a clinically more significant foreground part of the image designating white blood cells. However, we do this without sacrificing the quality of background reconstruction, so the haematologists can still be comfortable working with such images.

This separate encoding of white blood cell let us estimate the confidence of a generation of this part of the image; thus, the score is not corrupted with background information.

6.1 Disentanglement of image components in the latent space

Our methodology consists of two steps: first, building a network that generates images based on disentangled style and structure representations and second, generating multiple samples with perturbed style and structure representations to estimate the model's confidence. The disentanglement representations are constructed with the help of segmentation masks obtained from a pre-trained U-Net. Given image $x \in X$ and domain $y \in Y$, we assume hidden latent representation s, t for style and structure, respectively. We randomly sample latent code z_s and z_t from a normal distribution for structure and style generation. Our model consists of (1) a style module, (2) a background module and (3) a WBC structure module depicted in Fig. 6.1. The style module is composed of style mapping network $M_S(z_s; \theta_{M_S})$ and style encoder $En_S(x; \theta_{En_S})$. The WBC structure module is composed of WBC structure mapping network $M_T(z_t; \theta_{M_T})$ and WBC structure encoder $En_T(x; \theta_{En_T})$. Finally, the background module is the main generation module and is composed of background encoder $En_B(x; \theta_{En_B})$, generator $G(x, s, t; \theta_G))$, and discriminator $D(x; \theta_D)$. The

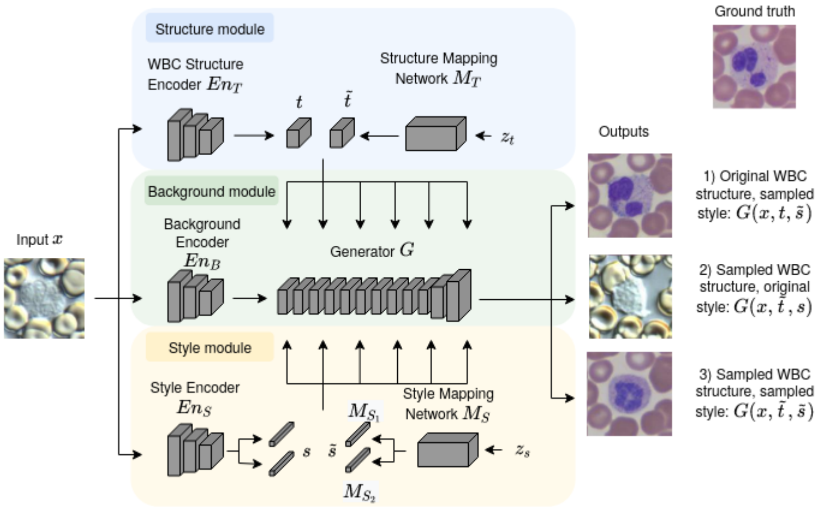
Structure module
WBC Structure Encoder En_T
Structure Mapping Network M_T
t
$\tilde{t}$
z_t
Background module
Background Encoder En_B
Generator G
Input x
Style module
Style Encoder En_S
Style Mapping Network M_S
M_{S_1}
M_{S_2}
s
$\tilde{s}$
z_s
Ground truth
Outputs
1) Original WBC structure, sampled style: $G(x, t, \tilde{s})$
2) Sampled WBC structure, original style: $G(x, \tilde{t}, s)$
3) Sampled WBC structure, sampled style: $G(x, \tilde{t}, \tilde{s})$

Figure 6.1: Overview of the architecture of our model. The network consists of three encoders, two blocks mapping sampled code into latent representations, a bottleneck and a generator.

first step of our method is building an adversarial neural network. The generator receives three different latent representations: one of the white blood cell, one of the background and one of the styles in which the sample is to be generated. We propose a model optimized with several objectives to obtain such three distinct representations.

Adversarial loss During the training, we randomly sample a target domain $\widetilde{y} \in Y = \{1, 2\}$, style latent code z_s and latent structure code z_t. They are used to generate structure code $\widetilde{t} = M_T(z_t)$ and a target style code $\widetilde{s} = M_{S_{\widetilde{y}}}(z_s)$. The $S_{\widetilde{y}}$ indicates the head of the style mapping network M_S, choosing the one that corresponds to domain $\widetilde{y}$ as indicated in Fig. 6.1. The generator G can produce three kinds of outputs. It takes an image x, latent code for style $\widetilde{s}$ and structure $\widetilde{t}$ as inputs and learns to generate:

- $G(x, \widetilde{s}, \widetilde{t})$: an image belonging to the distribution with generated style and structure,
- $G(x, s, \widetilde{t})$: an image with original style and generated structure and
- $G(x, \widetilde{s}, t)$: an image with generated style and original structure.

We constrain the model in multiple ways to generate these three different outputs. The first training objective is minimizing the classical adversarial loss on all of the three generated samples:

$$\begin{aligned}\mathcal{L}_{adv}(\theta_D, \theta_G, \theta_{M_T}, \theta_{M_S}, \theta_{En_T}, \theta_{En_S}, \theta_{En_B}) = \mathbb{E}_{x,y}[log D_y(x)]\\ +\mathbb{E}_{x,\widetilde{y},z_s,z_t}[log(1 - D_{\widetilde{y}}(G(x, \widetilde{s}, \widetilde{t})))]\\ +\mathbb{E}_{x,\widetilde{y},z_t}[log(1 - D_{\widetilde{y}}(G(x, s, \widetilde{t})))]\\ +\mathbb{E}_{x,\widetilde{y},z_s}[log(1 - D_{\widetilde{y}}(G(x, \widetilde{s}, t)))]\end{aligned} \tag{6.1}$$

where θ are the parameters of the given subnetwork, the style mapping network M_S provides style code $\widetilde{s}$ for target domain $\widetilde{y}$ and the structure mapping network M_T provides structure code $\widetilde{t}$ likely for the input image x.

Style reconstruction The second objective is constraining the style generation. Generator G utilized the style code $\widetilde{s}$ to produce the images with the original structure and generated style and both style and structure. We enforce the style reconstruction by punishing the difference between the output of style encoder En_x for generated images and the output of style mapping network for a given domain $M_{S_{\widetilde{y}}}$.

$$\begin{aligned}\mathcal{L}_{sty}(\theta_G, \theta_{M_T}, \theta_{M_S}, \theta_{En_T}, \theta_{En_B}) = \mathbb{E}_{x,\widetilde{y},z_s}[||\widetilde{s} - En_s(G(x, \widetilde{s}, t))||_1]\\ +\mathbb{E}_{x,\widetilde{y},z_s,z_t}[||\widetilde{s} - En_s(G(x, \widetilde{s}, \widetilde{t}))||_1]\end{aligned} \tag{6.2}$$

Structure reconstruction The third objective is to preserve the structural information of the input. To this end, we use pseudo-segmentation masks obtained from pre-trained U-Net and thresholding. They are obtained on stained images x_{match} and transferred to unstained images using transformations obtained by a pixel-accurate alignment between stained and unstained images. These segmentation masks are the only use of the matching image pairs produced by pixel-accurate image alignment. We define four segmentation classes: background (b), red blood cells (r), cytoplasm (c) and nucleus (n). For all the outputs, we penalize if the generated cell segmentation mask deviated from the input segmentation mask in terms of background and red blood cells. Additionally, we punish the nucleus and cytoplasm segmentation classes for the output where the white blood cell structure is to be preserved. Our objective is to minimize cross-entropy loss.

$$\begin{aligned}\mathcal{L}_{str}(\theta_G, \theta_{M_T}, \theta_{M_S}, \theta_{En_T}, \theta_{En_B}) = CE_{i \in b,r,c,n}(Seg(x_{match}),\\ Seg(G(x, \widetilde{s}, t))) + CE_{j \in b,r}(Seg(x_{match}), Seg(G(x, \widetilde{s}, \widetilde{t}))\end{aligned} \tag{6.3}$$

where CE is Cross Entropy loss, and Seg is a pre-trained segmentation network (U-Net). The gradients from this loss function are propagated through this segmentation network, but their weights are not updated. It only influences the weights of WBC structure encoder En_T, background encoder En_B and generator (G).

Diversification As we don't want to lose the possibility of generating diverse outputs, we follow StarGANv2 ([23]) and include the diversification loss where the distance between images generated with two different style codes $\widetilde{s}_1$ and $\widetilde{s}_2$, as well as two different structure codes $\widetilde{t}_1$ and $\widetilde{t}_2$ is maximized.

$$\begin{aligned}\mathcal{L}_{ds}(\theta_G, \theta_{M_T}, \theta_{M_S}, \theta_{En_T}, \theta_{En_B}) = \lambda\Big(\mathbb{E}_{x,\widetilde{y},z_s,\hat{z}_s}[||G(x,\widetilde{s}_1,t) -\\ G(x,\widetilde{s}_2,t))||_1] + \mathbb{E}_{x,\widetilde{y},z_t,\hat{z}_t}[||G(x,s,\widetilde{t}_1) - G(x,s,\widetilde{t}_2))||_1]\\ +\mathbb{E}_{x,\widetilde{y},z_s,\hat{z}_s,z_t,\hat{z}_t}[||G(x,\widetilde{s}_1,\widetilde{t}_1) - G(x,\widetilde{s}_2,\widetilde{t}_2))||_1]\Big)\end{aligned} \tag{6.4}$$

As the diversification of outputs can be an issue, we introduce an λ regularization term that allows us to weigh it appropriately.

Preserving source characteristics For the preservation of other domain invariant characteristics, we use standard cycle consistency loss:

$$\begin{aligned}\mathcal{L}_{cyc}(\theta_G, \theta_{M_T}, \theta_{M_S}, \theta_{En_T}, \theta_{En_B}) = \mathbb{E}_{x,y,\widetilde{y},z_s,z_t}(||x - G(G(x,\widetilde{s},\widetilde{t}),s,t)||_1]\\ +\mathbb{E}_{x,y,\widetilde{y},z_s}(||x - G(G(x,\widetilde{s},t),s,t)||_1]\\ +\mathbb{E}_{x,y,\widetilde{y},z_t}(||x - G(G(x,s,\widetilde{t}),s,t)||_1]\end{aligned} \tag{6.5}$$

6.2 Confidence estimation using latent representation

The second part of our method tackles confidence estimation. We use the disentangled style and structure representations to estimate how confident the network is about its generated image. We generate k samples $(g_1, ..., g_k)$ with the structure code corrupted with Gaussian noise and k samples $(h_1, ..., h_k)$ with the style code corrupted with Gaussian noise, where the standard deviation used to generate the noise can be regulated by factors α and β as multiples of the standard deviation of the latent representation: $\sigma_t = \alpha\Omega(t)$ and $\sigma_s = \beta\Omega(\widetilde{s})$.

$$\begin{aligned}\eta_{1,...,k} &\sim \mathcal{N}(0,\sigma_t^2)\\ \psi_{1,...,k} &\sim \mathcal{N}(0,\sigma_s^2)\\ g_1, ..., g_k &= G(x,\widetilde{s},t+\eta_1), ..., G(x,\widetilde{s},t+\eta_k)\\ h_1, ..., h_k &= G(x,\widetilde{s}+\psi_1,t), ..., G(x,\widetilde{s}+\psi_k,t)\end{aligned} \tag{6.6}$$

Next, we use the Mutual Information ($\mathcal{MI}$) to quantify the differences between the digitally stained target sample $x_{trg} = G(x,\widetilde{s},t)$ where the original structure is enforced and noisy generated images $(g_1, ..., g_k)$ and $(h_1, ..., h_k)$:

$$\delta = \frac{1}{k}\sum_{i=1}^{k}\Big(\mathcal{MI}(x_{trg}, g_i) + \mathcal{MI}(x_{trg}, h_i)\Big) \tag{6.7}$$

and interpret the score δ as the confidence of the network of generated image x_{trg}. We encapsulated the whole procedure as Algorithm 2.

We assume that the more robust the style and structure representations are, the more confident the network is in its generated output. The score measures

Algorithm 2 Confidence estimation procedure

```
Require: $x$: input image, $k$: number of corrupted outputs, $\alpha$: style noise regularizer, $\beta$: structure noise regularizer, $\mathcal{MI}()$: function calculating mutual information between images, $\Omega()$: standard deviation
$t \leftarrow En_t(x)$
$\widetilde{s} \leftarrow M_{S_{\widetilde{y}}}(z_s)$
$x_{trg} \leftarrow G(x, \widetilde{s}, t)$
$\delta \leftarrow 0$
for $i \leftarrow 1$ to $k/2$ do
    $\eta_i \sim \mathcal{N}(0, (\alpha(\Omega(t)))^2)$
    $g_i \leftarrow G(x, \widetilde{s}, t + \eta_i)$
    $\psi_i \sim \mathcal{N}(0, (\beta(\Omega(\widetilde{s})))^2)$
    $h_i \leftarrow G(x, \widetilde{s} + \psi_i, t)$
    $\delta \leftarrow \delta + \Big(\mathcal{MI}(x_{trg}, g_i) + \mathcal{MI}(x_{trg}, h_i)\Big)$
end for
$\delta \leftarrow \frac{1}{k} * \delta$
return Confidence score $\delta$
```

the information contained between images when their structure and style representations are corrupted. We chose the mutual information ($\mathcal{MI}$) here as it is not as biased by the structural information as Structure Similarity Index (SSIM) and not as detail-oriented as Mean Squared Error (MSE). Intuitively speaking, $\mathcal{MI}$ measures the dependence of the outputs. Therefore, after sampling outputs with corrupted structure or style representations, we measure how much information is retained between them. It captures how much the corruption of style and structure representation influences the output image, so how robust these representations are.

6.3 Implementation and experiments

Our model was implemented in PyTorch 1.7. The input and output image size is fixed to 128x128 pixels (resized from 256x256 crops). The overview of the framework is presented in Fig. 6.1. For all the experiments, we follow the StarGAN [22] in replacing Eq. (6.1) with Wasserstein GAN[4] objective with the gradient penalty. We used mixed-precision training for all of the experiments. We introduce λ term that weights the diversity part of our objective function (Eq. (6.4)) and set it to 3, based on the ablation study reported in Table 6.1. We define parameters α and β that regulate the amount of noise added to generate noisy samples and k that represents the number of noisy samples for confidence estimation. We set $\alpha = 3$, $\beta = 7$, and $k = 20$ based on ablation study reported in 6.4. We train our model for 200k iterations. The style representation is introduced to the generator using the AdaIN layer as in StyleGAN[56], and structure representation is introduced using SPADE[84] as it contains structural information.

Table 6.1: Comparison of diversity measured with LPIPS and quality values quantified with SSIM for different values of the regularization weight λ used in the Eq. (6.4). The quality deteriorates as the diversity increases. Note that for small values λ (e.g. 1 and 2), the network only generates samples belonging to the most numerous classes, namely LY and SNE. For the rest of experiments we fixed the λ value to 3 as a good balance between quality and diversity.

λ	LPIPS ↑	SSIM ↑
1	0.256 (0.254) ± 0.078	0.839 (0.829) ± 0.041
2	0.332 (0.329) ± 0.082	0.838 (0.820) ± 0.044
3	0.393 (0.391) ± 0.099	0.838 (0.812) ± 0.064
4	0.420 (0.417) ± 0.100	0.811 (0.810) ± 0.069
5	0.423 (0.420) ± 0.102	0.782 (0.773) ± 0.070

Dataset For this part of the work, we used a subset of the dataset described in section 4.4, namely the two domains: DIC and BF. We chose this subset because of the accessibility of the segmentation masks for the stained images and the complete known correspondences between images in both domains.

Evaluation In this chapter, we calculate FID in two ways: (1) for the generated images and the stained ground truth test set images (FID_{test}) and (2) for the generated images and the stained part of the train set (FID_{train}). It captures the discrepancy between learning well the train set and test set distributions while having a patient-wise split. We also use LPIPS in two ways; first, to evaluate image quality, we extract features from pre-trained AlexNet and calculate the difference between the features of the generated image and its stained ground truth version. In this case, the lower the value, the better. Secondly, we use LPIPS to evaluate the diversity of generated outputs. In this case, we generate 20 images with sampled style and structure codes and calculate the feature difference between them. In this case, the greater the LPIPS value, the more diverse the network's outputs are.

6.3.1 Image quality

First, to make sure our network generates good-quality images, we compare our results in terms of image quality to state-of-the-art style adaptation and image translation networks: Pix2PixHD, StarGANv2, CycleGAN, DRIT++ and U-GAT-IT (see Tables 6.2 and 6.3, and Fig 6.5). We chose the Pix2PixHD model to represent the upper bound for this translation problem, as it uses the matched image pairs. In more realistic scenarios, such pairs are not always available due to the costly preprocessing needed to obtain them. As our model was trained without direct reconstruction loss, we also compare our method to networks that work with unmatched data. The CycleGAN is the most widespread baseline. StarGANv2 offers the possibility of transferring between multiple domains. Although not included in our application, it is particularly interesting for haematological images, as different staining protocols could be

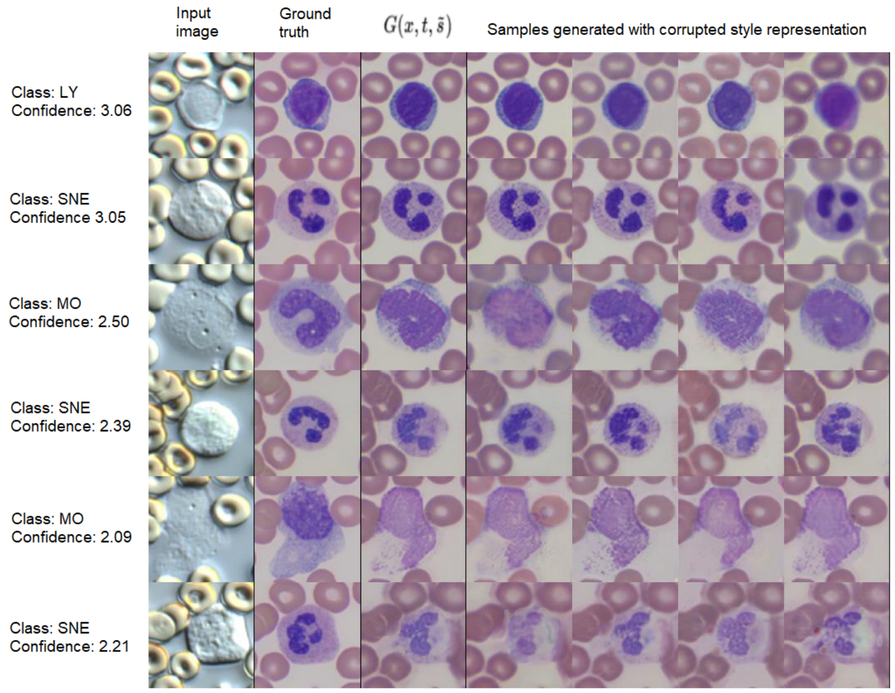

Figure 6.2: Visualisation of generated cells used for confidence estimation. Examples include cells generated with noisy style representation. We present the input to the network, its corresponding ground truth image and a generated sample with the original structure preserved and sampled style $G(x, \tilde{s}, t)$. Next, we show four samples generated with corrupted style representation (see equation 6.6). This illustrates with what samples the style part of the confidence score is calculated.

represented. The DRIT++ relies on style-structure disentanglement, and the U-GAT-IT model uses an attention mechanism to preserve significant features. StarGANv2, CycleGAN, DRIT++, and U-GAT-IT are style transfer networks; they learn the distribution perfectly, generating realistically looking images and achieving low FID scores compared with the training set. However, they do not preserve the class-specific details, resulting in changing the cell type during the generation, making the results clinically inadmissible, which can be seen in Fig. 6.4, and low FID score for the test set. For a fair comparison, we generate 20 samples with StarGANv2 and compare the average score among the 20 samples and the minimum value of MSE, LPIPS and maximum SSIM. We do it to accommodate the possibility of a well-generated sample among the diverse outputs. Still, even the best possible option does not score as well as competitors. Additionally, to the best of our knowledge, there is no way to predict a priori which sample would be most faithful to the input. Pix2PixHD is an image-to-image translation network benefiting from the direct supervision of paired images. It performs best in terms of MSE, SSIM and LPIPS; however, it depends on the pixel-wise alignment that, in practice, is a very expensive pre-

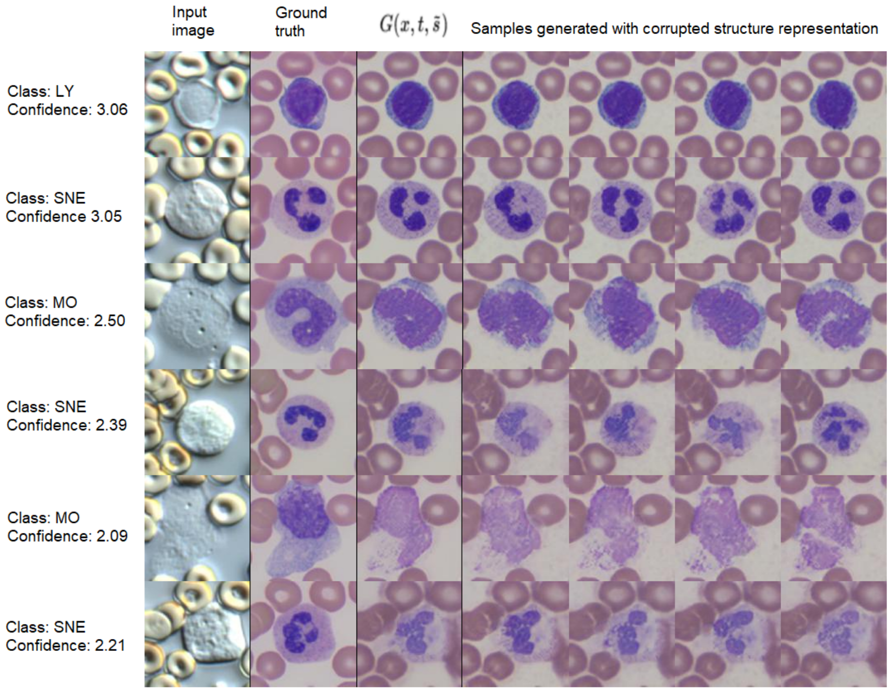

Figure 6.3: Visualisation of generated cells used for confidence estimation. Examples include cells generated with noisy structure representation. We present the input to the network, its corresponding ground truth image and a generated sample with the original structure preserved and sampled style $G(x, \tilde{s}, t)$. Next, we show four samples generated with corrupted structure representation (see equation 6.6). This illustrates with what samples the structure part of the confidence score is calculated.

processing step. Also, it does not correctly learn the distribution representation that we can measure with FID (wrt. DRIT++ when compared to the training set, and wrt. our model when compared to the testing set), and it offers very little flexibility not representing staining variations. Our model performs en par with Pix2PixHD in terms of structure preservation (SSIM), slightly worse comparing MSE and LPIPS (which is expected as we do not use the matching pairs straightforwardly but only to obtain the segmentation masks) and better when it comes to distribution modelling achieving the best FID_{test} score.

Next, we confirm we are still able to generate diverse samples. We conduct an ablation study on a regularisation term λ that decides on the magnitude of the diversity loss term. The results are presented in Table 6.1. As the regularisation term, λ, increases, the diversity of generated samples increases. For values lower than three, we have observed mode dropping, with the generator being able only to produce cells belonging to the most numerous classes: SNE and LY. For values greater than three, the quality of images starts to deteriorate, as the values of diversity term dominate the total loss. To preserve the quality of generated images but not give up on the possibility

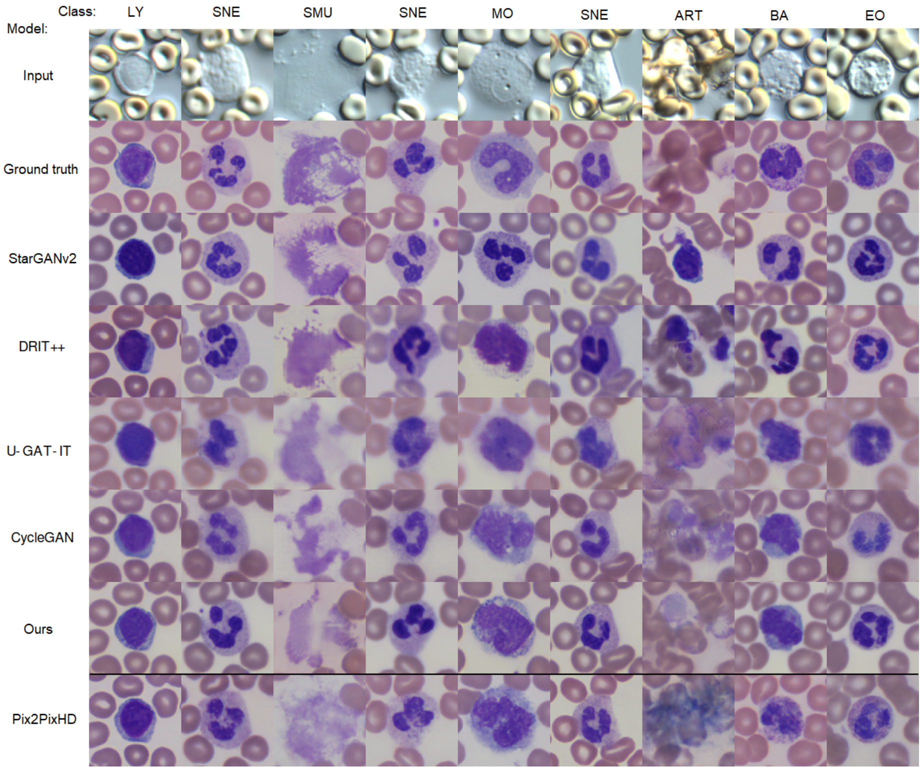

Figure 6.4: Comparison of the images generated by our model to related approaches StarGANv2 [23], DRIT++ [70], U-GAT-IT [61], CycleGAN [138] and Pix2PixHD [123]. We present the translation results on all classes in the test set. StarGANv2, DRIT++, U-GAT-IT and CycleGAN change the shape and intracellular structures of the white blood cell, making it impossible to differentiate classes of generated outputs. Pix2PixHD and our model successfully stain the samples from well-represented classes such as LY, SNE, SMU, MO and ART, but both fail for underrepresented BA and EO.

of generating diverse samples, we set the λ to be equal to three. We also investigate diversity qualitatively. Fig. 6.2 and 6.3 present the cells with sampled style and structure. We can see slight differentiation in colours and sharpness between images in the sample styles, representing our training set. However, it also happens that the shape of the nucleus differs slightly, especially when it is not visible on the unstained input image. The main reason to include the corrupted style representation in the confidence score is that when there is not enough structural information available, the network uses the style code to compensate for it.

We validated our style-structure disentanglement by training a network without structure preservation loss (the structure modelling component is not optimised). We report the results in Tables 6.3 and 6.2. We evaluate it by comparing the ground truth stained cell with the images generated with the original image structure code. We notice that the SSIM of a network trained

Table 6.2: Comparison of the quality of generated samples with other models in terms of MSE and SSIM. We report Mean (Median) ±Std for all values. We mark the best result in bold and underline the second best. Our model performs similarly to Pix2PixHD, although we do not rely on direct reconstruction with paired images.

Model	MSE ↓	SSIM ↑
CycleGAN	0.007 (0.006) ±0.004	0.798 (0.821) ±0.081
DRIT++	0.009 (0.008) ±0.006	0.774 (0.791) ±0.079
U-GAT-IT	0.007 (0.006) ±0.004	0.787 (0.802) ±0.072
StarGANv2-min/max	0.008 (0.007) ±0.003	0.743 (0.751) ±0.061
StarGANv2-average	0.021 (0.016) ±0.013	0.596 (0.600) ±0.108
Ours (without Eq. 6.2)	0.013 (0.013) ±0.004	0.595 (0.596) ±0.045
Ours (without Eq. 6.3)	0.009 (0.008) ±0.004	0.755 (0.770) ±0.066
Ours (without Eq. 6.4)	0.008 (0.007) ±0.003	0.708 (0.716) ±0.057
Ours	0.0044(0.004) ±0.003	0.838 (0.812) ±0.064
Pix2PixHD	**0.003 (0.003) ±0.003**	**0.846 (0.869) ±0.074**

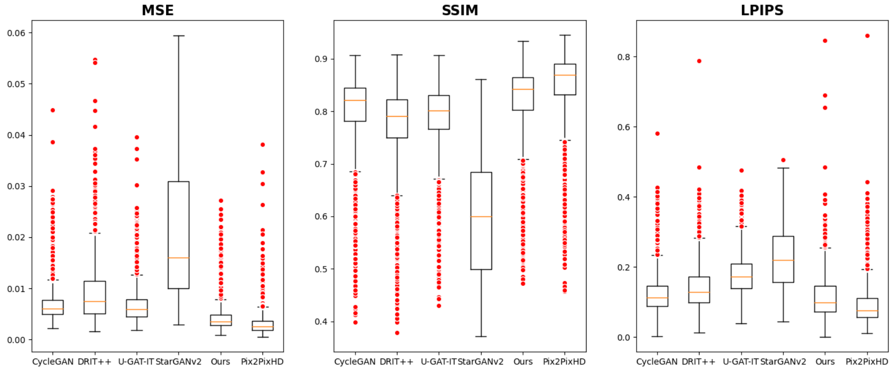

Figure 6.5: Box plot presenting image quality results comparison in terms of MSE, SSIM and LPIPS, with median marked in orange and outliers in red. The Pix2PixHD model defines our upper bound, while our method outperforms all the baselines trained on unpaired data.

without the additional loss term is comparable to the original StarGANv2 and significantly lower than when the structure preservation loss term is used, proving the necessity for this part of the model.

Finally, we validated the two loss terms responsible for learning the style distribution in terms of style code and style diversification. We report the numbers in Table 6.3. Without the style learning component (Eq.6.2), the network cannot correctly represent a sample's staining, resulting in poor performance in terms of MSE, SSMI and LPIPS. On the other hand, without the diversification term (Eq.6.4), the staining distribution is not represented, as reflected in very poor FID scores on both the training and test sets.

Table 6.3: Comparison of the quality of generated samples with other models in terms of LPIPS and FID. We calculate FID both with the train and test set. We report Mean (Median) ±Std for all values. We mark the best result in bold and underline the second best.

Model	LPIPS ↓	FID_{train} ↓	FID_{test} ↓
CycleGAN	0.125 (0.112) ±0.059	126.172	146.274
DRIT++	0.143 (0.129) ±0.063	**80.695**	203.292
U-GAT-IT	0.178 (0.173) ±0.058	265.851	301.027
StarGANv2-min/max	0.141 (0.131) ±0.054	110.640	218.041
StarGANv2-average	0.222 (0.217) ±0.082		
Ours (without Eq. 6.2)	0.206 (0.193) ±0.077	434.123	413.562
Ours (without Eq. 6.3)	0.166 (0.155) ±0.055	231.666	245.876
Ours (without Eq. 6.4)	0.176 (0.160) ±0.081	536.000	629.508
Ours	0.115 (0.099) ±0.062	152.011	**134.854**
Pix2PixHD	**0.090 (0.076) ±0.051**	127.425	150.649

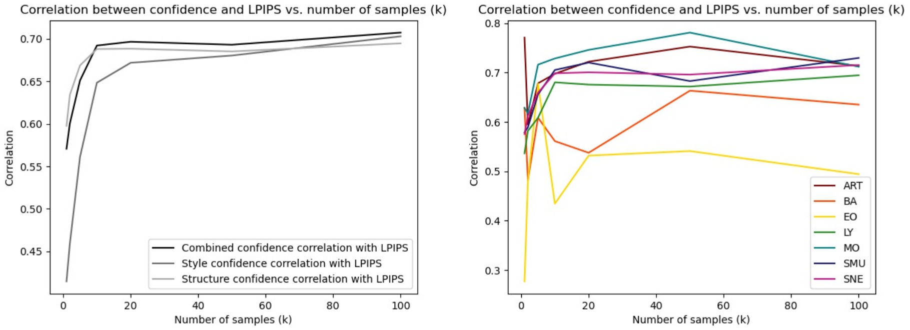

Figure 6.6: Line plots depicting the influence of the number of samples on the correlation of confidence and LPIPS. The overall correlation increases while more samples are drawn. However, looking more closely at the class dependence (plot on the right), we can see that this is the trend of the most numerous classes (LY, SMU, SNE). Classes with fewer samples (MO, ART) peak at 50 samples and decrease when 100 samples are drawn. In the case of underrepresented classes (BA, EO), the correlation is much weaker and unstable.

6.3.2 Confidence estimation

We generate multiple samples with corrupted style and structure representation to estimate the confidence. We present some examples in Fig. 6.2 and 6.3. We can see changes introduced by injecting the Gaussian Noise into the latent representation. Our intuitive assumption is that the more robust the latent representation is to the noise, the more mutual information ($\mathcal{MI}$) is retained between corrupted samples. Therefore, by measuring the $\mathcal{MI}$ between our target sample (one with stained style and original structure) and its noise-

Table 6.4: Correlation of confidence (δ from Eq. 6.7) and LPIPS. We show how the correlation value changes with increasing noise corruption of style component α, structures component β (see Eq. 6.6) and the number of generated sample k. The more samples are drawn for the estimation, the stronger the correlation between confidence value and LPIPS score.

$\alpha \backslash k$	2	5	10	20	50	100
1	0.490	0.567	0.630	0.640	0.644	0.660
3	0.459	0.561	0.648	0.672	0.680	0.703
5	0.392	0.503	0.612	0.652	0.675	0.704

$\beta \backslash k$	2	5	10	20	50	100
1	0.592	0.610	0.615	0.618	0.609	0.617
3	0.637	0.656	0.667	0.669	0.661	0.671
5	0.646	0.671	0.681	0.682	0.676	0.686
7	0.645	0.667	0.687	0.688	0.682	0.692
10	0.634	0.669	0.688	0.688	0.685	0.695

Table 6.5: Comparison of the image quality and the absolute value of correlation between confidence (δ from Eq. 6.7) and image quality metrics per class.

Image quality	MSE ↓	SSIM ↑	LPIPS ↓
ART	0.019	0.829	0.155
BA	0.015	0.842	0.155
EO	0.021	0.808	0.163
LY	0.017	0.828	0.142
MO	0.017	0.830	0.152
SNE	0.018	0.830	0.149
SMU	0.018	0.824	0.147

Correlation	MSE vs. δ	SSIM vs. δ	LPIPS vs. δ
ART	0.282	0.378	0.709
BA	0.620	0.449	0.636
EO	0.596	0.655	0.487
LY	0.421	0.439	0.693
MO	0.519	0.579	0.710
SNE	0.534	0.429	0.732
SMU	0.447	0.449	0.713

corrupted versions, we can quantify the confidence of our structure encoder and style mapping network. To evaluate the confidence estimation quality, we first correlate it with image quality score MSE, SSIM and LPIPS values, calculated for the generated stained image and the ground truth stained image. The results are presented in Table 6.5. The most important metric, in this case, is LPIPS, as it captures the similarity of the deep features. As shown by [133], it cannot be fooled by changes in sharpness, contrast or noise. The Pearson Correlation Coefficient between the confidence δ and LPIPS score across all the classes is 0.71. We can see a higher correlation between the confidence score and with LPIPS metric in the well-represented classes, such as LY, SMU and SNE. The correlation is lower for underrepresented classes, which is to be expected, as the few samples present in the test set do not allow fair estimation. We also investigated the correlation between MSE and SSIM metrics, but it was quite weak. The possible reason for it is that MSE is too fine-grained to be representative, and SSIM is, by definition, susceptible to changes in brightness and contrast, which occur when sampling the style (see Fig. 6.2 and 6.3).

Secondly, we perform an ablation study varying the number of samples k needed and the scaling of the standard deviation (α for structure and β style from section 6.2) of the Gaussian noise needed for confidence estimation. We investigate the influence of the structure standard deviation scaling factor β varying the values from 1 to 10 and the style standard deviation scaling factor α ranging from 1 to 5. The greater values introduced too much corruption for the results to be analysed. We present the results in Table 6.4. As the number of samples k increases, so does the Pearson Correlation Coefficient, and this trend is independent of the standard deviation of injected noise. The best values are obtained with $k = 100$. However, sampling more than 100 samples per data point is time-wise not feasible. We analyzed the influence and dependence of the number of samples on the correlation. Fig. 6.6 shows the results. With 10 to 20 samples, the correlation almost reaches its maximum value for 100 samples, and the increase between 20 and 100 can be neglected. It shows that using 10 samples per data point, which is less computationally demanding, is enough to obtain a reliable estimation. Also, the optimal number of samples depends on how well the given class was during training. For underrepresented classes (BA, EO), the correlation is unstable independently of the number of drawn classes (also seen on scatter plots in Fig. 6.7). Nevertheless, this could potentially change for more complex pathological cell classes. Therefore, we set the parameters $\alpha = 3$, $\beta = 7$ and $k = 20$ (due to time efficiency).

6.3.3 Latent space

To better understand what kind of information is encoded in the structure encoder En_T and the style encoder En_S we plot T-SNE embeddings in Fig. 6.8 of both structure and style representations extracted from the test samples. Although they are not clustered w.r.t. the ground truth classes, we can see some agglomerations of SMU and MO in the structural representation and SMU in the style representation. It is promising because these cell types differ significantly in structural information. Smudges do not have an apparent nucleus, and Monocytes are significantly bigger than the other cell types. Obtaining a more pronounced clustered representation w.r.t. ground truth classes cannot be expected. A correct classification would need both, style information, such as the colour of the cytoplasm and granulation, and structural information, such as the shape of the nucleus and size of the cell.

6.3.4 Discussion

As promising as the approach is, certain limitations are to be considered. The first and most significant one is the definition of the structure. We used the help of a segmentation network to specify components of the image. Potentially, the approach could work just as well for different images, as long as the structure is as easily definable as on the microscopic images, where the perspective and lighting are fixed. There are also unsupervised ways to disentangle structure and style, but since the method in this paper is significantly dependent on the disentanglement quality, any hindrance on that side would undoubtedly be reflected in the proposed confidence score. Following this thought, we could conclude that better quality segmentation masks, with additional classes, for example, granulation and nucleoli (a small structure inside the nucleus), would lead to improved confidence estimation.

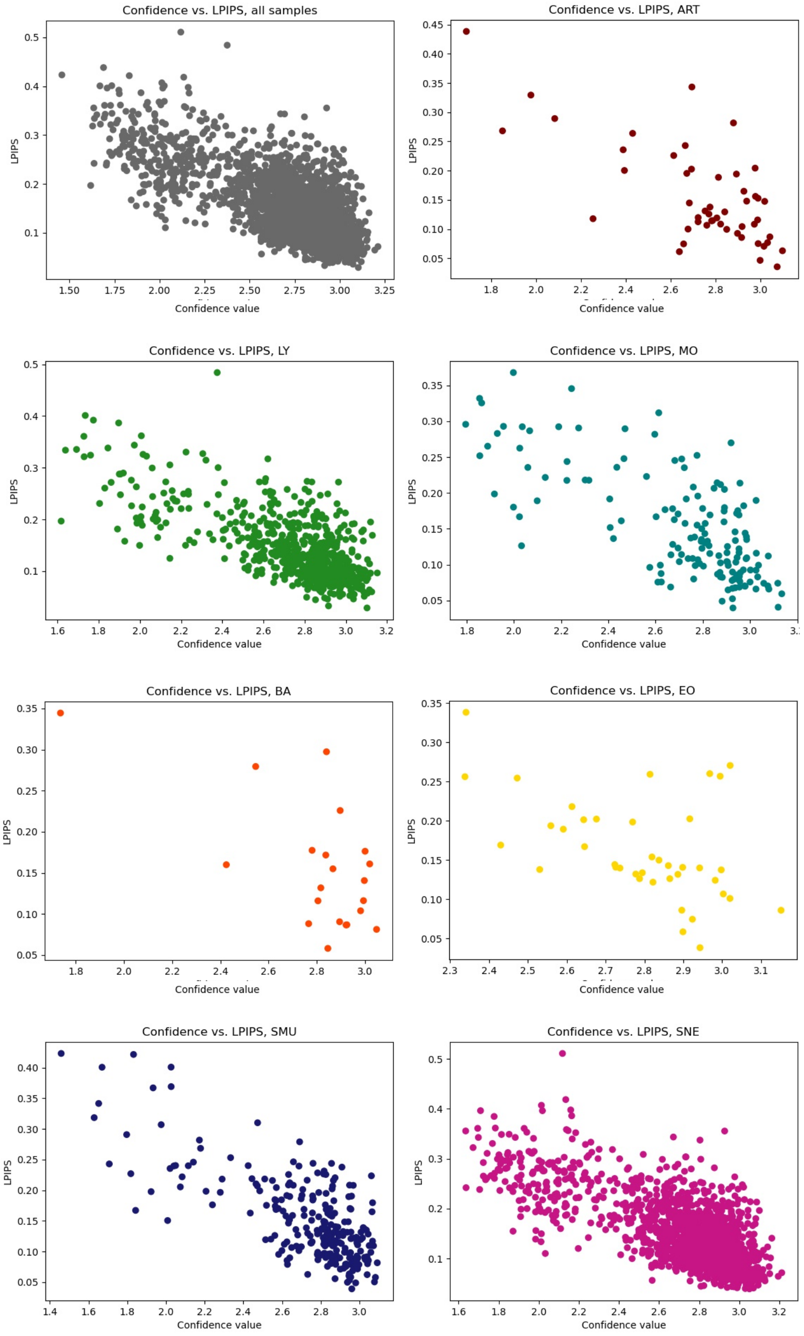

Figure 6.7: Scatter plots showing the correlation between LPIPS and confidence score δ. As the overall correlation is 0.7, we could consider all the generated samples with $\delta < 2.50$ low quality. From the detailed class-specific plots, we can see it would work for well and middle-represented classes (ART, LY, MO, SMU, SNE), but the correlation is practically nonexistent in the unrepresented classes (BA, EO).

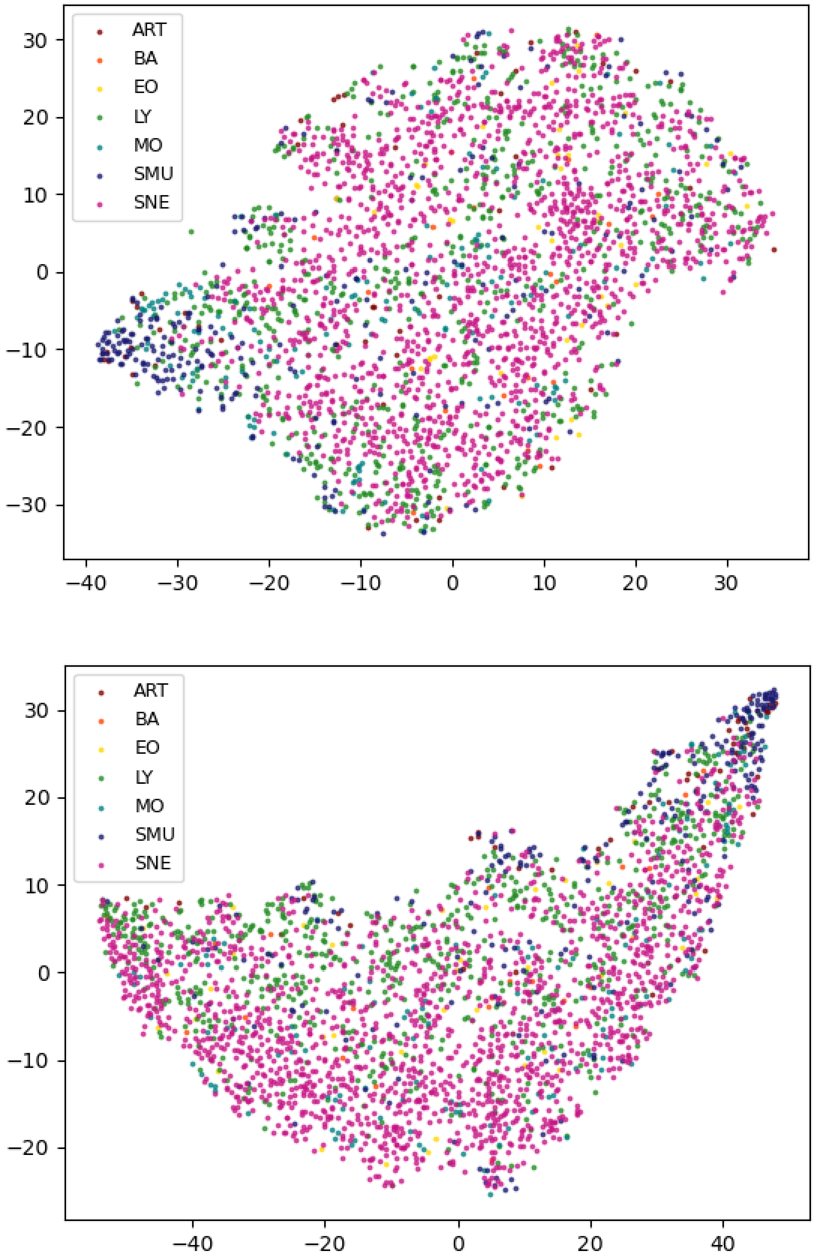

Figure 6.8: TSNE-embbeding images of Structure Encoder (top) and Style Encoder (bottom).

7 Corrupting latent representation with noise

After the encouraging results on the white blood cell dataset described in Chapter 6, we set to investigate whether the concept also holds for a less constrained problem. To this end, we let go of the assumption that we can disentangle the latent space to style and structure representations. Additionally, we decide to test the idea on different datasets containing CT and MRI scans. Visual adaptation of CT images to MRI is a problem essentially the same as digital staining of blood cells, which remains of primary interest in this thesis: the same object is captured with different modalities, and preserving its anatomical features while translating the image from one domain to another is critical.

CT-MRI translation has also been addressed as an image-to-image translation problem [19, 135, 27, 20, 33, 128]. Some of the approaches concentrate on shape consistency and the preservation of anatomical features [130, 135, 27, 33] and others proposed a multimodal approach to address the scalability concerns [126, 50, 103]. Nevertheless, determining when a generative adversarial network can faithfully preserve the structure remains the main challenge. This issue considerably impacts the medical field, where generated images with fabricated features have no clinical value. Upadhyay et al. [120] tackled this problem by predicting the output images with the corresponding aleatoric uncertainty and then using it to guide the GAN to improve the final output. This method requires many changes in the optimization process (additional loss terms) and network architecture (additional output). They showed that uncertainty guidance improves the generated image quality. However, they did not address the point that in medical imaging, the visual quality of images does not always transfer to the performance on a downstream task. Our goal was to examine this problem from a different perspective and test the hypothesis: the more robust the image representation, the better the quality of the generated output and the result.

Similarly to the method introduced in Chapter 6, here we aim to estimate

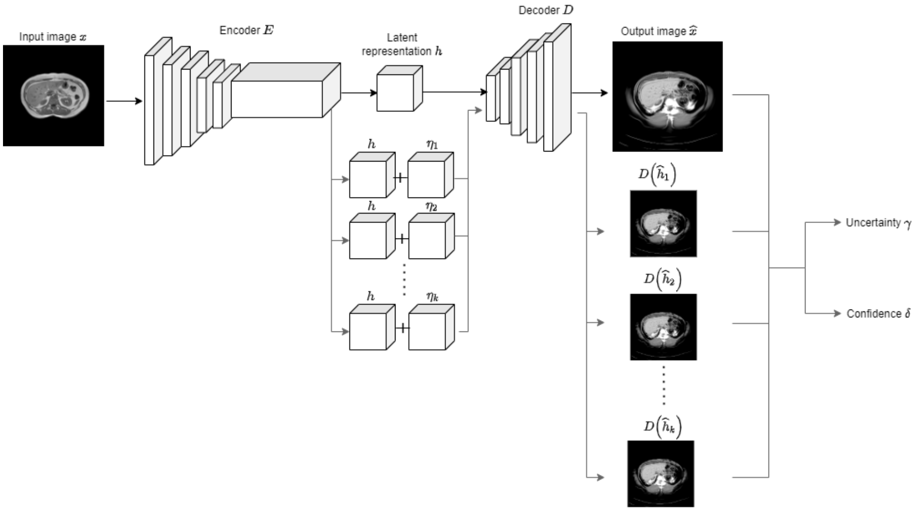

Figure 7.1: Method overview. Multiple noise injections into the latent space enable the generation of multiple outputs.

the confidence of the latent space representation. However, this time using the whole image, omitting the disentanglement step. In such a case, we opt to use a binary mask on the latent space representation to filter out the changes in image space that do not concern the object. It is trivial to do on CT and MRI images, as the background is black. We aim to estimate the confidence of the generated structure in the target domain and investigate it in the context of an end task - segmentation.

7.1 Method of injecting the noise into latent representation

We design a method to test further the assumption that the stronger the latent representation, the better the quality of a generated image. To check the validity of this statement, we again corrupt the latent representation of an image with noise drawn from a normal distribution and see how it influences the generated output image. However, we do not have access to the pseudo-segmentation masks this time, so we use the whole latent representation, as depicted in Fig. 7.1. In other terms, given an image $x \in X$, domain $y \in Y$ and a Generative Adversarial Network G, we assume a hidden representation $h = E(x)$ with dimensions n, m, l, where E stands for the encoding part of G and D for the decoding part. We denote the generated image as $\hat{x}$. Next, we construct k corrupted representation latent codes $\hat{h}$, adding to h noise vector η:

$$\eta_{1,..,k} \sim \mathcal{N}(0, \alpha\sigma^2_{h_{1,...,l}}) \tag{7.1}$$

where $\sigma^2_{h_{1,...,l}}$ is channel-wise standard deviation of input representation h. As opposed to the previous method in Section 6.2, taking the standard deviation channel-wise helps to control the noise level. We can further adjust it with factor α.

Another additional step is eliminating the background noise before the noise injection using the *bin* operation: masking it with zeros for all the channels where the output pixels are equal to zero so they do not contain any information.

$$\begin{gathered} bin(h) = h[\hat{x} > 0] \\ \hat{h}_{1,...,k} = h_1 + bin(h_1)\eta_1, ..., h_k + bin(h_k)\eta_k \end{gathered} \tag{7.2}$$

Having multiple representations for a single input image, we pass them to decoder D and generate multiple outputs:

$$\hat{x}_{1,...,k} = D(\hat{h}_1), ..., D(\hat{h}_k) \tag{7.3}$$

Next, we use the multiple outputs to quantify the uncertainty connected with the representation of a given image. We calculate two scores: the variance (the average of the squared deviations from the mean) γ of our k generated images

$$\gamma = Var(\hat{x_1}, .., \hat{x_k}) \tag{7.4}$$

and, as in the previous section 6.2, the Mutual Information (MI) between the multiple outputs and our primary output $\hat{x}$ produced without noise injection.

$$\begin{gathered} MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log(\frac{p(x,y)}{p(x)p(y)} \\ \delta = \frac{1}{k} \sum_{0}^{k} MI(\hat{x}, (\hat{x_i})) \end{gathered} \tag{7.5}$$

We interpret the γ and δ as the measures of the representation quality. The variance γ can be considered an uncertainty score - the higher the variance of generated outputs with the corrupt representations, the more uncertain the encoder is about producing representation. On the other hand, the MI δ score can be interpreted as a confidence score, quantifying the information preserved between the original output $\hat{x}$ and the outputs produced from corrupted representations $\hat{x_1}, ..., \hat{x_k}$. We calculate the MI based on a joint (2D) histogram, with a number of bins equal to $\lfloor \sqrt{n/5} \rfloor$, where n is a number of pixels per image as proposed by [16].

We conducted several experiments using state-of-the-art architectures to demonstrate the effectiveness of our proposed method and confirm our hypothesis that the stronger the latent representation, the better and more reliable the image quality - also using the whole images, and not only after disentangling the latent space, as described in Chapter 6. We evaluate this method on two publicly available datasets, namely CHAOS [58] and LiTS [10] datasets.

7.2 Network architectures and implementation details

TarGAN As our primary baseline, we use the TarGAN [19] network, which uses a shape-consistency loss to preserve the shape of the liver while translating an image from one modality to another. We trained the model for 100 epochs. We kept all the parameters unchanged as in the official implementation provided by the authors of TarGAN. We use PyTorch 1.10 to implement all the models and experiments. During inference, we constructed $k = 10$ corrupted representation with noise level $\alpha = 3$ and used them to evaluate our method.

UP-GAN We adopted the UP-GAN network from [120] to run on an unpaired dataset as shown in [118]. UP-GAN uses an uncertainty-guided loss along the standard cycle loss during training. The uncertainty loss defined for UP-GAN was used in every component of cycleGAN - identity loss and cycle loss for the training of both generators. We kept the learning rate at 10^{-4} for T1 to T2 transfer and at 10^{-3} for CT to T1 and T2 transfer. We tuned the hyperparameters in the following manner: 0.5 for each discriminator loss. At the same time, the generators had a factor of 1 with their cycle losses, 0.01 with the uncertainty cycle loss and factors of 0.1 and 0.0015 with identity losses. We trained all three models for 100 epochs.

Datasets. We use data of each modality (CT, T1 and T2) from 20 patients provided by the publicly available CHAOS19 dataset [58]. We randomly selected 50% of the dataset as the training set and used the rest for testing. We followed [19] in setting the liver as the target area, as the CT scans in the CHAOS19 dataset only have liver annotations.

Additionally, we used LiTS [10] dataset to evaluate our method on the pathological samples. The dataset contains CT scans of patients with liver tumours and corresponding segmentation masks. We resized all images to the size of 256×256.

7.3 Identification of uncertain parts of synthesised images

First, we conduct a sanity check experiment by blacking a random 50×50 pixel patch from the input images (we refer to it as perturbed input) and measuring the proposed uncertainty and confidence scores on the corresponding synthesised images. Table 7.1 reports the mean, median and variance of both uncertainty score (γ in eq.7.4) and confidence score (δ in eq.7.5) on both the original and perturbed images. Perturbed input has significant variance and low confidence compared to the original input. Fig. 7.2 shows such a case, where the confidence scores of the perturbed corrupted images are much lower than the corresponding ones for the original images. It demonstrates the effectiveness of our proposed method in detecting uncertain parts of synthesised images. The results suggest the possibility of finding a potential confidence threshold to eliminate uncertain synthesised images; however, such a threshold would always be dataset-specific. Surprisingly, the model could synthesise perturbed images hallucinating and replacing the masked regions with reasonably healthy tissues. Nevertheless, the uncertainty heatmaps captured the missing parts as shown in Fig. 7.3.

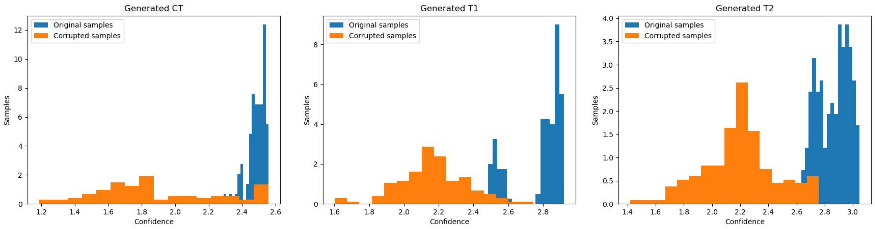

Figure 7.2: Histograms of confidence values for original and corrupted by blacking 50x50 pixel square images. We can see a significant drop in confidence for the corrupted inputs in all target modalities.

Table 7.1: The variance of multiple outputs with corrupted representation increases as we introduce a black patch on the input.

	Variance (γ)	
Source → Target	Original input	Perturbed input
CT → T2	0.018 (0.018)±0.002	0.020 (0.020)±0.002
T2 → CT	0.016 (0.016)±0.001	0.020 (0.020)±0.003
T1 → T2	0.002 (0.002)±0.000	0.003 (0.003)±0.001
T2 → T1	0.005 (0.004)±0.001	0.010 (0.012)±0.005
T1 → CT	0.018 (0.018)±0.002	0.019 (0.019)±0.003
CT → T1	0.057 (0.057)±0.003	0.068 (0.067)±0.007
	Confidence (δ)	
Source → Target	Original input	Perturbed input
CT → T2	2.261 (2.274)±0.052	1.940 (1.953)±0.094
T2 → CT	2.733 (2.741)±0.035	2.272 (2.168)±0.382
T1 → T2	2.028 (2.002)±0.070	1.823 (1.836)±0.174
T2 → T1	2.265 (2.257)±0.036	1.929 (1.864)±0.230
T1 → CT	2.774 (2.797)±0.059	2.112 (2.083)±0.381
CT → T1	2.181 (2.182)±0.052	1.722 (1.694)±0.112

To validate the proposed method in a more realistic clinical setting, we also run the inference on the LiTS dataset, which consists of CT scans with tumours. While we expected to see high uncertainty for input images with tumours as out-of-distribution samples, we only observed this for the translation from CT to T1 and only for small tumours (see first two columns in Fig. 7.5). More extensive tumours (third column) and CT to T2 translation (last two columns) did not increase the uncertainty value. It seems that the network was confidently preserving such tumours with the T1 target modality and all of the pathologies with the T2 target modality. On the other hand, small lesions in translation from CT to T1 - the ones that were not generated and probably filtered out as artefacts - caused a spike in the uncertainty value.

Visualisation of corrupted representations in the latent space. To better understand the noise-corrupted representations, we visualise the latent space

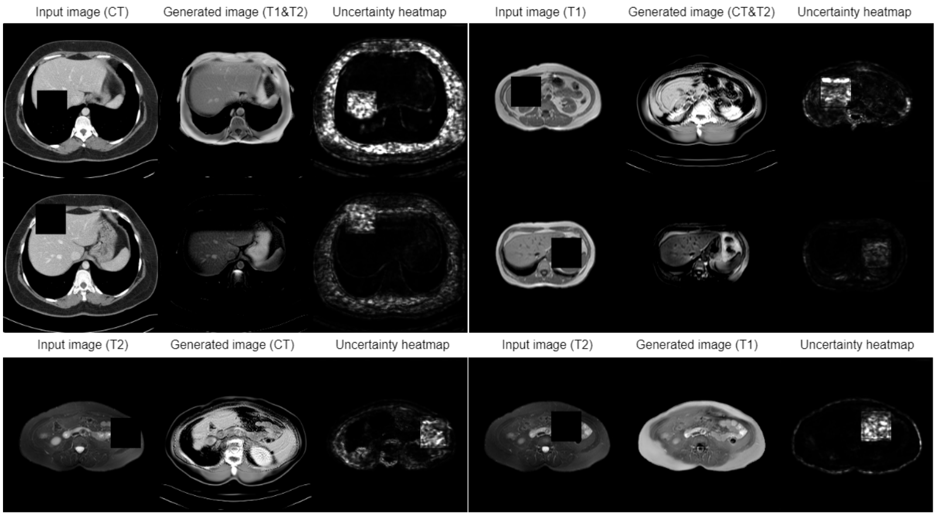

Figure 7.3: Samples generated using input with randomly blacked out 50 × 50 pixel patch. We can see on uncertainty heat maps the lack of confidence when generating this part of the images. The output image does not reflect it.

of the GAN network, including the original representation and the corrupted ones with different noise levels $\alpha = \{1, 3, 5\}$. Fig. 7.4 shows how the corrupted representations stay in the proximity of the original representations. The higher the noise level, the wider the distribution of corrupted representations spread. We expected such observation as the noise level injection happens into the bottleneck representations.

7.4 Improving the quality of a synthesised image

Our next experiment involved injecting noise into latent representation during training to check if it would create a more robust representation and consequently improve image quality. As shown in Table 7.2, injecting a small amount of noise (α=0.5) into half of the synthesised samples during the training process slightly improved the final image quality. Nevertheless, it did not seem to translate to the end task: segmentation accuracy did not improve. Also, we found that introducing excess noise ($\alpha > 0.5$) or corrupting most samples during training leads to deteriorated performance.

7.5 Confidence score and the performance on the downstream task

To address this question, we train three U-Net [95] networks to perform liver segmentation on three imaging modalities, namely CT, T1 and T2, and then run the inference on both the same imaging modality and the transferred synthesized ones and report the results in Table 7.3. On the diagonal, we

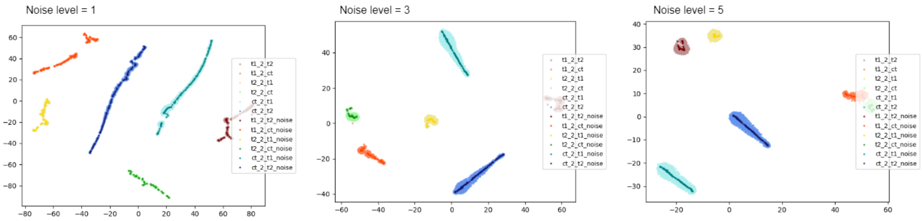

Figure 7.4: Latent space visualisation with different noise levels corrupting the representations. We can see approximately which corrupted representations are passed to the decoder to calculate the confidence. The assumption is that confident, well-represented samples will not be influenced by corrupting their latent representation as much as poorly represented samples.

Table 7.2: Slightly corrupting the latent representation in GANs bottleneck leads to improved image quality. The lower the FID score, the better.

Image Quality [FID]

Source → Target	TarGAN	TarGAN ($\alpha = 0.5$)	TarGAN ($\alpha = 0.8$)	TarGAN ($\alpha = 1.0$)
T1 → CT	0.065	**0.061**	0.062	0.044
T2 → CT	**0.148**	0.150	0.170	0.179
CT → T1	0.065	**0.051**	0.058	0.056
T2 → T1	0.120	**0.114**	0.128	0.145
CT → T2	**0.047**	0.047	0.060	0.065
T1 → T2	0.060	**0.055**	0.058	0.067

present the scores for the original modality, which range from 0.95 for CT to 0.82 for T1, which are slightly different from the ones reported in [19] due to the fact of using a standard 2D U-Net and no enrichment technique [43]. Nevertheless, the segmentation results are acceptable for the CT to T1, CT to T2 and T1 to CT transferred images. However, the performance deteriorates for images where T2 scans are the source modality. It is reflected in the correlation scores as well (*cf.* Table 7.3 and Table 7.4). There is a correlation around and higher than 0.5 for translations where the segmentation network worked well. It suggests that our method can be used most efficiently in cases where the generated images are of good enough quality in the first place for the downstream task network to perform well. If the generated images are of such a low quality that the segmentation network fails (DICE < 0.5), the confidence value does not correlate with the DICE score.

7.6 Other uncertainty estimation techniques

We compare our method to the existing way of estimating aleatoric uncertainty, described in UP-GAN [120]. We measure the quality of generated images with FID scores and the correlation between the DICE coefficient and the mean of

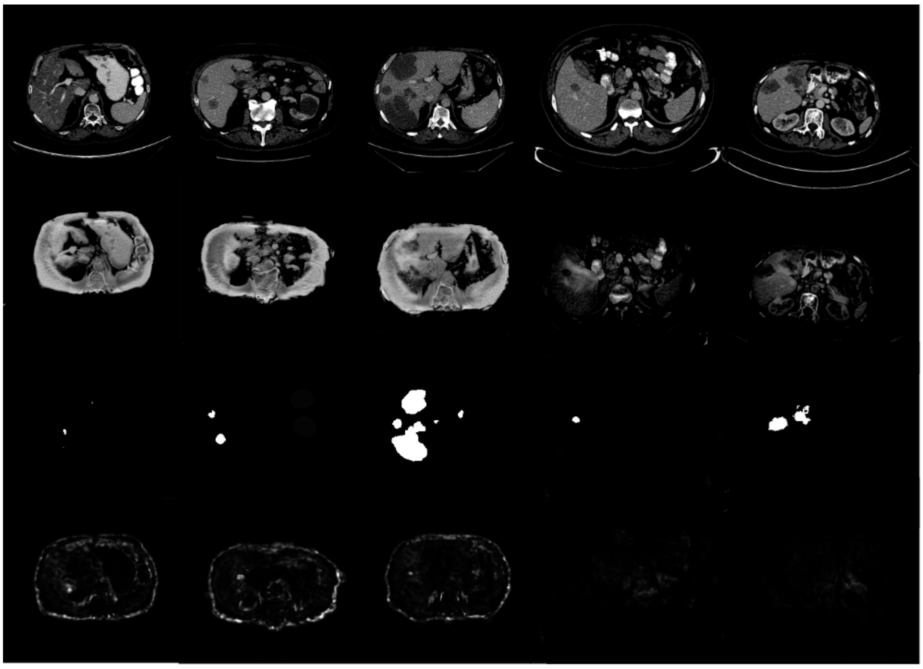

Figure 7.5: CT slides from LiTS dataset containing tumour pathologies. From top to bottom, we show the input CT slice, corresponding tumour segmentation map, generated T1 or T2 images, the segmentation mask, and the uncertainty heat map.

the estimated aleatoric uncertainty values as defined in [118]. The FID scores are slightly lower than those of a TarGAN because the shape-optimizing loss term is not a part of UP-GAN. Furthermore, the aleatoric uncertainty does not correlate well with the DICE score, indicating that even though the aleatoric uncertainty might help improve image quality, as demonstrated in the paper, it does not translate directly into the downstream task of segmentation and cannot be used to indicate unsuitable samples. We also emphasize that our method only affects the inference stage and can be used with any architecture. The UP-GAN model involves significant differences in the architecture (extra outputs of the network) and the optimization process (additional loss terms requiring parameter tuning).

Table 7.3: Segmentation results on original input images (diagonal) and images transferred with TarGAN.

Segmentation quality [DICE]			
from\to	CT	T1	T2
CT	0.951 (0.971)±0.100	0.681 (0.732)±0.223	0.730 (0.753)±0.169
T1	0.690 (0.855)±0.374	0.828 (0.958)±0.313	0.527 (0.667)±0.406
T2	0.409 (0.532)±0.365	0.509 (0.530)±0.366	0.835 (0.954)±0.278

Table 7.4: The image quality evaluated with FID score and the absolute value of a correlation between the confidence score and DICE coefficient for our method and UP-GAN [120].

	Noise injections		UP-GAN [120]	
Source → Target	FID	Correlation	FID	Correlation
CT → T1	0.065	0.542	0.202	0.003
T1 → CT	0.065	0.544	0.162	0.219
CT → T2	0.047	0.495	0.156	0.254
T2 → CT	0.148	0.254	0.654	0.001
T1 → T2	0.060	0.055	0.140	0.102
T2 → T1	0.120	0.311	0.166	0.083

7.7 Conclusions

To summarize, a robust latent representation results in a higher quality of a generated image and higher performance on a downstream segmentation task. There are indicators that the quality of latent representation corresponds to the final quality of a generated image. If the downstream task network performs well, it correlates weakly with the latent representation's quality. We also showed that small noise injections during the training phase produce a more robust representation and slightly higher image quality. However, this does not translate to better segmentation results. We compared the noise injections to the aleatoric uncertainty estimation method proposed by [118]. Although our approach has a more negligible impact on image quality, it indicates performance on downstream tasks more accurately. Our method is easier to incorporate as it does not require changes in the model's architecture or the optimization process. To place this chapter in the context of the whole thesis: we observed that the noise injection technique could also be used without the latent representation disentanglement. The method also generalizes well, as it can be used on different types of data than the microscopic images of the blood cells. Unfortunately, as the dataset was completely unpaired, it was impossible to investigate the correlation between the image quality and the confidence score on this task. However, we saw that it could be indicative of the downstream task performance. structural

Part V

Clinical Validation Study

8 Clinical validation study

To conclude this work, we present the results of a validation study conducted with haematology experts. We wanted to address whether there is a real possibility of replacing the chemical staining with an artificial one. Could the generated white blood cell images be correctly classified by a haematologist?

8.1 Clinical setup and dataset

This study was designed to prove the concept of artificial staining. Through the years of this project's progress, significant changes were introduced both on the hardware and software side, resulting in better quality DIC images and an efficient processing pipeline. These changes resulted in a new, more elaborate dataset, captured fully with the updated version of the DIC microscope that contained both healthy and pathological cells from hundreds of donors. The dataset is completely aligned and paired and fully labelled by haematologists, providing a new value for the data-driven deep learning models. We used this sophisticated dataset to prove the concept of digital staining.

The main objective of this study was to design an image-to-image translation system that could potentially replace chemical staining. The reason to generate artificially stained blood cells in clinical practice instead of directly classifying the unstained images is twofold: (1) the first step in the morphological review of the peripheral blood smear is to highlight through staining those relevant features to characterise the different cell types, and (2) the clinical pathologist could classify the cell images to predict a diagnosis, or these images could be uploaded into a system to obtain the assistance of automatic classification. The system development and assessment involve two main stages:

1. In the first stage, we develop a GAN-based model to automatically generate artificially stained images of 10 cell groups: neutrophils, eosinophils, basophils, monocytes, lymphocytes, reactive lymphocytes, immature granulocytes (promyelocytes, myelocytes and metamyelocytes), blasts (myeloblasts and lymphoblasts), abnormal lymphoid cells and erythroblasts.

2. Based on the results obtained when testing this model, in the second stage, we evaluate the quality of the artificial staining using a manual classification by two clinical experts and an automatic classification with a neural network.

The remaining part of this section will describe the most relevant issues of the GAN model development, including the used image database and how to evaluate the model.

8.1.1 Dataset acquisition

For this study, we constructed a new, much more elaborated dataset consisting of DIC images and corresponding stained images captured with a BF microscope. The data processing pipeline followed the one described in Section 4.4. The only difference was the crop size which was increased to 360x360 pixels following a request of the clinical experts.

8.1.2 Dataset preparation

Pathologists classified stained blood cell crops according to their morphological characteristics. We considered these labels the ground truth for training and evaluating the models. As the dataset consists exclusively of paired images, the labels of the stained cells could be easily transferred to the unstained cells. Without alignment, preparing the ground truth for the unstained crops would be nearly impossible. We used a dataset containing paired unstained and stained images of 92 healthy donors and 92 patients with haematological diseases. Pathological samples were shipped from Spain to Germany, which caused a delay in the scanning procedure. Therefore, some pathological samples resulted in bad quality, and we excluded them from the study. The final dataset was arranged with 92 healthy donors from the Hospital of Erlangen and 83 patients compiled during the daily work in the Hospital Clinic of Barcelona, including patients with viral infections and patients with myeloproliferative, patients with acute leukaemia and patients with lymphoid neoplasia. Table 1 details the number of images corresponding to the different cell classes in training and testing sets. Images were grouped into 12 classes for each dataset: neutrophils, eosinophils, basophils, monocytes, lymphocytes, reactive lymphocytes, immature granulocytes (promyelocytes, myelocytes and metamyelocytes), blasts (myeloid and lymphoid blasts), abnormal lymphoid cells (ALC), erythroblasts, smudges, and artefacts. This dataset consisted of 22 thousand cell images distributed into the above interest groups. The training set was arranged with 19.5 thousand images from 162 samples (85 normal and 77 pathological samples). The testing set for the final assessment included 28 thousand cell images from 13 samples (seven normal and six pathological samples).

Table 8.1: The table shows the structure of full dataset. It contains images from both healthy donors and diseased patients. There are five normal cell types: Neutrophils, Eosinophils, Basophils, Monocytes, Lymphocytes, six pathological types: Reactive lymphocytes, Immature granulocytes, Blasts, Abnormal lymphoid cells, Erythroblasts, Smudges, and always present Artifacts.

	Number of samples	
Cell type	Training	Testing
Neutrophils	105440	16217
Eosinophils	4221	778
Basophils	1497	280
Monocytes	11496	1334
Lymphocytes	40350	4713
Reactive lymphocytes	9152	1373
Immature granulocytes	2172	341
Blasts	6050	597
Abnormal lymphoid cells	449	125
Erythroblasts	1170	541
Smudge	12138	1518
Artifacts	1219	171
Total	195354	27988

8.2 Adapted model

The baseline GAN architecture for image-to-image translation problems on paired datasets is Pix2PixHD [53]. Our dataset contains high-resolution images, so we chose its state-of-art extension for fixed view one-to-one translation problems: Pix2PixHD [123]. Moreover, to introduce additional class information to the training process, we followed the auxiliary classifier-GAN (AC-GAN) approach of including in the discriminator an additional fully connected layer [81] to perform classification. The discriminator gives a probability distribution over the generated image and the class label. The objective of the discriminator is to maximize how similar is the generated image to the original image and how likely it belongs to the correct class. It is a way to optimize the generator to preserve the cell class-specific features.

Implementation details We trained our model for 100 epochs until full convergence, using cross-entropy as a loss function, jointly with the standard adversarial loss. We compared the output of our augmented PIx2PixHD to the baseline model in terms of image quality and classification results using an automated classifier. Additionally, we used data augmentation techniques such as random cropping and image resizing during training. We abstained from using random flips and rotations because an underlying, gradient-based structure within unstained images could be distorted while flipping or rotating. We adopted the following training strategy to address the imbalanced class distribution problem. Firstly, we ensured that each image came from a different

Table 8.2: Class-wise evaluation of generated image quality.

Cell class	MSE	SSIM	LPIPS
Neutrophils	0.002	0.903	0.044
Eosinophils	0.002	0.912	0.036
Basophils	0.003	0.873	0.056
Monocytes	0.002	0.891	0.042
Lymphocytes	0.002	0.902	0.044
Reactive lymphocytes	0.003	0.881	0.038
Immature granulocytes	0.003	0.881	0.055
Blasts	0.002	0.914	0.059
Abnormal lymphoid cells	0.001	0.922	0.051
Erythroblasts	0.003	0.880	0.047
Smudge	0.003	0.882	0.069
Artifacts	0.008	0.749	0.118
Model			
Ours	0.002	0.898	0.046
Pix2PixHD	0.003	0.891	0.047

cell class in each training batch (batch size of six). In addition, in the first 80 epochs, we let the network converge without class balancing, whereas frequency balancing was used in the last 20 epochs. Our model was implemented using PyTorch 1.10 and trained on four Nvidia Titan X GP102 GPUs.

Evaluation To evaluate the quality of artificially stained images, we earlier discussed metrics (for details, please see Section 4.5): Mean Squared Error (MSE), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [133].

8.3 Image quality

All the image quality metrics slightly improved when using additional class information during the optimization process, as shown in Table 8.2. This means the network can generate images of higher visual quality if trained to recognize the class they belong to. The most problematic classes from the point of image quality are Basophils, Immature granulocytes (Promyelocytes, Myelocytes and Metamyelocytes), Reactive lymphocytes and Blasts. Now, we will have a closer look at the results by analyzing the classes separately:

- Basophiles are generated poorly in terms of all the metrics: pixel-wise, structure-wise and feature-wise. It is probably caused by the problems with the generation of the basophilic granules, the most significant feature of the Basophils, that is unfortunately not always captured on the DIC images.

- Immature granulocytes are also generated poorly according to all the metrics. The main problem is generating the right colour of the cytoplasm, which can differ significantly. We suppose there is insufficient information on the DIC images to learn the mapping correctly.

- Reactive lymphocytes have a high pixel-wise error and low SSIM and one of the lowest LPIPS scores. Although the cells are not generated precisely in terms of accuracy and structure, most features are preserved.
- Blasts results are opposite to reactive lymphocytes: very good MSE and SSIM scores but one of the highest LPIPS. The perceptual features are not preserved, so it may be difficult to classify the generated cells.
- Smudges have a high LPIPS score which means some of the features may be difficult to generate properly, for example, the border of the smudged cell, which is not always marked strongly in DIC images.
- Artifacts are the class with the worst metrics performance. However, they are not diagnostically significant. Where does this poor result come from? Most artefacts are created during staining, and there is no indication in the DIC images. Therefore, the network generates arbitrary artefacts that usually are not similar to the original one when measuring the similarity against it. This causes a huge discrepancy between generated images and ground truth.
- The best class in terms of image quality metrics is Eosinophiles. It is a distinct class with specific features preserved on the DIC images (pronounced granulation).
- The two other well-generated classes are Segmented neutrophils and Lymphocytes. The better quality here is probably due to the dominating number of samples in the train set, so the dataset has better distribution coverage of these classes.

Overall, the image quality results show the absolute dominance of the pair-based image-to-image translation over the cycle-consistency-based methods. The effort to create a well-tailored dataset was rewarded with images of superior quality. We present stained cell samples in Fig. 8.1 and 8.2. In some cases (as in the fourth row of Fig. 8.1), we can observe how the artificial staining mitigated variability in the stain that is so characteristic of the process of chemical staining.

8.4 Clinical experts validation results

To answer the question of whether artificially stained images could be used in clinical practice, we conducted a study with two experts that agreed to label the cells stained by our model. Additionally, we used an automatic classifier with ResNet architecture trained on stained images to classify the artificially stained cells. We present the percentage-wise and instance-wise results in Tables 8.3 and 8.4. We also included the detailed expert's confusion matrices in Tables 8.5 and 8.6. Overall accuracy higher than 85% was obtained

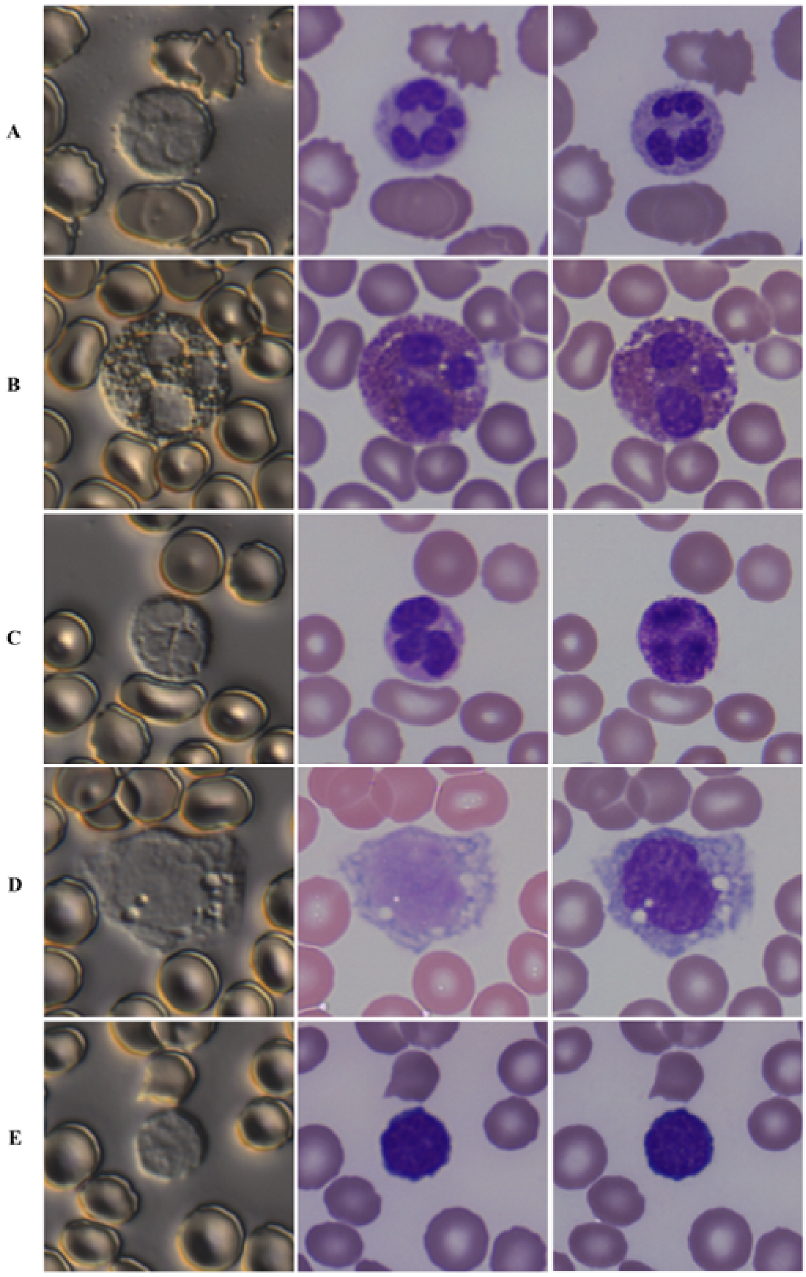

Figure 8.1: Sample artificial staining on normal cells. From left to right: the unstained image captured with interference contrast (DIC), the brightfield (BF) image stained with May Grünwald-Giemsa also serving as ground truth (GT), and the corresponding artificially stained image, From top to bottom: (A) Neutrophil, (B) Eosinophil, (C) Basophil, (D), Monocyte, (E) Lymphocyte.

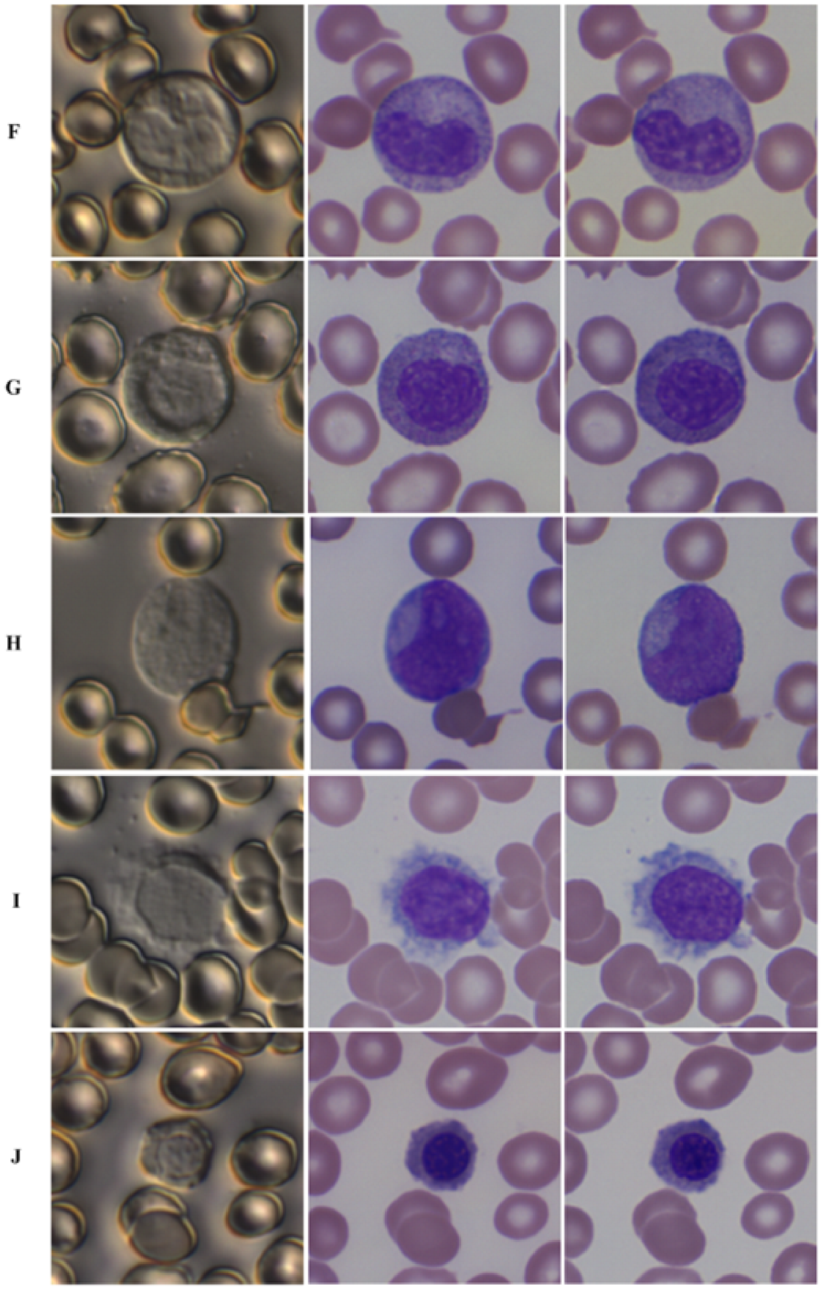

Figure 8.2: Sample artificial staining on abnormal cells. From left to right: the unstained image captured with interference contrast (DIC), the brightfield (BF) image stained with May Grünwald-Giemsa also serving as ground truth (GT), and the corresponding artificially stained image, From top to bottom: (F, G), Immature granulocytes, (H) Blast, (I) Abnormal lymphoid cell, and (J) Erythroblast. Blasts and monocytes appeared difficult to distinguish because of their morphological similarities. It was challenging to artificially generate the characteristic granulation of immature granulocytes and basophils.

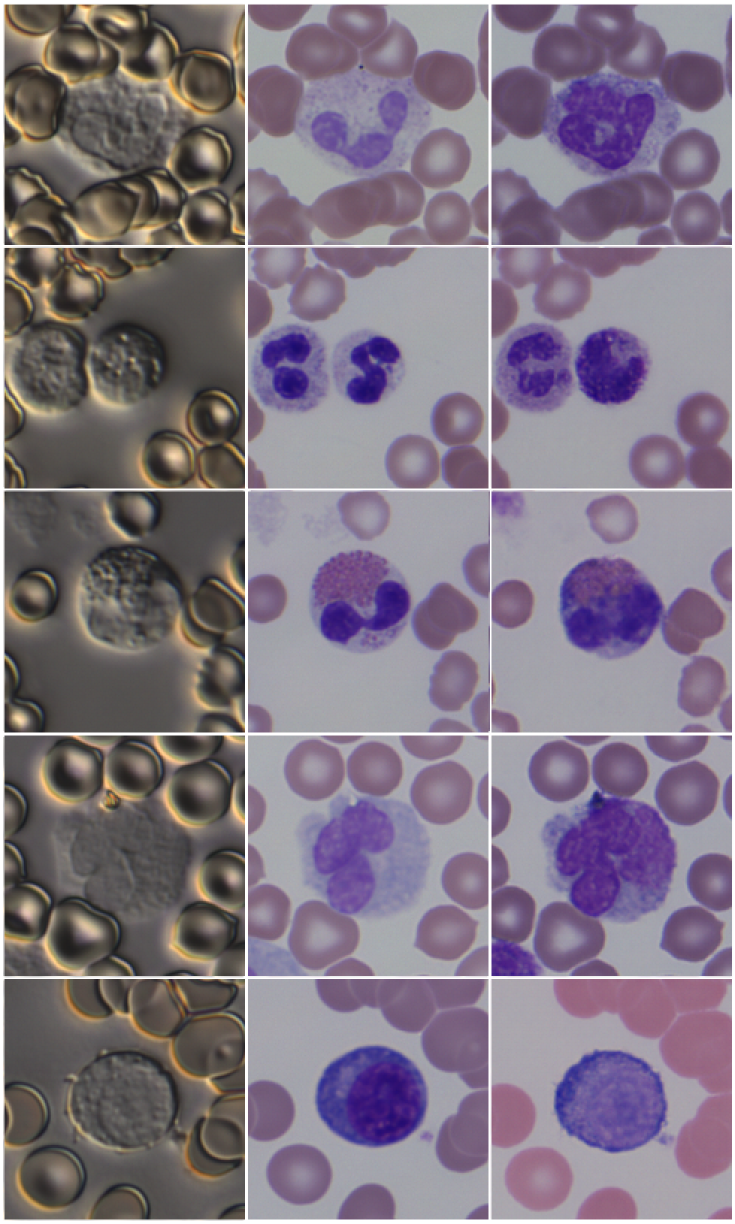

Figure 8.3: Samples of failed artificial staining. Both experts rejected only 1% of artificially stained images because of their quality. From the left: input, chemically stained cell, digitally stained cell.

Table 8.3: Classification of artificially stained images results (percentage-wise) by experts and an automatic classifier.

	Expert 1	Expert 2	Automatic
Overall accuracy	94.5%	85.9%	87.2%
Neutrophils	97.1%	97.5%	96.6%
Eosinophils	97.8%	97.8%	93.5%
Basophils	41.6%	46.8%	40.3%
Monocytes	74%	58.6%	70.1%
Lymphocytes	95.5%	74.2%	77.4%
Immature granulocytes	52.1%	53.3%	59.4%
Blasts	56%	31.5%	81.5%
Abnormal lymphocytes	93.5%	94.4%	68.4%
Erythroblasts	94.1%	95.9%	92.4%
Smudge	82.7%	89.9%	85.6%

for every one of the three ways of classification, with expert 1 reaching an impressive 94.5%. The highest accuracy values were obtained for Segmented neutrophils, Eosinophils and Erythroblasts. The experts also had no problem with Atypical lymphocytes, even though they proved difficult for the automatic model, where we can observe a 25% accuracy drop with respect to the experts' scores. Lymphocytes were classified correctly by Expert 1, reaching 95% accuracy. However, they were challenging for Expert 2 and the Automatic model, with their accuracy values oscillating around 75%. In the confusion matrix of Expert 2 (Table 8.6) we can see that they were mostly confused with Atypical Lymphocytes.

One of the classes that proved the hardest to classify was Basophils, with an accuracy of classification between 40 and 45%. The low accuracy values obtained for Basophils could be explained by the challenging basophilic granulation generation, which is one of the most characteristic features. The low accuracy scores are consistent with the previously observed low performance in image quality metrics. As we can see in the confusion matrices, experts frequently confused Basophiles with Lymphocytes, which supports the claim of the failure of artificially generating the basophilic characteristic granules. Moreover, from our observation during these years working on the project, the granules of Basophiles can be washed out when chemically staining the smears. This means that in the training set, we have images of Basophiles without prominent and characteristic basophilic granules, which could also explain the results of the testing set, where the prominent granulation was not generated. The resulting artificially stained image is a cell with a nucleus with mature chromatin, which seems to have slightly basophilic granulation but could be misclassified as a Lymphocyte.

Overall, only 1% of the generated images were classified by the experts as poor quality (PQ) (see Tables 8.5 and 8.6), which underlines the great quality of the samples. We present the failed cases in Fig. 8.3.

Table 8.4: Classification of artificially stained images results done by both experts and an automatic. Reporting exact numbers.

	Expert 1	Expert 2	Automatic
Overall accuracy	94.5%	85.9%	87.2%
Neutrophils	3,428/3,530	3,440/3,530	3,409/3,530
Eosinophils	45/46	45/46	43/46
Basophils	32/77	36/77	31/77
Monocytes	270/365	214/365	256/365
Lymphocytes	2,275/2,383	1,767/2,383	1,845/2,383
Immature granulocytes	127/244	130/244	145/244
Blasts	459/821	259/821	669/821
Abnormal lymphocytes	386/413	390/413	275/413
Erythroblasts	160/170	163/170	157/170
Smudge	587/710	638/710	608/710

Table 8.5: Confusion matrix detailing classification results of Expert 1.

	SNE	EOS	BAS	MO	LY	ALC	BL	IG	NRBC	SMU	PQ
SNE	**97.1**	0	0.34	0.7	0.3	0.03	0.03	0.7	0	0.1	0.6
EOS	2.2	**97.8**	0	0	0	0	0	0	0	0	0
BAS	9.1	1.3	**41.6**	7.8	15.6	0	0	0	0	1.3	23.3
MO	0.6	0	0	**74**	13.4	1.4	2.2	1.5	0	4.4	2.5
LY	0.1	0	0	1.7	**95.5**	1	0.13	0.04	0	0.8	0.6
ALC	0	0	0	1.5	4.1	**93.5**	0	0	0	0	0.7
BL	0.5	0	0.2	4	35.9	0.4	**55.9**	1.3	0	0	1.8
IG	20.1	0	0.4	10.3	12.7	0	1.6	**52.1**	0	0.4	2.4
NRBC	0	0	0	0.6	4.1	0	0	0	**94.1**	0	0.6
SMU	3.2	0.1	0	0	4.7	0.1	0.3	0	3.2	**82.7**	1

Table 8.6: Confusion matrix detailing classification results of Expert 2.

	SNE	EOS	BAS	MO	LY	ALC	BL	IG	NRBC	SMU	PQ
SNE	**97.5**	0	0.4	0.2	0.3	0.1	0	0.6	0	0.5	0.4
EOS	2.2	**97.8**	0	0	0	0	0	0	0	0	0
BAS	7.8	0	**46.8**	3.9	20.8	3.9	0	1.3	0	6.5	9
MO	0.3	0	0	**58.6**	9.6	17	0.6	4.3	0	6.3	3.3
LY	0.1	0	0	0.9	**74.2**	23.1	0	0.4	0	0.8	0.5
ALC	0	0	0	0.2	4.1	**94.5**	0	0	0	0.5	0.5
BL	0.5	0	0.9	2.3	10	51.4	**31.6**	1.5	0	0	1.8
IG	22.5	0	1.1	2.9	6.6	6.2	3.3	**53.3**	0.4	0.4	3.3
NRBC	0	0	0	0	1.8	0	0	1.8	**95.9**	0.5	0
SMU	2.7	0	0	0	3	0.7	0.1	0.1	2.5	**89.9**	0.7

8.5 Limitations and conclusions

There are still some limitations to this method which need to be considered and improved. The first and most significant one is the quality of the definition of the nucleus and cell structure. From what we have seen in previous models, we can assume that additional structural constraints with high-quality segmentation masks for the nucleus, cytoplasm, granulation, and nucleoli would lead to improved results. Another important limitation is a still lack of colour faithfulness in the generated staining, which is essential for properly assessing blood cells (immatureness of the chromatin, presence or absence of cytoplasmic granulation, basophilia of the cytoplasm).

To sum up, the proposed model revealed the possibility of haematologists working directly with artificially stained blood cell images. It has been proven to properly preserve most of the normal cell features and pathological cell features. It could provide standardization of the staining protocol among clinical laboratories worldwide, reducing tedious work, chemical residues, costs and environmental impact.

Part VI

Conclusion and Outlook

9 Conclusion

Here I will summarize our methods and findings, analyze their advantages and limitations, propose directions for future research, and outline the potential clinical impact.

9.1 Summary

In this work, we proposed multiple approaches to staining and classifying white blood cells. We showed how segmentation pseudo-masks help to preserve the structure during image-to-image translation and that even partial matches between domains significantly improve the domain adaptation and final classification result. Next, we designed a method to estimate the generated image quality based on the analysis of the latent space. We introduced a technique of noise injections into the latent space, which allows to generate multiple outputs. Quantifying the similarities between these outputs gave us a confidence score correlated with generated image quality. Additionally, we showed how to disentangle white blood cells image components in terms of style and structure using segmentation pseudo-masks and how to use such disentangled structure and style representation for more interpretable generator confidence estimation. Furthermore, we demonstrated that the method of injecting noise into latent representation to estimate the confidence of the generator also holds for the whole images without latent space disentanglement and for different modalities. We analysed its correlation with the downstream task of segmentation. Last but not least, we conducted a validation study with clinical experts, where we leveraged an elaborate white blood cell dataset with pairs of unstained and stained images. Additionally, we used provided class information to improve generated image quality. We showed that there exists indeed a possibility of replacing the chemical staining process with a digital one and integrating such change into clinical workflow.

9.2 Limitations and future Work

While the main focus of our work has been to develop the concepts and methods that allow for efficient and realistic image-to-image translation, this problem is also limited by the imaging modalities used. At the current stage, it seems that not all of the features relevant for classification are captured with DIC microscopy. Although there was a huge improvement in data quality throughout this project, further advances would also surely be reflected in the quality of generated images, especially in pathological samples. Moreover, time is also a limiting factor for this approach. Scanning a blood smear with the current method takes around forty minutes, which needs to be significantly reduced before entering clinical practice. Forty minutes is still close to the time required for staining, so only the dye cost and waste reduction would be the benefits in this case.

What is more, although the current method of confidence estimation correlates with image quality, it translates only slightly to a correlation with the downstream task of segmentation. Further investigation into more sophisticated manners of noise corruption could enhance this technique. Another aspect is the disentanglement approach which relies on strong assumptions about the structure of the cell. On the one hand, it would make the method generalize better if disentanglement was achieved in an unsupervised fashion. On the other hand, better segmentation masks, including classes such as nucleoli and granulation, would allow for more targeted noise corruption and more precise estimation at the end.

Furthermore, there still exists a significant gap between the results obtainable on paired and unpaired datasets. In reality, paired datasets are rather rare, mostly because of the constrains they impose during the data collection phase and the required processing. However, the undeniable image quality improvement such a dataset offers makes it a preferred option for this problem. In most cases, no parts of the blood cells were either added or omitted while training using the direct reconstruction loss on a large dataset. The gap between models trained with direct reconstruction loss and those trained with cycle-consistency loss can be partially closed by additional constrained (for example, the segmentation module for structure preservation) but not fully eliminated. As a sample cannot be stained twice, this problem will persist for translation from one staining protocol to another.

Therefore, future work should be concentrated on further improvement of the imaging aspect and developing robust methodologies for image-to-image translation that can additionally output their confidence in the correctness of the generated image. In a broader view, investigating unstained red blood cells could also be of interest to confirm whether diagnosing diseases such as anaemia and malaria would be feasible with unstained cells. Also, applying digital staining to histology data would indicate if chemical staining can be eliminated from the laboratories.

9.3 Potential clinical impact

If we could replace the chemical staining of blood cells with digital processing, it would change the world of haematology, open the door to transforming histology, and revolutionize diagnostics.

It is difficult to imagine how much would change with a detailed and successful analysis of unstained blood samples. First, the examination would no longer be limited to the laboratories. It would allow ambulatory blood analysis, which would speed the diagnosis significantly.

What is more, it would reduce costs on many levels. First, the costs of the staining equipment and dyes. Secondly, the work force of laboratory staff would no longer be required to execute the tiring staining procedure. Finally, it would reduce chemical waste and the costs of disposing of it. Consequently, the overall cost reduction would significantly increase the availability of blood examinations.

Increased availability opens the door to creating a big database of blood cell images, including rare diseases. In general, people seem less resistant to sharing their blood samples for research than, for example, whole-body scans, which could potentially lead to a dataset including millions of patients. The diagnostic benefits would be enormous.

Current systems such as CellaVision [1] concentrate mostly on healthy cells and still fail in many cases in the presence of diseases. The system presented in this thesis considered pathological classes of blood cells. We showed there potentially, it would be possible to diagnose patients with digitally stained cells. With confidence estimation algorithms in place, the experts would have to manually examine the cells only in special cases.

To conclude, a system automatically classifying and staining white blood cells could help to normalize quick blood cell tests and significantly speed up the diagnostics process.

9.4 Epilogue

The methods presented in this thesis have proven to be a real possibility for the future of haematology. We believe that perfecting these methods and collecting more exhaustive datasets leads to replacing the process of chemical staining with a digital one. Although the exciting journey taken to reach these last lines of the dissertation is now coming to an end, we believe that the developed methodology and obtained results will serve as a stepping stone for future research and, ultimately, for making precise and realistic artificial staining of any white blood cell possible.

A Authored and Co-authored Publications

2021

- Tomczak, A., Ilic, S., Marquardt, G., Engel, T., Forster, F., Navab, N., Albarqouni, S.: Multi-task multi-domain learning for digital staining and classification of leukocytes. IEEE Transactions on Medical Imaging (2021). doi: 10.1109/TMI.2020.3046334

2022

- Tomczak, A., Boldu, L., Brock, J.P., Merino, A., Engel, T., Marquardt, G.: Diagnosing artificially stained images of peripheral blood cells. International Symposium on Technical Innovations in Laboratory Hematology (2022)
- Tomczak, A., Gupta, A., Ilic, S., Navab, N., Albarqouni, S.: What can we learn about a generated image corrupting its latent representation? MICCAI (2022)
  – *Oral presentation*

2023

- Tomczak, A., Ilic, S., Marquardt, G., Engel, T., Navab, N., Albarqouni, S.: Digital staining of white blood cells with confidence estimation
  – *Under review*

Bibliography

[1] CellaVision System: improving microscopy workflows for greater diagnostic certainty. `https://www.cellavision.com/en/`. Accessed: 2021-04-29

[2] Acevedo Lipes, A., Alferez, S., Merino, A., Puigvi, L., Rodellar, J.: Recognition of peripheral blood cell images using convolutional neural networks. Computer Methods and Programs in Biomedicine **180**, 105,020 (2019). doi: 10.1016/j.cmpb.2019.105020

[3] Almezhghwi, K., Serte, S.: Improved classification of white blood cells with the generative adversarial network and deep convolutional neural network. Computational Intelligence and Neuroscience **2020**, 1–12 (2020). doi: 10.1155/2020/6490479

[4] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (2017). URL `http://arxiv.org/abs/1701.07875`

[5] Bain, B.: Blood Cells: A Practical Guide. Wiley (2014)

[6] Bashir, R.M.S., Qaiser, T., Raza, S.E.A., Rajpoot, N.M.: Hydramix-net: A deep multi-task semi-supervised learning approach for cell detection and classification. In: Interpretable and Annotation-Efficient Learning for Medical Image Computing, pp. 164–171. Springer International Publishing (2020)

[7] Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nature (2019). doi: 10.1038/s42256-018-0004-1

[8] Benaim, S., Galanti, T., Wolf, L.: Estimating the success of unsupervised image to image translation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)

[9] Bentaieb, A., Hamarneh, G.: Adversarial stain transfer for histopathology image analysis. IEEE Transactions on Medical Imaging **37**, 792–802 (2018)

[10] Bilic, P., Christ, P., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., Kadoury, S., Konopczynski, T., Le, M., li, F., Li, X., LipkovÃ , J., Lowengrub, J., Meine, H., Moltz, J., Wu, J.: The liver tumor segmentation benchmark (lits) (2019)

[11] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York, NY, USA (1995)

[12] Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural networks. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15 (2015)

[13] Borji, A.: Pros and cons of gan evaluation measures. Computer Vision and Image Understanding (2018). doi: 10.1016/j.cviu.2018.10.009

[14] Bradbury, S., Bracegirdle, B.: Introduction to Light Microscopy. Microscopy handbooks. Bios Scientific Publishers (1998)

[15] Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019)

[16] Cellucci, C.J., Albano, A.M., Rapp, P.E.: Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. Phys. Rev. E **71**, 066,208 (2005). doi: 10.1103/PhysRevE.71.066208

[17] Chang, W.L., Wang, H.P., Peng, W.H., Chiu, W.C.: All about structure: Adapting structural information across domains for boosting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[18] Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 865–872 (2019). doi: 10.1609/aaai.v33i01.3301865

[19] Chen, J., Wei, J., Li, R.: Targan: Target-aware generative adversarial networks for multi-modality medical image translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer (2021)

[20] Chen, S., Qin, A., Zhou, D., Yan, D.: Technical note: U-net-generated synthetic ct images for magnetic resonance imaging-only prostate intensity-modulated ratiation therapy treatment planning. Medical Physics **45** (2018)

[21] Cheng, M.M., Liu, X.C., Wang, J., Lu, S.P., Lai, Y.K., Rosin, P.L.: Structure-preserving neural style transfer. IEEE Transactions on Image Processing **29**, 909–920 (2020). doi: 10.1109/TIP.2019.2936746

[22] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[23] Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)

[24] Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing (2018)

[25] Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5707–5715. IEEE Computer Society, Los Alamitos, CA, USA (2017). doi: 10.1109/ICCV.2017.608

[26] Dong, N., Zhai, M., Chang, J., Wu, C.: White blood cell classification. CoRR **abs/2008.07181** (2020). URL `https://arxiv.org/abs/2008.07181`

[27] Emami, H., Dong, M., Nejad-Davarani, S., Glide-Hurst, C.: Sa-gan: Structure-aware generative adversarial network for shape-preserving synthetic ct generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2021)

[28] Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(10), 1858–1865 (2008)

[29] Gabbay, A., Hoshen, Y.: Improving style-content disentanglement in image-to-image translation. arXiv preprint arXiv:2007.04964 (2020)

[30] Gadermayr, M., Appel, V., Klinkhammer, B.M., Boor, P., Merhof, D.: Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pp. 165–173. Springer International Publishing (2018)

[31] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16 (2016)

[32] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research (2016)

[33] Ge, Y., Wei, D., Xue, Z., Wang, Q., Zhou, X., Zhan, Y., Liao, S.: Unpaired mr to ct synthesis with explicit structural constrained adversarial learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1096–1099. IEEE (2019)

[34] Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: European Conference on Computer Vision (ECCV) (2016)

[35] Ghosh, M., Das, D., Mandal, S., Chakraborty, C., Pala, M., Maity, A.K., Pal, S.K., Ray, A.K.: Statistical pattern analysis of white blood cell nuclei morphometry. In: 2010 IEEE Students Technology Symposium (TechSym), pp. 59–66 (2010). doi: 10.1109/TECHSYM.2010.5469197

[36] Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics (2010)

[37] Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). URL `https://proceedings.neurips.cc/paper/2018/file/dc6a70712a252123c40d2adba6a11d84-Paper.pdf`

[38] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press (2016)

[39] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Advances in neural information processing systems (NIPS) (2014)

[40] Graham, S., Vu, Q.D., Jahanifar, M., Raza, S.E.A., Minhas, F., Snead, D., Rajpoot, N.: One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. Medical Image Analysis **83**, 102,685 (2023). doi: https://doi.org/10.1016/j.media.2022.102685

[41] Guan, S., Loew, M.: Measures to evaluate generative adversarial networks based on direct analysis of generated images (2020)

[42] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)

[43] Gupta, L., Klinkhammer, B., Boor, P., Merhof, D., Gadermayr, M.: GAN-Based Image Enrichment in Digital Pathology Boosts Segmentation Accuracy, pp. 631–639 (2019). doi: 10.1007/978-3-030-32239-7_70

[44] Habibzadeh, M., Jannesari, M., Rezaei, Z., Baharvand, H., Totonchi, M.: Automatic white blood cell classification using pre-trained deep learning models: Resnet and inception. In: International Conference on Machine Vision (ICML) (2018)

[45] Habibzadeh, M., Jannesari, M., Rezaei, Z., Totonchi, M., Baharvand, H.: Automatic white blood cell classification using pre-trained deep learning models: Resnet and inception. p. 105 (2018). doi: 10.1117/12.2311282

[46] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015). URL `http://arxiv.org/abs/1512.03385`

[47] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS. Curran Associates Inc., Red Hook, NY, USA (2017)

[48] Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation (2018)

[49] Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR **abs/1608.06993** (2016). URL `http://arxiv.org/abs/1608.06993`

[50] Huang, P., Li, D., Jiao, Z., Wei, D., Li, G., Wang, Q., Zhang, H.: CoCa-GAN: Common-Feature-Learning-Based Context-Aware Generative Adversarial Network for Glioma Grading, pp. 155–163 (2019). doi: 10.1007/978-3-030-32248-9_18

[51] Ibrahim, S.: Comparative analysis of support vector machine (svm) and convolutional neural network (cnn) for white blood cells classification. International Journal of Advanced Trends in Computer Science and Engineering **8**, 394–399 (2019). doi: 10.30534/ijatcse/2019/6981.32019

[52] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15 (2015)

[53] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

[54] Ivakhnenko, A., Lapa, V., Lapa, V., McDonough, R.: Cybernetics and Forecasting Techniques. Modern analytic and computational methods in science and mathematics. American Elsevier Publishing Company (1967). URL `https://books.google.de/books?id=rGFgAAAAMAAJ`

[55] Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)

[56] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)

[57] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proc. CVPR (2020)

[58] Kavur, A.E., Gezer, N.S., Baris, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Ozkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nurnberger, A., Maier-Hein, K.H., Bozdagi Akar, G., Unal, G., Dicle, O., Selver, M.A.: Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101,950 (2021). doi: https://doi.org/10.1016/j.media.2020.101950

[59] Kazemi, H., Iranmanesh, S.M., Nasrabadi, N.: Style and content disentanglement in generative adversarial networks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 848–856 (2019). doi: 10.1109/WACV.2019.00095

[60] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, p. 5580â€"5590. Curran Associates Inc., Red Hook, NY, USA (2017)

[61] Kim, J., Kim, M., Kang, H., Lee, K.H.: 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`

[62] Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2014)

[63] Kotovenko, D., Sanakoyeu, A., Lang, S., Ommer, B.: Content and style disentanglement for artistic style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

[64] Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems **25** (2012). doi: 10.1145/3065386

[65] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: F. Pereira, C. Burges, L. Bottou, K. Weinberger (eds.) Advances in Neural Information Processing Systems, vol. 25. Curran Associates, Inc.

(2012). URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[66] Kutlu, H., Avci, E., Ã–zyurt, F.: White blood cells detection and classification based on regional convolutional neural networks. Medical Hypotheses **135**, 109,472 (2020). doi: https://doi.org/10.1016/j.mehy.2019.109472

[67] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, p. 6405â€"6416. Curran Associates Inc., Red Hook, NY, USA (2017)

[68] Lecun, Y.: Une procedure d'apprentissage pour reseau a seuil asymmetrique (a learning scheme for asymmetric threshold networks). In: Proceedings of Cognitiva 85, Paris, France, pp. 599–604 (1985)

[69] Lecun, Y.: Generalization and network design strategies. Elsevier (1989)

[70] Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M.K., Yang, M.H.: Drit++: Diverse image-to-image translation viadisentangled representations. International Journal of Computer Vision pp. 1–16 (2020)

[71] Liang, G., Hong, H., Xie, W., Zheng, L.: Combining convolutional neural network with recursive neural network for blood cell image classification. IEEE Access **6**, 36,188–36,197 (2018)

[72] Liang, G., Hong, H., Xie, W., Zheng, L.: Combining convolutional neural network with recursive neural network for blood cell image classification. IEEE Access **6**, 36,188–36,197 (2018). doi: 10.1109/ACCESS.2018.2846685

[73] Liu, X., Yin, G., Shao, J., Wang, X., Li, H.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: Advances in Neural Information Processing Systems (2019)

[74] Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Shao, L.: Structure preserving stain normalization of histopathology images using self supervised semantic guidance. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020. Springer International Publishing (2020)

[75] Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. bioRxiv (2019). doi: 10.1101/564039

[76] Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. bioRxiv (2019). doi: 10.1101/564039. URL https://www.biorxiv.org/content/early/2019/02/28/564039

[77] Mikolov, T.: Statistical language models based on neural networks. Disertační práce, Vysoké učení technické v Brně, Fakulta informačních technologií (2012). URL https://www.fit.vut.cz/study/phd-thesis/283/

[78] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016). doi: 10.1109/3DV.2016.79

[79] Mundhra, D., Cheluvaraju, B., Rampure, J., Dastidar, T.R.: Analyzing microscopic images of peripheral blood smear using deep learning. In: DLMIA/ML-CDS at MICCAI (2017)

[80] Murphy, D.: Fundamentals of Light Microscopy and Electronic Imaging. Wiley (2002)

[81] Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. Proceedings of Machine Learning Research, pp. 2642–2651. PMLR, International Convention Centre, Sydney, Australia (2017)

[82] Öztürk, S., Akdemir, B.: Cell-type based semantic segmentation of histopathological images using deep convolutional neural networks. International Journal of Imaging Systems and Technology **29** (2019)

[83] Park, S.H., Park, C., Choi, M.O., Kim, M.J., Cho, Y.U., Jang, S., Chi, H.S.: Automated digital cell morphology identification system (cellavision dm96) is very useful for leukocyte differentials in specimens with qualitative or quantitative abnormalities. International Journal of Laboratory Hematology **35** (2013)

[84] Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)

[85] Parker, D.B.: Learning logic. Tech. Rep. TR-47, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology (1985)

[86] Pei, Z., Cao, Z., Long, M., Wang, J.: Multi-adversarial domain adaptation (2018)

[87] Peng, X., Huang, Z., Sun, X., Saenko, K.: Domain agnostic learning with disentangled representations. In: ICML (2019)

[88] Postels, J., Ferroni, F., Coskun, H., Navab, N., Tombari, F.: Sampling-free epistemic uncertainty estimation using approximated variance propagation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2931–2940 (2019)

[89] Prellberg, J., Kramer, O.: Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks. In: ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging. Springer Singapore (2019)

[90] Rana, A., Yauney, G., Lowe, A., Shah, P.: Computational histological staining and destaining of prostate core biopsy RGB images with generative adversarial neural networks. IEEE International Conference on Machine Learning and Applications (ICMLA) (2018)

[91] Redmon, J., Divvala, S., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[92] Regmi, K., Borji, A.: Cross-view image synthesis using conditional gans. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[93] Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Unsupervised domain adaptation for classification of histopathology whole-slide images. Frontiers in Bioengineering and Biotechnology **7** (2019)

[94] Rivenson, Y., Liu, T., Wei, Z., Zhang, Y., Haan, K., Ozcan, A.: Phasestain: the digital staining of label-free quantitative phase microscopy images using deep learning. Light: Science and Applications (2019)

[95] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer International Publishing (2015)

[96] Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Cornell Aeronautical Laboratory. Report no. VG-1196-G-8. Spartan Books (1962). URL `https://books.google.de/books?id=7FhRAAAAMAAJ`

[97] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)

[98] Salehi, P., Chalechale, A.: Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. 2020 International Conference on Machine Vision and Image Processing (MVIP) pp. 1–7 (2020)

[99] Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[100] Schoenauer-Sebag, A., Heinrich, L., Schoenauer, M., Sebag, M., Wu, L.F., Altschuler, S.J.: Multi-domain adversarial learning. In: International Conference on Learning Representations (2019)

[101] Shaban, M.T., Baur, C., Navab, N., Albarqouni, S.: Staingan: Stain style transfer for digital histological images. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) pp. 953–956 (2018)

[102] Shahin, A., Guo, Y., Amin, K., Sharawi, A.A.: White blood cells identification system based on convolutional deep neural learning networks. Computer Methods and Programs in Biomedicine **168**, 69–80 (2019). doi: https://doi.org/10.1016/j.cmpb.2017.11.015. URL `https://www.sciencedirect.com/science/article/pii/S016926071730411X`

[103] Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J.M., Turkbey, B., Harmon, S.A., Sanford, T.H., Mehralivand, S., Choyke, P.L., Wood, B.J., Xu, D.: Multi-domain image completion for random missing input data. IEEE Trans. Med. Imaging **40**(4), 1113–1122 (2021). doi: 10.1109/TMI.2020.3046444

[104] Shmelkov, K., Schmid, C., Alahari, K.: How good is my gan? In: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds.) Computer Vision – ECCV 2018. Springer International Publishing (2018)

[105] Sinha, N., Ramakrishnan, A.: Automation of differential blood count. In: TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, vol. 2, pp. 547–551 Vol.2 (2003). doi: 10.1109/TENCON.2003.1273221

[106] Song, T.H., Sanchez, V., EI Daly, H., Rajpoot, N.M.: Simultaneous cell detection and classification in bone marrow histology images. IEEE Journal of Biomedical and Health Informatics **23**(4), 1469–1476 (2019). doi: 10.1109/JBHI.2018.2878945

[107] Su, M.C., Cheng, C.Y., Wang, P.C.: A neural-network-based approach to white blood cell classification. In: The Scientific World Journal (2014)

[108] Su, M.C., Cheng, C.Y., Wang, P.C.: A neural-network-based approach to white blood cell classification. TheScientificWorldJournal **2014**, 796,371 (2014). doi: 10.1155/2014/796371

[109] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS (2014)

[110] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR **abs/1409.4842** (2014). URL `http://arxiv.org/abs/1409.4842`

[111] Tang, H., Bai, S., Sebe, N.: Dual attention gans for semantic image synthesis. In: ACM MM (2020)

[112] Tang, H., Qi, X., Xu, D., Torr, P.H., Sebe, N.: Edge guided gans with semantic preserving for semantic image synthesis. arXiv preprint arXiv:2003.13898 (2020)

[113] Tiwari, P., Qian, J., Li, Q., Wang, B., Gupta, D., Khanna, A., Rodrigues, J.J., Albuquerque, V.H.C.: Detection of subtype blood cells using deep learning. Cognitive Systems Research (2018)

[114] Tomczak, A., Boldu, L., Brock, J.P., Merino, A., Engel, T., Marquardt, G.: Diagnosing artificially stained images of peripheral blood cells. International Symposium on Technical Innovations in Laboratory Hematology (2022)

[115] Tomczak, A., Gupta, A., Ilic, S., Navab, N., Albarqouni, S.: What can we learn about a generated image corrupting its latent representation? MICCAI (2022)

[116] Tomczak, A., Ilic, S., Marquardt, G., Engel, T., Forster, F., Navab, N., Albarqouni, S.: Multi-task multi-domain learning for digital staining and classification of leukocytes. IEEE Transactions on Medical Imaging (2021). doi: 10.1109/TMI.2020.3046334

[117] Tomczak, A., Ilic, S., Marquardt, G., Engel, T., Navab, N., Albarqouni, S.: Digital staining of white blood cells with confidence estimation

[118] Upadhyay, U., Chen, Y., Akata, Z.: Robustness via uncertainty-aware cycle consistency (2021)

[119] Upadhyay, U., Chen, Y., Akata, Z.: Uncertainty-aware generalized adaptive cyclegan. arXiv preprint arXiv:2102.11747 (2021)

[120] Upadhyay, U., Chen, Y., Hebb, T., Gatidis, S., Akata, Z.: Uncertainty guided progressive gans for medical image translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer (2021)

[121] Vatathanavaro, S., Tungjitnob, S., Pasupa, K.: White blood cell classification : A comparison between vgg-16 and resnet-50 models (2018)

[122] Wang, Q., Bi, S., Sun, M., Wang, Y., Wang, D., Yang, S.: Deep learning approach to peripheral leukocyte recognition. In: PloS one (2019)

[123] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)

[124] Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Disentangling content and style via unsupervised geometry distillation (2019). URL `https://openreview.net/forum?id=SkgsQ8LK_E`

[125] Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: Geometry-aware unsupervised image-to-image translation. In: CVPR (2019)

[126] Xin, B., Hu, Y., Zheng, Y., Liao, H.: Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis. In: The IEEE International Symposium on Biomedical Imaging (ISBI) (2020)

[127] Yampri, P., Pintavirooj, C., Daochai, S., Teartulakarn, S.: White blood cell classification based on the combination of eigen cell and parametric feature detection. In: 2006 1ST IEEE Conference on Industrial Electronics and Applications, pp. 1–4 (2006). doi: 10.1109/ICIEA.2006.257341

[128] Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S.: Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2019, pp. 255–263. Springer International Publishing, Cham (2019)

[129] Yildirim, M., Ã‡inar, A.: Classification of white blood cells by deep learning methods for diagnosing disease. Revue d'Intelligence Artificielle **33**, 335–340 (2019). doi: 10.18280/ria.330502

[130] Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P.: Ea-gans: Edge-aware generative adversarial networks for cross-modality mr image synthesis. IEEE Transactions on Medical Imaging **38**(7), 1750–1762 (2019). doi: 10.1109/TMI.2019.2895894

[131] Zanjani, F.G., Zinger, S., Bejnordi, B.E., van der Laak, J.: Histopathology stain-color normalization using deep generative models (2018)

[132] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Proceedings of the 36th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 97, pp. 7354–7363. PMLR (2019)

[133] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

[134] Zhang, Z., Sun, L., Zheng, Z., Li, Q.: Disentangling the spatial structure and style in conditional vae. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 1626–1630 (2020). doi: 10.1109/ICIP40778.2020.9190908

[135] Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. pp. 9242–9251 (2018). doi: 10.1109/CVPR.2018.00963

[136] Zhao, H., Zhang, S., Wu, G., Moura, J.M.F., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. In: Advances in Neural Information Processing Systems (NIPS) (2018)

[137] Zhou, Y.T., Chellappa, R.: Computation of optical flow using a neural network. In: IEEE 1988 International Conference on Neural Networks, pp. 71–78 vol.2 (1988). doi: 10.1109/ICNN.1988.23914

[138] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)

[139] Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 465–476. Curran Associates, Inc. (2017)