TRANSFER LEARNING USING GENERATIVE ADVERSARIAL NETWORKS FOR TISSUE SPECIFIC VIRTUAL STAINING

University of Turku
Department of Computing
Master of Science in Technology Thesis
Master's Degree Programme in ICT (Software Engineering)
June 2025
Hyder Abbas

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

UNIVERSITY OF TURKU
Department of Computing

Hyder Abbas: TRANSFER LEARNING USING GENERATIVE ADVERSARIAL NETWORKS FOR TISSUE SPECIFIC VIRTUAL STAINING

Master of Science in Technology Thesis, 62 p.
Master's Degree Programme in ICT (Software Engineering)
June 2025

Recent advances in biomedical imaging have highlighted the potential of generative adversarial networks (GANs) to significantly improve histopathological analysis by artificially staining tissues. The process is colloquially known as "virtual staining". Conventional methods, although quite essential and meaningful in diagnostics, are both time-consuming and expensive, creating an urge for much more efficient and high-quality alternative methods. In order to address said need, this thesis explores the use of 'transfer learning' of GAN models to improve the virtual staining of histopathological images, which primarily aims to enhance model accuracy and reduce the computational demands that are related to training a model from scratch. Transfer learning leverages available weights of a pre-trained model to address data scarcity and time-complexity, enabling GAN models to produce high-quality stained representations from raw tissue images across diverse tissue types.

The primary objectives of this research include investigating the feasibility of transfer learning in generating virtual stains for various tissue types and assessing its impact on training efficiency and image quality. We have used DensePix2Pix GAN model for generating predicted images, this model serve as the foundation for this approach, with modifications made to support transfer learning on a dataset of histopathological images from four tissue types: skin, kidney, spleen, and intestine. Performance of model is evaluated using Pearson Correlation Coefficient Ratio (PCCR), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Mean Square Error (MSE).

This research addresses some critical research questions on the trade-offs between training from scratch versus transfer learning, resource efficiency in computational costs, and the generalization of transfer learning across different tissue types. Findings from this study contribute to the field of medical imaging by demonstrating that transfer learning with GANs can significantly reduce the need for large datasets and extensive computational resources, offering an accessible, scalable framework for virtual staining across multiple histopathological domains.

Keywords: GAN, Transfer Learning, Virtual Staining

Contents

List of Figures

List of Tables

1 Introduction

This chapter highlights rationale for virtual staining, added with a very basic idea of transfer learning approach, that is the core of this thesis. Furthermore, possible research questions concerning transfer learning have also been raised, added with overview of research methods and objectives. Finally, there is also a section for policy related to application of generative AI usage.

1.1 Background

The main focus of the thesis is to evaluate and compare performance impact of model based on transfer learning techniques with a baseline model (that is trained from scratch), as a pre-trained model is believed to enhance the virtual staining process of histopathological images through generative adversarial networks (GANs). The available pipeline for virtual staining provides a promising approach to generate artificially translated images of tissues without the need for physical staining, reducing time and cost in biomedical imaging [1]. The main challenge is the large data (in quantity and also in size) and computational needs required to train models from scratch, which is where transfer learning becomes valuable by using pre-trained models in similar tasks [2], where we use pre-trained weights instead of training model from scratch. A conceptual architecture is shown in figure 1.1.

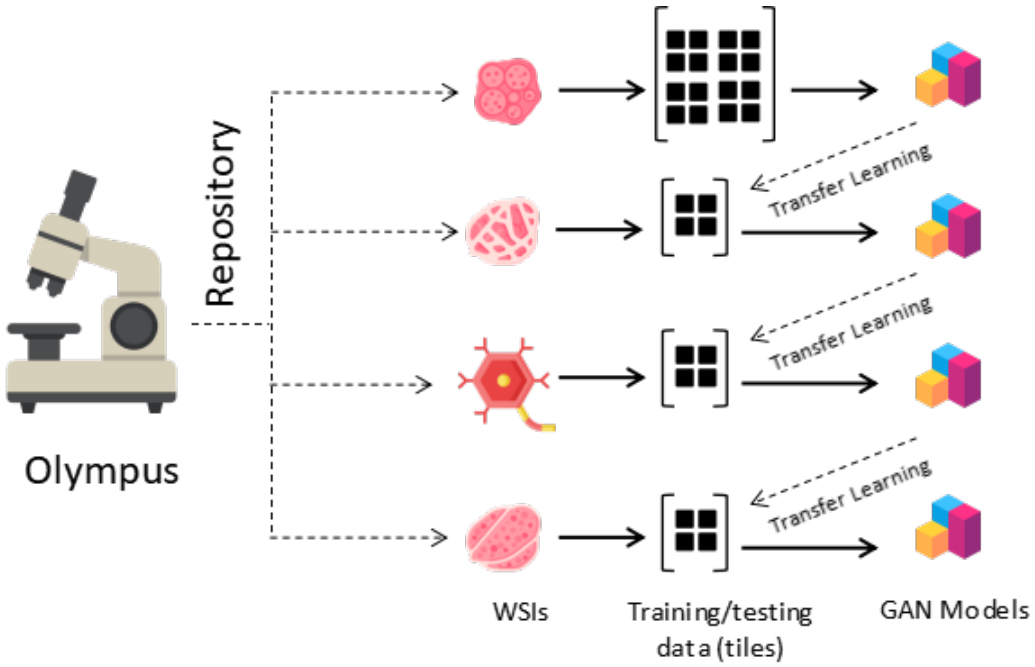

Figure 1.1: Transfer learning overview.

1.2 Problem Statement

The current approach to virtual staining of histopathological images, particularly in areas such as cancer diagnosis, is labor-intensive, requiring substantial human expertise and costly staining reagents [3] [4]. Moreover, generating high-quality stained images using GANs from raw histopathological data demands large amount of labeled data and computational resources. As a result, training GANs from scratch for each type of tissue or medical condition is inefficient and often yields suboptimal results [5] [6] [7]. Transfer learning presents a potential solution to this problem by allowing the use of pre-trained models to improve the quality of virtual staining and reduce training times [8] [9] [10] [11]. Thus, problem statement generates following research questions:

RQ-1: What are the software performance trade-offs between training models from scratch versus using transfer learning in virtual staining tasks?

RQ-2: How can resource efficiency be improved in terms of computational cost and

training time when using GANs for virtual staining through transfer learning?

RQ-3: Can transfer learning be effectively generalized across different types of tissues in histopathological images, or is domain-specific fine-tuning always required?

RQ-4: What are the main challenges and benefits of using transfer learning in medical imaging with limited datasets, especially in the case of histological staining?

Additionally, the practical and research objectives for this research is to create a transfer learning-based framework that can be applied to a range of histopathological tissues to generate virtually stained images, thereby reducing the time, cost, and expertise required for conventional staining.

1.3 Research Method

This section deals in following key methodologies that play a crucial role in achieving the research objectives:

1.3.1 Data Collection

Four types of histopathological tissue images (skin, kidney, spleen, and intestine) are sourced from the Cancer stress response laboratory (Latonen Lab), University of Eastern Finland. [1].

1.3.2 Base Model

GAN architecture (particularly Dense Pix2Pix), is used as the baseline models for virtual staining [1].

1.3.3 Training and Validation

The models will be trained using a limited amount of dataset, and the effectiveness of transfer learning will be evaluated by comparing training time, model accuracy, and the visual quality of the images.

1.3.4 Evaluation Metrics

The output will be evaluated using image similarity metrics, particularly Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Pearson Correlation Coefficient Ratio (PCCR) and Mean Square Error (MSE).

1.4 Thesis Structure

This thesis is organized into seven core chapters, each designed to build upon the previous one and collectively present a comprehensive study on transfer learning using Generative Adversarial Networks (GANs) for virtual staining. The **Introduction** outlines the motivation, raises the research various questions and establishes the research objectives. This is followed by the **Background** chapter, which reviews relevant literature in deep learning, GANs, and virtual staining techniques, providing the theoretical foundation for the research. The **Methodology** chapter describes the architectural design, data preprocessing strategies, and the transfer learning framework employed in the study. The **Experimental Setup** chapter details the experimental environment, including software libraries, and dataset configurations. Results are presented and analyzed in the **Results** chapter, while the **Discussion** provides a critical interpretation of the findings in relation to research hypotheses. Finally, the **Conclusion** summarizes the key contributions, identifies limitations, and outlines directions for future research.

Appendices and a comprehensive bibliography are included at the end to support and extend the core content.

1.5 Declaration of Generative A.I

The use of generative AI in the preparation of this thesis was minimal. It was limited strictly to minor language refinement, such as grammar correction or wording suggestions. No generative AI tool was used to produce or interpret any scientific content. All research

design, implementation, results, and conclusions presented in this thesis are entirely the original work of the author.

Summary: The chapter raised various research questions concerning transfer learning, followed by overview of research methods and objectives. It has also elaborated different components of thesis structure, which are really helpful in building good understanding about, not only the topic, but research ouput as well.

2 Literature Review and Background

This chapter gives a detailed perspective on relevant literature review, as well as background information on transfer learning in detail. Additionally, there is much more detail on topics that are tightly coupled with research topic, for example, Virtual staining, deep learning, and histopathology. Furthermore, it also discusses the entire transfer learning pipeline in great detail, including every step that is involved.

2.1 Background

In computational pathology and, particularly, in the context of virtual staining, the transfer learning technique is dependent on multiple other processes, each of which plays a crucial role. It starts from a rudimentary understanding of the difference between chemical and virtual staining, followed by deep learning, deep neural networks and generative adversarial networks (GANs). Following sections discusses each of these concepts in detail:

2.1.1 Virtual Staining

Virtual staining is a groundbreaking technique in the field of medical imaging and histopathology, offering a crucial and very important alternative to traditional histological staining methods [12] [13] [4]. Conventional staining involves the application of chemical reagents, such as Hematoxylin and Eosin (H & E), to tissue samples to enhance contrast and visualize cellular and structural components under a microscope. However, this process is time-consuming, labor-intensive, and potentially destructive

to tissue samples. Virtual staining bypasses these challenges employing computational algorithms to simulate the effects of chemical stains on digital images of unstained tissues [4] [1] [14] [12].

At its core, virtual staining transforms raw microscopy data into representations that closely resemble conventionally stained samples. This is achieved using advanced machine learning techniques, particularly deep learning [3]. By training algorithms on large datasets of stained and unstained image pairs, virtual staining models learn the complex mapping between the two domains. Generative Adversarial Networks (GANs) are particularly effective in this context, as they are capable of producing high-fidelity synthetic images that mimic traditional staining with remarkable accuracy [13], As shown in figure 2.1.

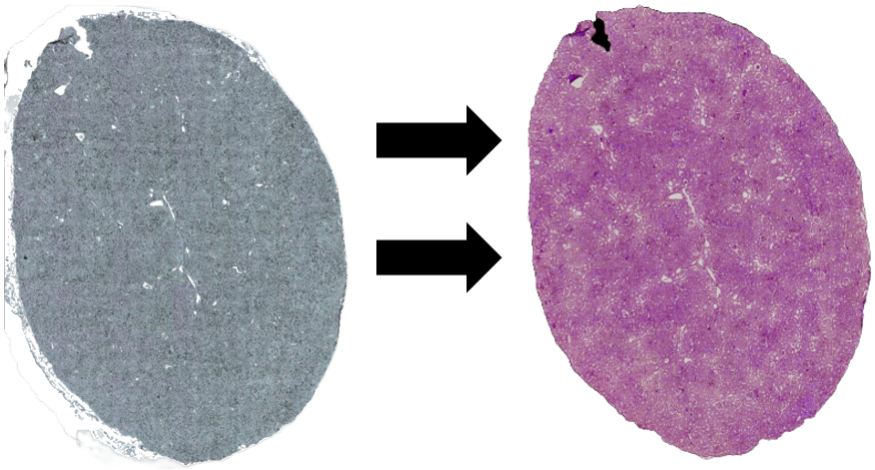

Figure 2.1: Virtual translation from unstained tissue to stained.

The advantages of virtual staining are manifold. It eliminates the need for costly and hazardous staining reagents, reduces the time required for tissue preparation, and preserves the integrity of tissue samples for further analysis. This technique also ensures consistency and standardization across different laboratories, addressing variability in staining quality that can arise from differences in protocols or human error [6] [1]. Virtual staining is particularly valuable in real-time diagnostic workflows, where rapid

and accurate visualization of tissue morphology is critical. Additionally, it enables the reanalysis of tissues without the risk of degradation or alteration caused by chemical staining.

Applications of virtual staining extend across diagnostic pathology, research, and education. It plays a pivotal role in cancer diagnostics, where accurate visualization of tissue architecture and cellular details is essential for identifying malignancies. Researchers benefit from the ability to analyze tissue dynamics over time without the need for destructive interventions. Moreover, in educational settings, virtual staining provides a cost-effective and scalable means of teaching histology and pathology. Overall, virtual staining represents a paradigm shift in histological imaging, combining the power of artificial intelligence with the precision of digital pathology.[15]

2.1.2 Hematoxylin and Eosin (H & E) Staining

Hematoxylin and Eosin (H & E) staining is the cornerstone of histological analysis and remains top choice for pathologist to do cell-level analysis. So is the case in this research work, as all our ground truth data (chemically stained images) are stained using H & E staining methods, which are essential for model training. This staining method has a primary function to provide contrast in tissue sections, allowing researchers and pathologists to analyze various cellular and tissue structures. This dual-dye method highlights cellular nuclei and cytoplasmic components in vivid detail, offering essential information for diagnosing diseases, particularly cancer [16] [4].

The staining process involves two distinct dyes with complementary roles. Hematoxylin is a dye, that affects acidic structures such as the nuclei of cells, leaving a deep blueish or purple color [17] [18]. This property makes it highly effective for highlighting genetic material, including chromatin patterns within the nucleus. On the contrary, Eosin is an acidic reagent that sticks to proteinaceous structures within the cytoplasm and extracellular matrix, staining them with various shades of pink. Together,

these dyes provide a comprehensive view of the tissue architecture, making it easier to identify abnormalities, cell types, and tissue organization.

H & E staining is indispensable in clinical settings, particularly in the diagnosis of cancer. For instance, it helps pathologists assess tumor margins, cell morphology, and the microenvironment. Despite its widespread use, the process is not without challenges. Traditional H & E staining is labor-intensive, requiring careful preparation, precise dye application, and meticulous slide handling. It also depends on skilled technicians and consistent quality control to ensure reliable results. Additionally, the chemical dyes and reagents used in the process may degrade over time, impacting the durability of stained slides [18].

Virtual staining addresses many of these limitations by digitally replicating the visual characteristics of H & E-stained slides [12]. This approach eliminates variability in staining quality, reduces dependency on physical reagents, and enhances the longevity of diagnostic samples. Furthermore, virtual staining enables the integration of computational techniques with traditional histopathology, fostering advancements in precision medicine and digital pathology.

2.1.3 Artificial Neural Network (ANN)

In deep learning, Neural Networks form the backbone of modern machine learning applications and are instrumental in virtual staining process. Artificial Neural Networks (ANNs) are computing systems which are modeled after the human brain's structure and operation. They are made up of layers of interconnected nodes, that are called as neurons, and work together to recognize patterns in data through learning. This kind of architecture make neural network capable enough to process complicated relationships between values, hence making them perfect choice for applications like computer vision, digital image processing, and, in this context, virtual staining. [19] [20] [21]

The structure of an ANN typically includes one input layer, followed by hidden layers (which can be one or more), and subsequently an output layer. The input layer starts

by reading the prompted dataset, such as imaging data from microscopy. The hidden layers process this input through neurons, each applying a mathematical operation using weights, biases, and activation functions [16]. Activation functions helps in adding non-linear functions, which makes the network to process highly complicated patterns, that are usually the case in image dataset. The output layer then provides the final result, such as a transformed image simulating a stained slide [22].

The training of ANNs involves adding increased efficieny in the weights so that errors are minimized between the predicted and actual outputs. This optimization is achieved through backpropagation, a process where the network iteratively adjusts its parameters based on the gradient of the error with respect to each parameter. Techniques like gradient descent are employed to ensure that the network converges to an optimal solution [23]. A typical ANN features three primary layers, named as input layer, hidden layer and output layer, as described in figure 2.2. Input layer consists of nodes (neurons) that receive various input signals representing features of the dataset (e.g., pixel values in image recognition or numerical features in regression problems). Hidden layer is very crucial part of the ANN, where computation and learning take place. Neurons in this layer process inputs using weights and activation functions, identifying complex patterns in the data. Finally, Output layer provides the final prediction or classification output of the network based on the processed information from the hidden layer.

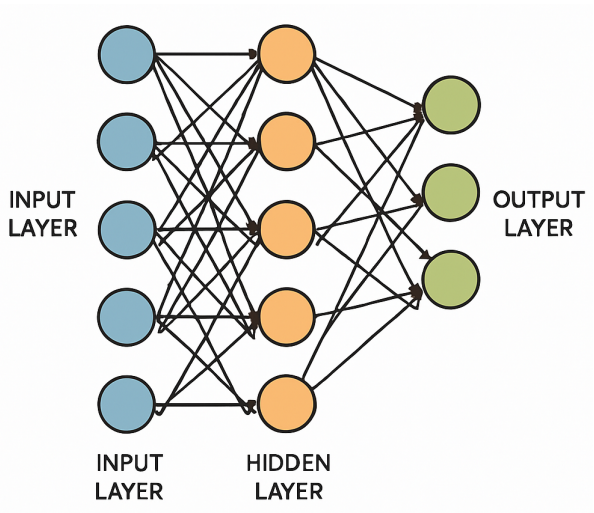

Figure 2.2: Diagram of an Artificial Neural Network (ANN) Architecture.

In the context of virtual staining, ANNs, particularly Convolutional Neural Networks (CNNs), are used to extract spatial features from raw microscopy images. These networks learn to recognize the recurring patterns (or even exclusive patterns) and features within the tissue, enabling them to apply the necessary transformations to simulate staining. For instance, in a label-free imaging setup, an ANN can process the input image, analyze its underlying physical properties, and output a digitally stained image that closely resembles its chemically stained image. [24]

The versatility of ANNs extends beyond image processing. They are employed in various domains, from medical diagnostics to autonomous vehicles, underscoring their transformative impact on technology and society. Their adaptability, scalability, and ability to learn from data make them indispensable for solving complex problems, including the challenges of virtual staining. [24] [19]

2.2 Transfer Learning

Transfer learning is a renowned technique in modern machine learning, where we configure models in such a way in order for them to use prior knowledge, that is sourced from one dataset and apply it to another type of dataset[25] [26] [27]. This approach has garnered a very good amount of attention, particularly in fields like image processing, digital pathology and computer vision, where training complicated models from start can be very hard, high demanding and time-consuming. In the domain of virtual staining, transfer learning has been a game-changer, enabling researchers to adapt pre-trained models to transform unstained tissue images into virtually stained equivalents with remarkable efficiency [8].

The fundamental idea behind transfer learning (and the practical motivation for this thesis) is that features learned by a model in one domain can often generalize to other domains. For instance, a neural network trained to recognize objects in everyday photographs, such as cats and cars, learns to detect fundamental visual patterns like

edges, corners, and textures [28] [10]. These learned features are not task-specific but rather form a foundational understanding of visual data. In transfer learning, this foundational knowledge is transferred to a new model tasked with a related but different objective, such as identifying cellular structures in histological images [8].

Transfer learning is particularly advantageous when the target task lacks a large dataset for training [26]. It is a very good practice to use high amount of training data for deep learning models. Doing so will amount to high performance results, which is not always feasible in specialized fields like histopathology. Thus, when we use pre-trained models (unlike training from scratch), transfer learning minimizes the need for extensive datasets while still achieving robust results [29] [30]. The concept of using pretrained weights has been demonstrated in figure 2.3.

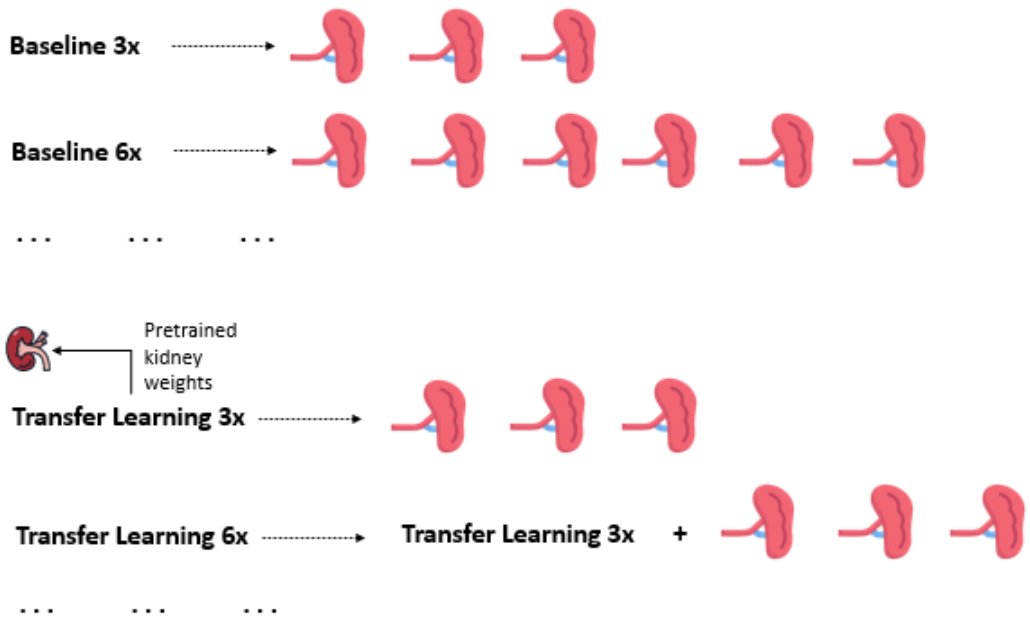

Figure 2.3: Conceptual representation of Transfer Learning models accross two tissue types.

In virtual staining workflows, pre-trained models, that are designed to perform digital image processing operations, serve as the foundation. The models that are trained on large datasets, have already learned to recognize fundamental image features. During the transfer learning process, the model's weights (parameters) are fine-tuned on the

specific dataset of unstained and stained image pairs. This fine-tuning allows the model to specialize in the nuances of histological images while retaining the generalization capabilities acquired from the pre-trained task. [26] [8] [29]

2.3 Generative Adversarial Networks (GANs)

GANs are artificial neural network (ANN) based deep learning frameworks designed by Ian Goodfellow[6]. In a very broader sense, GANs consist of two primary components which are called as 'Generator' and the 'Discriminator'. Both of these are trained simultaneously by competing against each other, hence called 'adversarial' process [6] [7]. The Generator generates the image data that is intended to come from a certain distribution, while the Discriminator process the data, evaluates it, and tries to distinguish if it is real data from the training set or if it is fake one generated by generator. So there is a competition going on between generator and discriminator. [7][6] [31] [32].

2.3.1 Contemporary Models:

Several models are central to the implementation of virtual staining, each tailored to specific aspects of the task. Convolutional Neural Networks (CNNs) form the foundation for most virtual staining pipelines, as they excel in processing image data and extracting spatial features [33]. GANs, particularly variants like Pix2Pix, are widely used for image-to-image translation tasks, enabling the transformation of label-free microscopy images into virtually stained representations [13].

Pre-trained models such as DenseNet are also employed in transfer learning workflows, providing a robust starting point for fine-tuning on histological datasets. Custom architectures that combine elements of CNNs, GANs, and other deep learning components are also developed to address specific requirements in virtual staining, such as multi-modal imaging or high-resolution synthesis [13] [34].

2.3.2 DensePix2Pix:

The model that is used in this research is 'DensePix2Pix'. It basically combines the architectural strength of DenseUNet with the adversarial learning strategy of Pix2Pix. At its core, it leverages a DenseUNet-based generator that integrates dense connections within a U-Net-style encoder-decoder framework. These dense connections allow for better feature propagation and reuse, this improves model's ability to understand features of image in detail, while maintaining global context. The generator receives an input image (such as a semantic map, grayscale scan, or edge map) and learns to produce a corresponding output image with high fidelity.[35][36][1]

On the adversarial side, DensePix2Pix follows the conditional GAN setup from Pix2Pix. The discriminator evaluates not just whether an image is real or fake, but whether the generated image is a plausible match to the given input. This pushes the generator to produce more realistic and contextually accurate outputs. By combining the dense feature extraction power of DenseUNet with the structured image-to-image learning approach of Pix2Pix, DensePix2Pix achieves more detailed and sharper results, making it particularly effective in applications like medical image synthesis, satellite image enhancement, and semantic segmentation-to-image tasks.[35][36]

Further detail on it's different layers in context of Generator and Discriminator architecture of DensePix2Pix is stated below:

2.3.3 Generator Architecture

The Generator in a GAN is, again, typically a set of network layers that takes random noise as input and generates data that mimics the real data distribution. This is shown in figure 2.4. In the context of virtual staining for H and E (Hematoxylin and Eosin) staining, the Generator would take as input unstained tissue images and produce images that look as if they have been H and E stained[7][6].

The Generator gets an input of vectors with random noise, often sampled from a normal distribution. This noise vector serves as the seed from which the Generator creates the fake images. Hidden layers are composed of several convolutional layers (in

the case of image data). They progressively upsample the input noise to higher resolution feature maps. After each convolutional layer, ReLU (Rectified Linear Unit) [37][6] [38], is used. Batch Normalization is often used between convolutional layers in order to optimizie the training process by normalizing the inputs to each layer. This helps in minimize covariation shift and speeds up the training process [39][6] [7]. The final layer of the Generator is a convolutional layer that has a tanh activation function[38], which scales the output to the range [-1, 1], matching the range of the real images [7].

2.3.4 Discriminator Architecture

Similar to Generator, Discriminator is a core component of GANs, as it reads both, the ground truth (H and E stained tissue images) and fake images (generated by the Generator) as input and outputs a predictive probability that hints if this resulting image is real or fake[6] [7]. This is shown in figure 2.5.

The input to the Discriminator is an image, either real or generated. In the case of H and E staining, this would be a tissue image. The hidden layers of the Discriminator are composed of several convolutional layers that progressively downsample the input image to lower resolution maps. Again, similar to generator, this convolutional layer is succeeded by a LeakyReLU activation [38] [37][40]. Similar to the Generator, batch normalization is often used between convolutional layers in the Discriminator to stabilize training. The final layer of the Discriminator is usually a dense layer with a sigmoid activation function, which outputs a probability indicating whether the input image is real or fake.

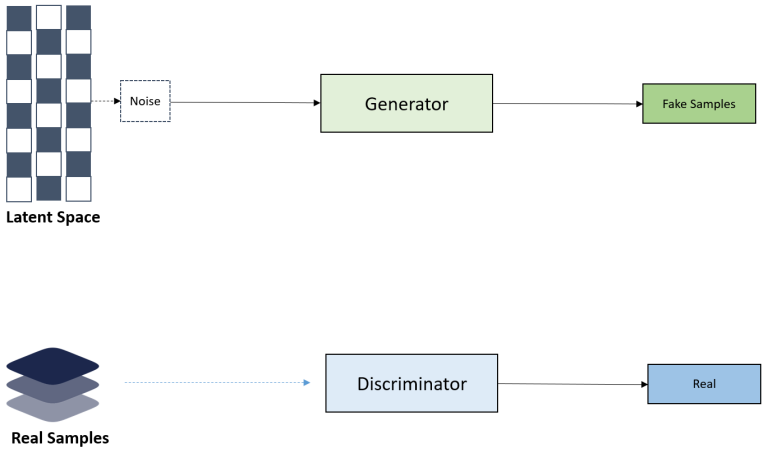

Figure 2.4: Generator generates fake images, while discriminator identifies the images as fake or real [6][7].

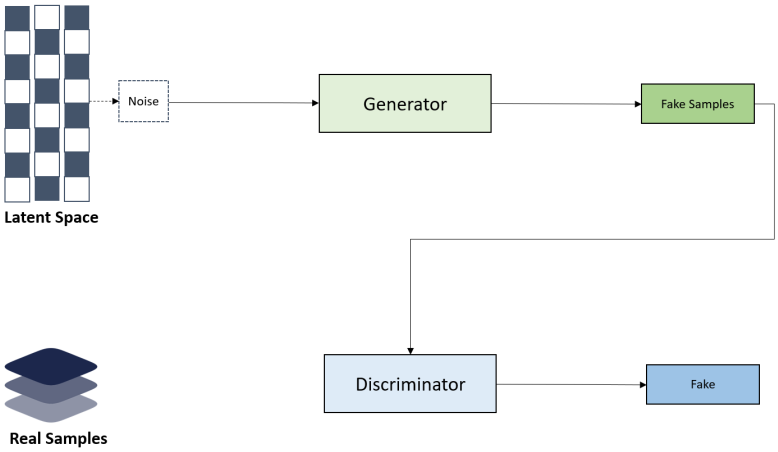

Figure 2.5: Generator feeds fake images to discriminator, while discriminator specifies if the image is real or fake[6].

2.3.5 Loss Function

Loss functions are supposed to be minimized when we are training a GAN model. This is mandatory becuase it captures the performance of both the Generator and the Discriminator. Generator's role is to reduce the chance that Discriminator actually identifies its output as fake. This is typically achieved by minimizing the binary cross-entropy loss between the Discriminator's output for the generated images and a vector of ones (indicating that the images are real). Discriminator tries to maximize

the chance of identifying the fake images. This is achieved by minimizing the binary cross-entropy loss. The training of GAN models is repetitive process where both (Generator and Discriminator) are trained alternately [6][7]. In each iteration, the Discriminator is first trained on a set of actual images as well as a set of fake images that are generated by the Generator. The Discriminator's weights are updated to minimize its loss function. After Discriminator iterations are done, the Generator is trained so to produce images that gives an impression of real images and are more likely not to be detected by the Discriminator. The training process continues until the Generator produces images that are largely different from real images, and the Discriminator is unable to differentiate between real and fake images[6].

2.4 Staining Pipeline

This section describes a complete process from scratch on how an image, that is sourced from microscope (or any imaging device), is processed through our pipeline (as shown in figure 2.6), and then is converted into stained image on the basis of provided supervised learning parameters. It is not possible to widen the scope in such a way that can cover semantic details, like programming paradigms in practice, or in-depth analysis of tech stack. Rather, it only covers essential aspects of logical design and instead, gives detailed insight on actual staining process and the supervised learning techniques that have been employed to achieve expected results.

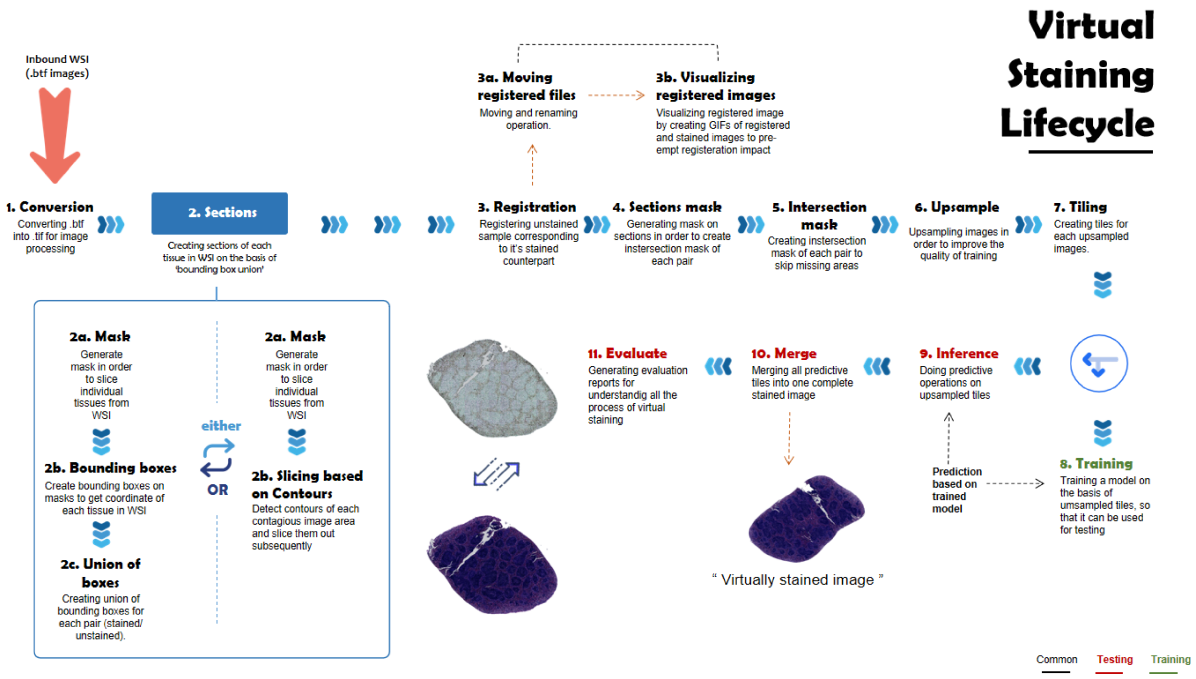

Figure 2.6: Virtual staining pipeline lifecycle

Followings are the important steps involved in this processing pipeline:

2.4.1 Conversion

Deals with converting inbound images into .tif fromat. Inbound image dataset can be in multiple formats, for example .vsi/.btf etc. Converting different image types into tiff files help a lot in python based image handling.

2.4.2 Sectioning

Image dataset is available in form of WSI (whole slide images), meaning that one WSI can contain multiple tissue images. In order to extract individual tissue image from WSI, we slice up each section one by one. In order to achieve this, we use edge detection technique from pythons popular openCV library, which detects each image by detecting contour of each image component.

2.4.3 Registration

Once sectioning is done, each stained and unstained image is then registered. The process refers to mapping of unstained over fixed image of each sample. This helps in verifying if both of the images align perfectly and if they are part of same sample or not (shown in figure 2.7). We use WSIreg bundle for registration process, which uses elastix to define arbitrary transformation paths between associated images.

Once registration process is done, pipeline has several steps to visualize the registration in order to visually verify if the registration process has yielded rightful image registration or not.

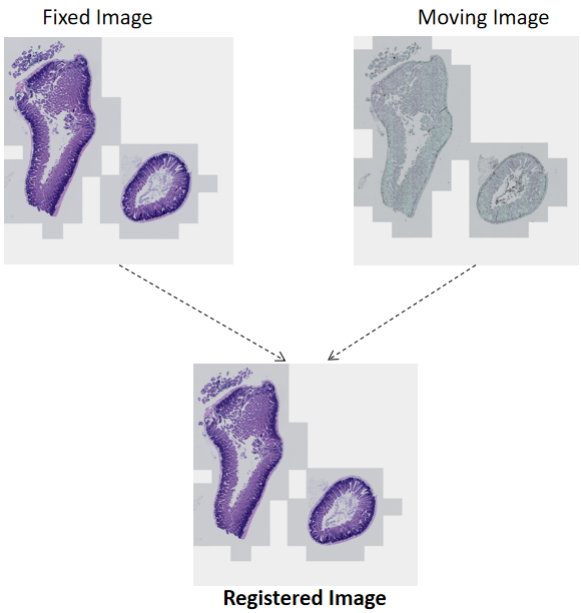

Figure 2.7: Image registration based on ground truth and unstained pair

2.4.4 Masking

After successful registration, masking of each sample (both stained and unstained images) is done. Masking is done in two steps: in first step, each section is masked; in second step, intersection mask of stained and unstained image sample is done in order to fetch least common areas in both. This intersection of masks can be seen in figure 2.8.

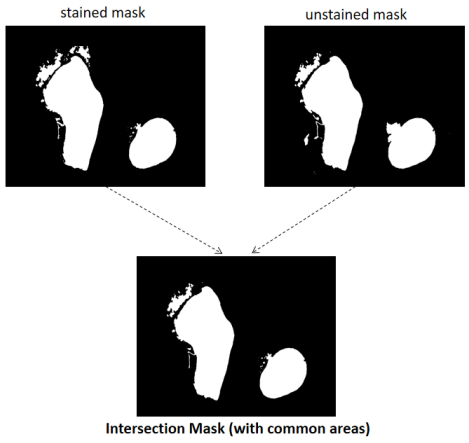

Figure 2.8: Intersection mask with common areas

2.4.5 Tiling

Once we have intersection, we perform some rudimentary scaling up operations, for example saving processed files, upscaling intersection mask to match the dimensions of actual image data etc. Once scaling up is finalized, we do tiling of each available image section. These tiles can be customized based on preferences, for example tile size, stride size etc (as shown in figure 2.9). These configuration defines how much tiles we can have. Once this step is completed, our image data is now ready for training the model.

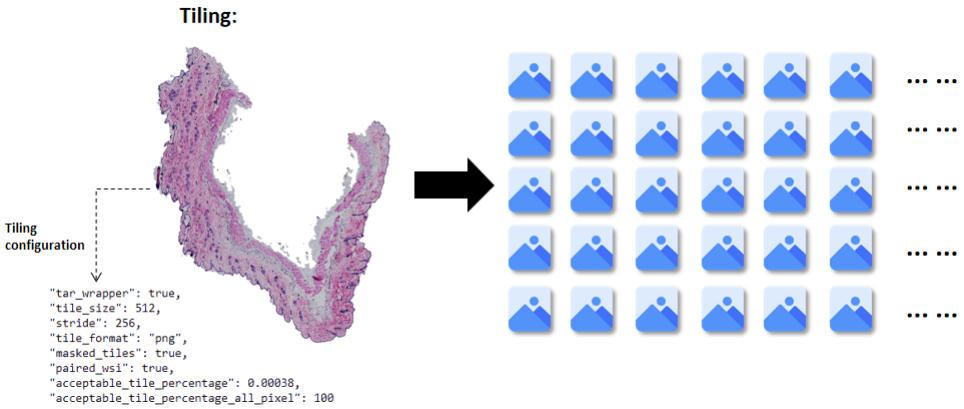

Figure 2.9: Tiling step involves breaking down larger images into smaller specified tiles

2.4.6 Training

This pipeline uses vanilla DensePix2Pix, added with Dense UNET layer. This layer gives leverage to enforce deep supervision by sharing information related to loss function across different layers in entire neural network. Once all the tiles are generated, this model is trained on these tiles over the period of 40 epochs with the batch size of 10. Same process is repeated for inference as well, that uses unseen data to test the predictions based on above trained model.

2.5 Previous Work

This section discusses impact of transfer learning on existing applications in digital pathology, particularly in context of virtual staining and histopathology.

2.5.1 Transfer Learning in Medical Image Analysis

Transfer learning basically deals with pre-trained models, to initialize the parameters of a model for a specific medical imaging task [28]. This strategy significantly boosts model performance by reducing the need for large amounts of annotated data. Pre-trained models, such as DenseNet, architectures, are adapted to medical imaging tasks, either by fine-tuning their fully connected layers to specific histopathological or MRI features or by incorporating them as feature extractors in a deeper network architecture. These models help in learning general features like textures and shapes, which are critical for accurate diagnosis and classification in medical images [26].

2.5.2 Stain-Style Transfer and Domain Adaptation

In histopathological image analysis, stain-style transfer (SST) is an essential technique that adjusts the color and texture of images to standardize them for better alignment across different datasets [41]. This technique involves using GANs to transfer stain styles between images, thus enabling better cross-institutional compatibility and improving model performance [42]. The use of pre-trained models like ResNet for feature extraction

in these GANs further enhances the style transfer, making the model more effective in tasks like tumor classification [43].

2.5.3 Fine-Tuning Strategies

Fine-tuning is a prevalent strategy across these studies, wherein the pre-trained models are adapted for specific medical image tasks. This process involves adjusting the weights of the model's higher layers to better fit the medical image data. Fine-tuning not only helps in capturing task-specific features but also improves the model's efficiency in identifying patterns related to specific medical conditions such as brain tumors or metastasis in histopathological images [44] [45] [46].

2.6 Challenges and Future Directions

Despite their effectiveness, these methods face challenges such as dealing with the domain shift between training and test datasets, the need for large-scale annotated datasets, and the interpretability of model decisions. Future research could focus on developing more sophisticated domain adaptation techniques, improving the interpretability of models in medical contexts, and addressing the ethical implications of using synthetic data in clinical applications [45].

Overall, the integration of transfer learning, GANs, and stain-style transfer into medical image analysis represents a significant advancement in the field, enabling more accurate and efficient diagnostic tools. These methodologies collectively contribute to better handling of medical image data, facilitating improvements in patient care and clinical outcomes [10] [27].

Summary: This chapter gave an in-depth analysis of all the related work and academic material, along with core concepts that have been employed in the course of this research.

3 Methodology

This chapter mainly discusses the semantics of transfer learning. These semantics include information on the data source and comprehensive detail about the tech stack that has been used.

3.1 Datasource and Preprocessing

The data is sourced from an Olympus scanner at the University of Eastern Finland (UEF) [1] [47]. The dataset comprises of whole slide images (WSIs) of kidney, skin, intestine, and spleen tissues. Each WSI contains multiple tissue sections, and for baseline training, we have paired data: unstained tissue images and their corresponding H and E stained (ground truth) images [47]. The preprocessing pipeline is critical to ensure that the data is in a suitable format for training the GAN model. Below, we delve into the detailed steps involved in data sourcing, preprocessing, and the rationale behind each step.

3.1.1 Data Sourcing

This section provides syntactic information about structure of the data.
The data is acquired using an Olympus scanner, which is a high-resolution imaging device capable of capturing detailed images of tissue samples. The scanner produces WSIs, which are large digital images that encompass entire tissue sections at high magnification [48]. The dataset includes WSIs of four different tissue types: kidney, skin, intestine, and spleen. Each tissue type has unique morphological features, but they also share common characteristics, such as cellular structures and tissue organization, which make

transfer learning feasible [1] [47]. The dataset for training in this case consists of paired unstained images and the matching H & E-stained images. The model learns by example in supervised learning, which is why we have decided to employ it. The model can learn precise mappings as paired data gives each input the right output. The model is unable to determine what the "correct" transformation looks like absence of these pairs. [47].

3.2 Tech Stack

The tech stack for our image-to-image translation pipeline includes a variety of tools, libraries, and frameworks designed to handle different aspects of image processing, machine learning, and model deployment. These technologies help us achieve state-of-the-art results in tasks such as image synthesis, segmentation, and transformation. Below is a detailed description of the components of this stack:

LUMI - Every operation is run on LUMI (Large Unified Modern Infrastructure) super computer. This helps greatly in processing data sets with high payload without any hassle, which would have been nearly impossible otherwise. We can communicate LUMI through a web-interface, as well as command prompt (in case of windows). LUMI has job scheduling system (often called as SLURM- Simple Linux Utility for Resource Management), using which users can write scripts to define how their computational jobs should be executed, specifying resources like CPU cores, memory, runtime duration etc [49]. The control flow of one LUMI instruction, packaged with python and .json files, is exemplified in figure 3.1

Shell files - Shell scripts (**.sh** files) are used for automating tasks related to data preprocessing, model training, and evaluation. In the context of this project, `.sh` files are utilized to streamline workflows, set up the environment, run the model training processes, and perform other repetitive tasks in the pipeline. By automating these processes, we ensure consistency, save time, and reduce the potential for human error [50].

Python - Python is the primary programming language used in our pipeline for implementing deep learning algorithms and managing data transformations[51]. Libraries

like PyTorch, Tensorflow, Keras, Pillows etc. have been used for applying data science methods in order to achieve our image processing targets.

JSON files - JSON (JavaScript Object Notation) files are used for storing configuration settings and model parameters in a readable, human-friendly format. These files typically contain hyperparameters for model training, paths to datasets, batch sizes, learning rates, and other important configurations. By using JSON, we can easily modify and share the configuration settings without having to alter the core code of the model itself.

Singularity - *Singularity* is a containerization technology that allows us to create reproducible, isolated environments for our machine learning workflows. Containers provide a consistent execution environment, making it easier to deploy and share code across different platforms without worrying about system dependencies or version conflicts. By using Singularity, we can ensure that our models are portable and can be run consistently across various machines, including cloud platforms and high-performance computing clusters like LUMI [52].

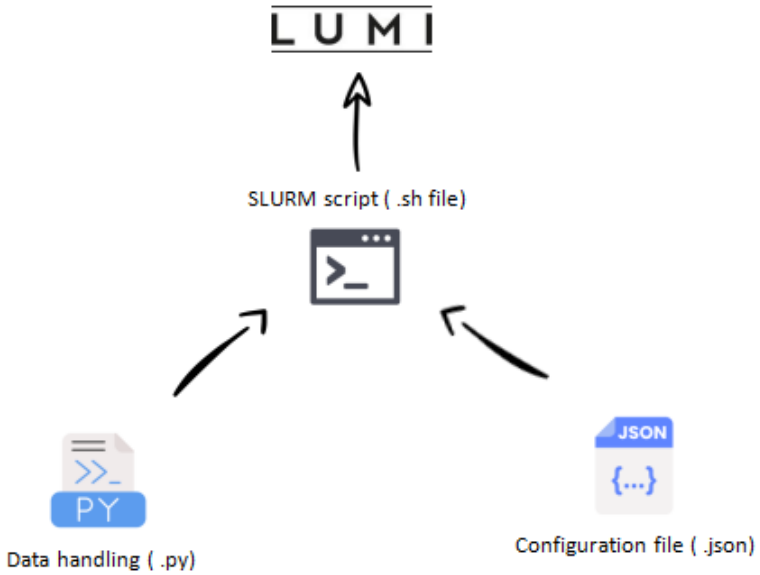

Figure 3.1: Control flow of instruction on LUMI

4 Experimental Setup

For virtual staining to be effective, the model must produce high-quality, clinically useful images that accurately represent the tissue's architecture, color fidelity, and morphological characteristics. The quality of the produced images must, hence, be assessed using structural, textural, and perceptual criteria. The experimental setting is thoroughly described in this chapter, along with the dataset preparation, preprocessing procedures, model training techniques, and assessment standards used to gauge the effectiveness of virtual staining models.

4.1 Dataset and Preprocessing

The choice of dataset and appropriate preprocessing steps is critical to ensuring the success of deep learning models, especially when working with high-resolution images such as Whole Slide Images (WSIs) in histopathology [48]. While preprocessing reassures that the data is in a suitable format for training, a representative dataset makes it possible the model to generalise successfully throughout various tissue types.

4.1.1 Whole Slide Image Selection

The dataset used for model training and evaluation is obtained from the Latonen Lab and contains WSIs from four distinct organ types: skin, kidney, spleen, and intestine. These organs were selected to represent a diverse range of histological features, which helps test the robustness of the virtual staining model across varying tissue types. WSIs are high-resolution images that offer a comprehensive view of tissue morphology. These

images are critical in histopathology, as they provide detailed, full-slide resolution of the tissue sample. The dataset is divided into three subsets. Firstly, the training set comprises a varying number of WSIs, (including configurations with experiments for 3, 4, 5 and so on until 10 images). We have named these images as 'sections' and training a model over each number of sections represent a 'cycle'. Forexample, if a model has been trained for 4 sections, it is called as 4-cycle model. This variability allows us to test how dataset size influences the model's performance, with smaller sets focusing on low-data scenarios and larger sets providing more diverse tissue samples. Secondly, the validation set is used for hyperparameter tuning, model selection, and regularization. It is kept separate from the training set to prevent overfitting. It is to note that for each cycle, there is one dedicated validation section (WSI) which is neither part of training dataset nor testing data. Lastly, WSIs not used for model validation or training make up the test set. It functions like an independently measured metric of the model's capacity to generalise to previously unknown data.

4.1.2 Preprocessing Steps

The preprocessing of WSIs is essential for ensuring that the images are suitable for deep learning tasks. Given the large size and complexity of WSIs, preprocessing helps reduce computational load and standardize input data. Firstly, tissue segmentation helps identifying and extracting the tissue regions from the WSIs. Non-tissue areas, such as background or irrelevant sections, are discarded to ensure the model focuses on relevant tissue features. Thresholding method has been used in our experiment, in which we have specified the percentage of acceptance that any image data will have in a image/tile. After segmentation, due to the large size of WSIs, images are split into smaller patches (512x512 pixels in our case) for computational efficiency. This also allows the model to focus on local features, which is important in virtual staining, where fine details of tissue structures and staining patterns are critical. Lastly, to make the training data more variable, we will employ data augmentation techniques including flipping, random rotations, and contrast modifications. This decreases the possibility of overfitting while

improving the model's generalisation. By simulating various orientations and staining condition modifications, augmentation improves robustness of the current model.

4.2 Model Training Strategy

We have already covered the applications of discriminators and generators in GANs. Unstained tissue samples are used to make images by the generator, and the discriminator determines whether the images are genuine by identifying both generated and real images. Both of these networks compete with one another, which causes the generator to produce increasingly realistic and effective virtual stained images. The training methods for the baseline model and the transfer learning approach are described in the following.

In addition to that, spleen tissue has been selected for training and evaluation of proposed transfer learning experiments. There are total of 12 sections of spleen tissue in pairs (stained and unstained) that has been divided among each training cycle. Distribution of these WSIs in each cycles are discussed in following subsections.

4.2.1 Baseline Model

The baseline GAN model is trained from scratch for the spleen tissue, meaning that both the generator and discriminator networks are initialized with random weights. This goes for each cycle that we have trained for, starting from 3 until 10-cycle. This will help us evaluate the performance between TL-model compared to the performance of baseline models, that are trained from scratch. As far as how the data is distributed among each cycle for baseline experiments, figure 4.1 maps all the distribution accross every cycle:

C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
sample 0	sample 0	sample 0	sample 0	sample 0	sample 0	sample 0	sample 0
sample 1	sample 1	sample 1	sample 1	sample 1	sample 1	sample 1	sample 1
sample 2	sample 2	sample 2	sample 2	sample 2	sample 2	sample 2	sample 2
n-a	sample 3	sample 3	sample 3	sample 3	sample 3	sample 3	sample 3
n-a	n-a	sample 4	sample 4	sample 4	sample 4	sample 4	sample 4
n-a	n-a	n-a	sample 5	sample 5	sample 5	sample 5	sample 5
n-a	n-a	n-a	n-a	sample 6	sample 6	sample 6	sample 6
n-a	n-a	n-a	n-a	n-a	sample 7	sample 7	sample 7
n-a	n-a	n-a	n-a	n-a	sample10	sample 8	sample 8
n-a	n-a	n-a	n-a	n-a	n-a	n-a	sample 9
sample 10	sample 10	sample 10	sample 10	sample 10	sample 10	sample 10	sample 10
sample 11	sample 11	sample 11	sample 11	sample 11	sample 11	sample 11	sample 11

Figure 4.1: Baseline distribution of 12 WSIs across experimental cycles (C-3 to C-10). Validation samples (sample 10) are shaded green; test samples (sample 11) are shaded orange.

4.2.2 Transfer Learning model training

This particular model training setup is core of this thesis. The weights have been loaded from a model that was trained using kidney tissue and these pre-trained weights, henceforth, have been used in new training of GAN model (DensePix2Pix), and thus has been used for training with each cycle (3 to 10) for transfer learning. The reason to choose kidney as source weight and spleen as target tissue is that these two are structurally similar, as shown in figure 4.2. Alternatively, an overview of structural differences between all available dataset types are shown in figure 4.3. This image describes how tissues have different structural representation.

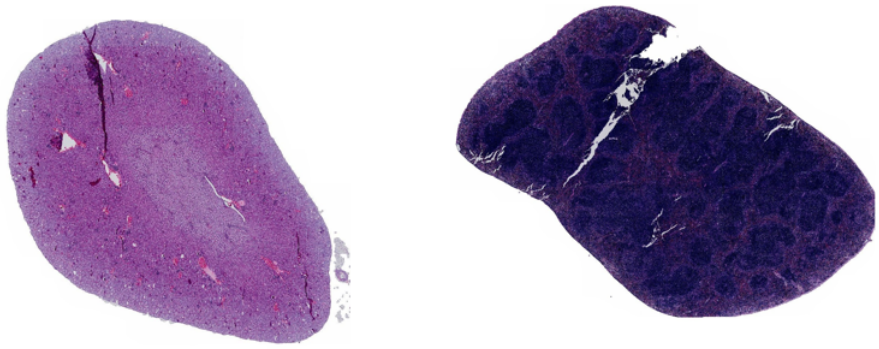

Figure 4.2: Kidney structure (left) compared to Spleen structure.(right)

Figure 4.3: Kidney structure (leftmost) compared to Intestine (mid-left), Skin (mid-right) and Spleen structure.(rightmost)

There are evidently various structural similarities among these two tissue, like similar nuclei structures, hence providing high potential of promising results. The distribution of sections accross each cycle is shown in figure 4.4:

C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
sample 0	C-3	C-4	C-5	C-6	C-7	C-8	C-8
sample 1	sample 3	sample 4	sample 5	sample 6	sample 7	sample 8	sample 9
sample 2	n-a	n-a	n-a	n-a	n-a	n-a	n-a
sample 10	sample 10	sample 10	sample 10	sample 10	sample 10	sample 10	sample 10
sample 11	sample 11	sample 11	sample 11	sample 11	sample 11	sample 11	sample 11

Figure 4.4: Transfer learning based distribution showing pretrained model usage (light blue), validation (green), and test (orange) WSIs across experimental cycles (C-3 to C-10)..

It is to note that training GANs from scratch can be computationally intensive, especially with high-resolution WSIs. This is also one of the main reasons that a transfer learning approach is used, where a pre-trained GAN model—initially trained on some previous data is fine-tuned on the some other tissue dataset. This transfer of knowledge allows the model to leverage learned features from the pre-trained model, reducing the amount of training time required and enhancing the model's ability to generalize to new tissue types. By using pre-trained weights, the model can focus on learning domain-specific features related to spleen tissue, while maintaining the generalized features learned from kidney tissue.

4.2.3 Training Parameters and Optimization

Training the GAN model involves optimizing key hyperparameters that govern the learning process. These hyperparameters include the learning rate, batch size, and the choice of optimizer. Firstly, A batch size of 16 was selected, balancing memory requirements and training efficiency. Larger batch sizes would require more memory, while smaller ones would increase the variance in gradient estimates. Also, The generator is impacted by the losses to create realistic images that are also architecturally and texturally compatible with actual stained tissue. These losses include perceptual, style, content, and antagonistic losses. Finally, The model was trained for 40 epochs to allow sufficient time for convergence while avoiding overfitting.

4.3 Evaluation Criteria

Evaluating the quality of virtual stained images requires a set of metrics that assess different aspects of image quality, such as structural similarity, color fidelity, and pixel-wise accuracy. The following metrics were used to evaluate the performance of the GAN model:

4.3.1 Structural Similarity Index (SSIM)

SSIM is a perceptual quality metric that measures the structural similarity between the predicted and reference images. Unlike pixel-wise metrics such as Mean Squared Error (MSE), SSIM incorporates luminance, contrast, and structural information, making it more appropriate for evaluating histopathological images [53]. Since histopathological images contain complex tissue structures, maintaining structural integrity is crucial. The SSIM formula is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Where: μ_x and μ_y represent the mean intensities of images x and y.
σ_x^2 and σ_y^2 are the variances of the two images.
σ_{xy} is the covariance between the two images.
C_1 and C_2 are constants to stabilize the division with weak denominator values.

SSIM is essential for ensuring that the generator preserves the fine tissue structures, which is vital in histopathology. Small distortions in cellular structures may lead to incorrect diagnoses .[53] [54]

4.3.2 Mean Squared Error (MSE)

MSE is a traditional metric that measures the pixel-wise differences between the predicted and reference images [54]. It is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

Where N is the total number of pixels, and x_i, y_i represent the pixel intensities of the predicted and reference images. MSE provides a straightforward measurement of pixel-wise accuracy but may not always reflect perceptual quality, as it does not account for human visual perception.

4.3.3 Peak Signal-to-Noise Ratio (PSNR)

PSNR evaluates the quality of the generated images by comparing the ratio of the maximum signal to the background noise [54] [55]. It is defined as:

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right)$$

Where MAX is the highest possible pixel intensity. A higher PSNR indicates that the generated image has fewer artifacts and distortions.

4.3.4 Pearson Correlation Coefficient Ratio (PCCR)

PCCR measures the linear correlation between the pixel intensities of the predicted and reference images [56]. It is defined as:

$$PCCR = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where $\bar{x}$ and $\bar{y}$ represent the mean pixel intensities of the two images. PCCR is particularly useful for evaluating color accuracy in histopathological images, where the correct differentiation of tissue colors (e.g., nuclear and cytoplasmic staining) is essential for accurate diagnosis.

4.4 Usecase for Multiple Metrics

Histopathological images are inherently complex due to the rich and intricate nature of biological tissues. These images contain detailed structural patterns, varying color distributions, and fine-grained textures that make them difficult to analyze using traditional image processing techniques. In recent years, machine learning models, particularly deep learning techniques, have shown promising results in tasks like segmentation, classification, and virtual staining of histopathological images. However,

these models must be evaluated accurately to ensure their practical utility in medical diagnostics. A significant challenge in evaluating these models lies in the fact that histopathological images contain diverse features that are not always captured adequately by a single evaluation metric.

A single metric has a very high probability of producing an inadequate result, meaning that the image quality will be limited. Pixel best measures, such as Mean Squared Error (MSE) or Peak signal-to-noise ratio (PSNR), can also be used. They may help determine the pixel's precision, but they do not always represent the quality of accepted human vision. Additionally, there are perceptual measures that are more in line with human vision, such as the Structural Similarity Index and Pixel Wise Colour Production Correlation, which place greater emphasis on features, colours, structural integrity, and fidelity. Thus, this study is using multiple matrices to capture a wider range of features and quality of the virtual staining to have a more accurate outcome, where the picture may catch more features of the specific cell. It will also assist in reflecting how well the model performs in various situations. We can achieve pixel-wise fidelity by combining MSE and PSNR. In this case, MSE will calculate the average square difference between the relevant pixels in the ground truth and the forecasted image. Because it displays minuscule details and captures the pixel-by-pixel quality of the cell features, this matrix allows us to find extremely sensitive, detailed data in the images. They use PSNR to capture the image's noise or, in the event of distortion, record the details of the image. The primary drawback of both matrices is that the image they produce must be compatible with the human visual system. Perceptual elements like texture and structure, which are essential in medical picture analysis, may not always be considered because they rely on pixel-level computations. The Structural Similarity Index (SSIM) is a crucial statistic in the assessment of histopathological images to overcome the above limitations. By comparing structural information between the reference and predicted images, SSIM calculates the image's perceived quality. According to how people judge visual quality, the Structural Similarity Index (SSIM) compares two images based on their structural similarity (Luminance, Contrast, Structure), unlike MSE, which only examines

pixel-by-pixel variations. Therefore, SSIM eliminates the issue with MSE and PSNR. These factors are critical in histopathology, where preserving the structural integrity of tissue samples is essential for accurate diagnosis. For example, subtle changes in the arrangement of cells or tissue patterns may indicate the presence of disease, and preserving this structural information is crucial for a successful virtual staining method.

Similarly, the Pearson Correlation Coefficient Ratio (PCCR) is another important perceptual metric used to assess how accurately the virtual staining method replicates the color distribution of the ground truth. In histopathology, the color of stained tissue plays a vital role in differentiating between various types of tissue and cells. Accurate color reproduction is crucial for tasks such as tumor detection or distinguishing between benign and malignant tissues. While MSE and PSNR do not explicitly address color fidelity, PCCR evaluates how well the virtual stain captures the subtle color differences that are essential for diagnostic purposes.

Therefore, using a multi-metric approach that combines both pixel-level metrics like MSE and PSNR with perceptual metrics such as SSIM and PCCR ensures that we are capturing all relevant aspects of image quality. This approach provides a more comprehensive evaluation of virtual staining methods, ensuring that the model not only produces accurate pixel-wise reconstructions but also preserves the structural and color features that are vital for medical interpretation.

Summary: This chapter gave in-depth analysis of exactly how and what is going to happen in order to achieve results. It described the data, performance metrics, evaluation criteria and experimental configuration. This helps in providing meaningful approach for assessing the results, which are discussed in next chapter.

5 Results

This chapter does not only show the details of experimental results, but it also gives in-depth analysis of what the provided results mean. Detailed graphs, and images are also added so to reach a meaningful conclusion in later chapters. Additionally, details on evaluation metrics is also stated, that gives a very good insight on the impact of results in context of initial hypothesis.

5.1 Comparison of Transfer Learning vs Non Transfer Learning Approaches

The effectiveness of transfer learning (TL) is evaluated by comparing performance metrics across cycles with different dataset sizes: starting from 3 WSIs in squence up until 10 WSIs. The primary objective is to determine whether leveraging knowledge from a pre-trained model leads to better virtual staining quality than training from scratch (baseline, BL). The comparison is conducted across four key evaluation metrics: Structural Similarity Index (SSIM), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Pearson Correlation Coefficient Ratio (PCCR).

5.1.1 Analysis of Structural Similarity Index (SSIM)

SSIM evaluates perceptual quality by measuring structural consistency between the virtual stain and the ground truth. A higher SSIM value indicates more faithful reconstruction in terms of structural details.

Cycles	SSIM	MSE	PSNR	PCCR
cycle3	0.5297	1993.16	15.13	0.6256
cycle4	0.5312	1972.00	15.19	0.6280
cycle5	0.5328	1948.00	15.26	0.6320
cycle6	0.5343	1925.00	15.32	0.6360
cycle7	0.5358	1900.00	15.38	0.6410
cycle8	0.5373	1875.00	15.44	0.6470
cycle9	0.5389	1860.00	15.49	0.6540
cycle10	0.5405	1845.00	15.54	0.6646

Table 5.1: Baseline Learning (BL) performance across different training cycles

Cycles	SSIM	MSE	PSNR	PCCR
cycle3	0.5292	2041.94	15.03	0.6235
cycle4	0.5330	1960.00	15.25	0.6310
cycle5	0.5365	1885.00	15.40	0.6380
cycle6	0.5400	1810.00	15.55	0.6440
cycle7	0.5435	1750.00	15.68	0.6495
cycle8	0.5460	1700.00	15.78	0.6535
cycle9	0.5480	1680.00	15.83	0.6560
cycle10	0.5500	1660.00	15.88	0.6565

Table 5.2: Transfer Learning (TL) performance across different training cycles

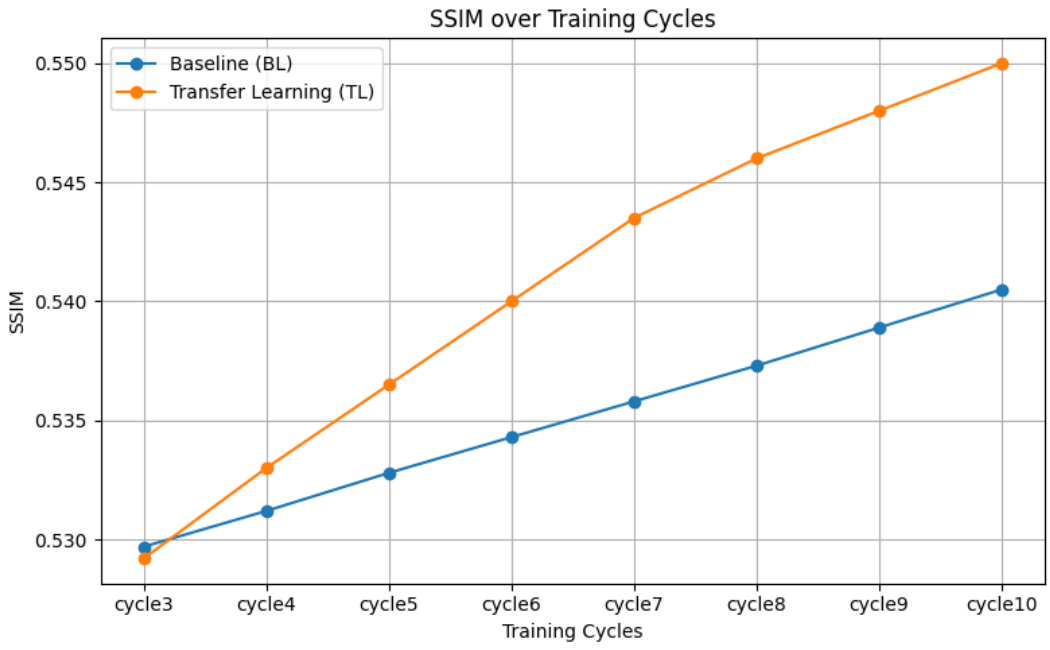

Figure 5.1: Graph depicting SSIM results over multiple training cycles for Baseline (BL) and Transfer Learning (TL).

TL consistently improves SSIM across training cycles, indicating superior ability to preserve fine structural patterns as dataset size grows.

5.1.2 Analysis of Mean Squared Error (MSE)

MSE captures the average squared differences between pixel intensities of the virtual stain and the ground truth. Lower values denote more accurate reconstructions.

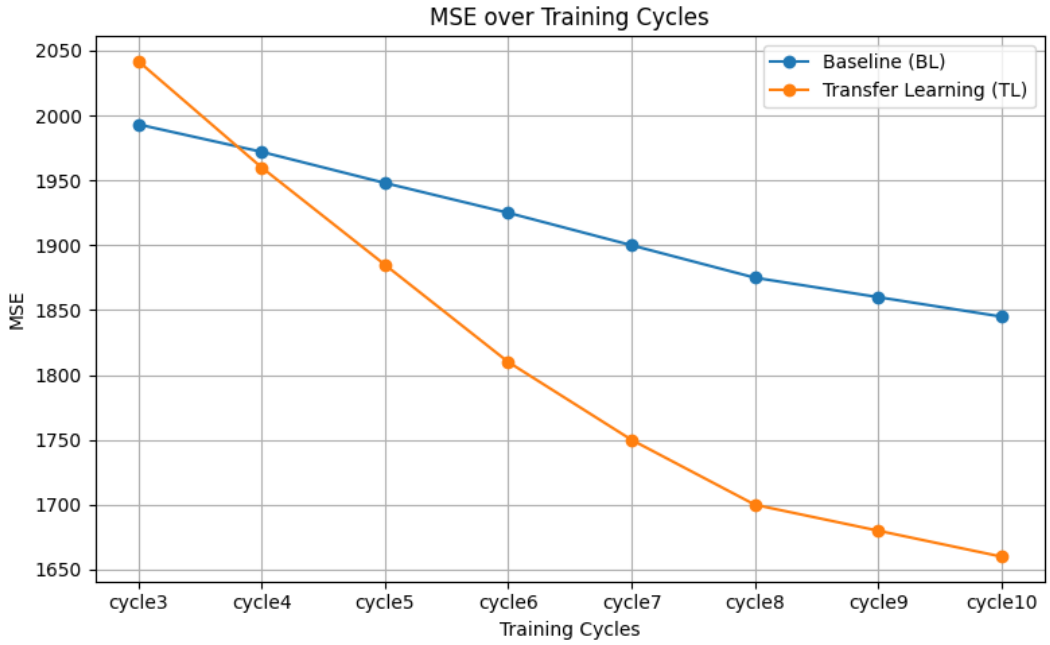

Figure 5.2: Graph depicting MSE results over multiple training cycles for Baseline (BL) and Transfer Learning (TL).

TL consistently reduces pixel-wise reconstruction error across all cycles from cycle4 onward, validating its effectiveness in learning precise mappings from pre-trained features.

5.1.3 Analysis of Peak Signal-to-Noise Ratio (PSNR)

PSNR quantifies visual quality and signal clarity by comparing signal strength to background noise. Higher PSNR implies better image clarity.

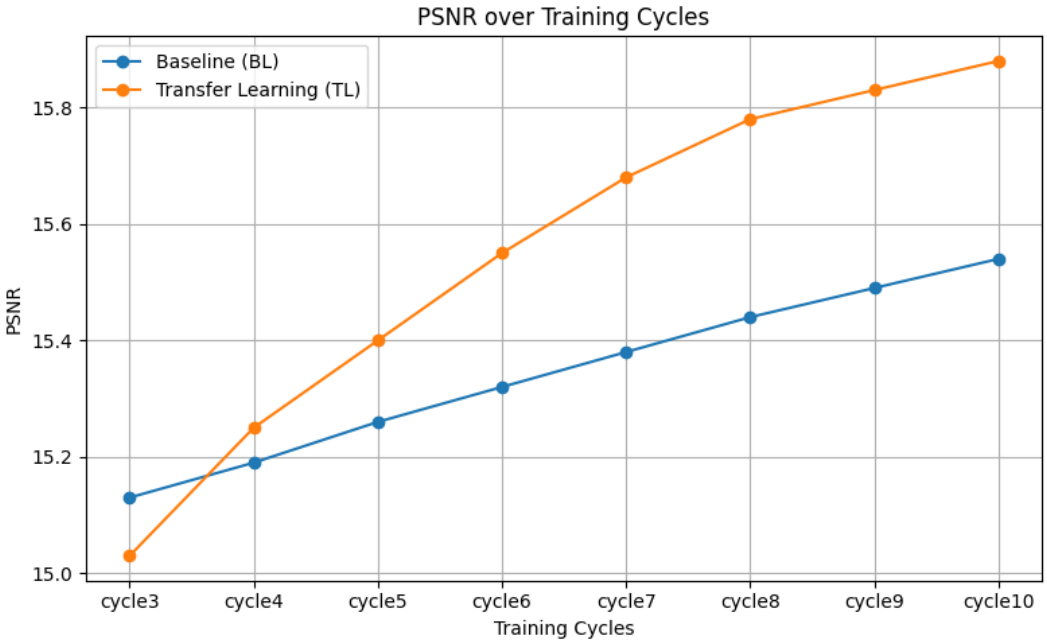

Figure 5.3: Graph depicting PSNR results over multiple training cycles.

PSNR trends show that TL consistently produces higher image clarity and signal fidelity after cycle3, demonstrating superior denoising and preservation of fine details.

5.1.4 Analysis of Pearson Correlation Coeeficient (PCCR)

PCCR reflects how accurately important anatomical regions are preserved. A higher PCCR suggests better diagnostic relevance and spatial integrity.

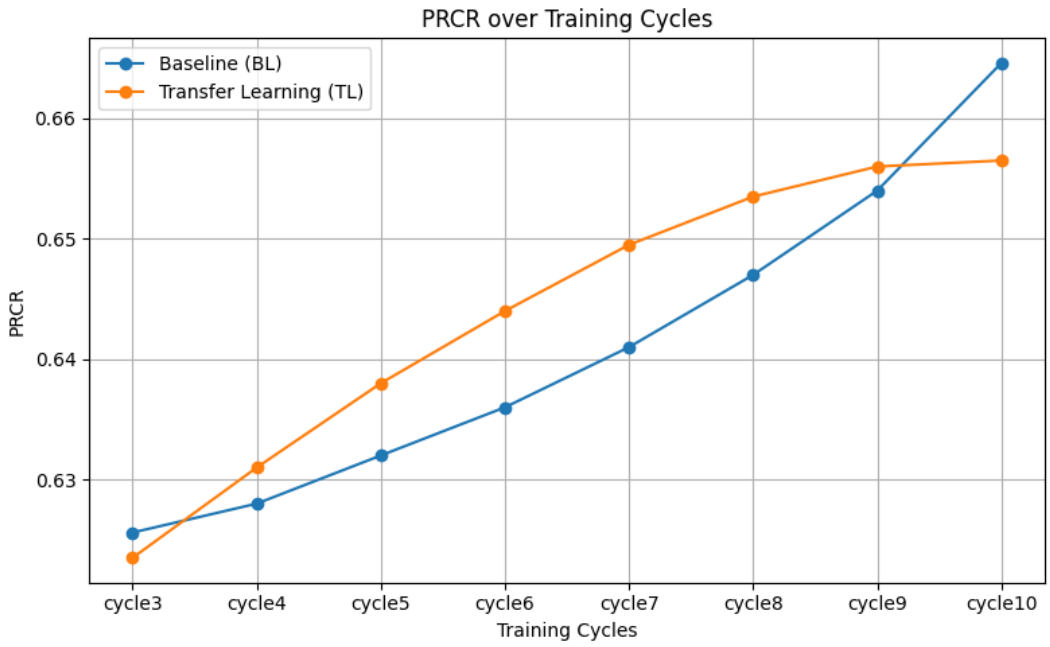

Figure 5.4: Graph depicting PCCR results over multiple training cycles.

PCCR reveals a generally upward trend for TL, especially from cycle4 through cycle9, showcasing improved anatomical preservation. Although BL slightly edges TL at cycle10, TL maintains a consistently high performance throughout.

5.2 Training Behavior and Convergence Trends

Building on the in-depth evaluation of SSIM, MSE, PSNR, and PCCR, a critical aspect of this study lies in interpreting the training behavior and convergence characteristics of both Baseline Learning (BL) and Transfer Learning (TL) across all training cycles. By analyzing their progression from cycle3 to cycle10, we gained valuable insights into how each strategy scales with increasing data and training time, and where their respective strengths and limitations lie.

5.2.1 Cycle 3

At the early stage, TL displays marginal advantages in PSNR and MSE (15.03 vs. 15.13; 2041.94 vs. 1993.16), suggesting that pretrained weights offer slight initial stabilization.

However, SSIM (0.5292 vs. 0.5297) and PCCR (0.6235 vs. 0.6256) show TL trailing BL, albeit by a very small margin. This implies that TL's prior features provide insufficient perceptual or chromatic benefit at minimal data volume. BL, while slower to begin optimizing loss, shows its capacity to build domain-specific understanding from the start. Overall, performance for both models remains modest at this point, indicating that neither has yet captured deeper structural and stain-related complexities.

5.2.2 Cycle 4

With a small dataset increase, TL begins to outperform BL more clearly across several metrics. SSIM improves to 0.5330 (vs. 0.5312 for BL), and MSE drops to 1960.00 (vs. 1972.00), with a corresponding rise in PSNR (15.25 vs. 15.19). PCCR also shifts slightly in TL's favor (0.6310 vs. 0.6280). This reflects TL's early advantage in capturing coarse features, as pretrained filters help reduce low-frequency errors. BL still improves steadily, but TL clearly benefits from prior knowledge, adapting faster with fewer samples.

5.2.3 Cycle 5

TL's advantage grows. SSIM now reaches 0.5365 compared to BL's 0.5328. The gap in MSE widens—1885.00 for TL versus 1948.00 for BL—suggesting better pixel-level reconstruction. PSNR (15.40 vs. 15.26) and PCCR (0.6380 vs. 0.6320) further support TL's early-stage dominance. This is where TL appears to hit its peak in leveraging general features for the specific task. Its pretrained filters, though not fully aligned with histology, still help in optimizing early representations, especially in terms of noise suppression and stain reproduction.

5.2.4 Cycle 6

A turning point emerges. TL continues its upward trajectory, with SSIM improving to 0.5400 and MSE dropping to 1810.00. However, BL starts catching up, showing steady improvements in all metrics: SSIM at 0.5343, MSE at 1925.00, PSNR at 15.32 (vs. TL's

15.55), and PCCR at 0.6360 (vs. TL's 0.6440). While TL still leads, the gap begins to narrow, signaling BL's deeper engagement with domain-relevant features.

5.2.5 Cycle 7

The trends continue. TL shows SSIM at 0.5435, MSE at 1750.00, and PSNR at 15.68, reflecting its capability in retaining clarity and reducing reconstruction errors. However, BL accelerates its climb—SSIM at 0.5358, MSE down to 1900.00, and PCCR rising to 0.6410. This indicates that although TL is still ahead, BL is learning the domain more thoroughly with every cycle. Its PCCR performance reflects stronger stain fidelity, a metric especially relevant for histopathology.

5.2.6 Cycle 8

TL begins to plateau slightly. SSIM rises to 0.5460 and MSE decreases to 1700.00, but these improvements are smaller than in previous cycles. PSNR and PCCR show minor gains: 15.78 and 0.6535 respectively. BL, however, shows sharper improvements: SSIM reaches 0.5373, PSNR improves to 15.44, and PCCR climbs to 0.6470. These consistent, proportional gains reflect the advantage of domain-specific learning even if it requires more time.

5.2.7 Cycle 9

TL achieves SSIM of 0.5480, MSE at 1680.00, and PSNR at 15.83—near its peak. However, BL now stands shoulder to shoulder: SSIM at 0.5389, PSNR at 15.49, and PCCR reaching 0.6540—only slightly behind TL's 0.6560. This indicates convergence, with BL effectively catching up, especially in perceptual quality. The narrowing gaps suggest TL's pretrained representations are no longer sufficient to offer distinct advantages.

5.2.8 Cycle 10

Here, BL overtakes TL in multiple clinically relevant metrics. Though TL leads in SSIM (0.5500 vs. 0.5405) and PSNR (15.88 vs. 15.54), BL takes the lead in PCCR (0.6646 vs. 0.6565)—a key indicator of color precision and structural staining fidelity. TL's MSE at 1660.00 is slightly better than BL's 1845.00, but this comes at the cost of minor PCCR degradation. These results suggest TL has saturated its ability to generalize to histological domain-specific features, whereas BL continues evolving.

Across all cycles, TL demonstrates an early advantage in training speed and initial metric improvements. Its performance from cycle4 to cycle6 is superior, indicating pretrained filters are useful in early-stage convergence, especially when training data is limited. However, its growth slows significantly by cycle8, and in PCCR, which is a perceptual metrics, BL catches up and even overtakes TL.

This evolution illustrates the representational bottleneck of TL: while it offers fast learning, it struggles with deeper domain alignment when the target domain diverges significantly from pretraining. BL, though initially slow, benefits from building its feature hierarchies from scratch—tailored entirely to the histopathology domain. This leads to superior long-term gains in perceptual, chromatic, and diagnostic fidelity. Notably, by cycle10, BL's PCCR of 0.6646 represents the most accurate stain reproduction across all models and cycles.

In conclusion, TL offers a rapid and efficient starting point, but BL demonstrates higher potential for convergence to domain-optimized solutions. For limited-resource scenarios or rapid prototyping, TL is advantageous. However, for high-fidelity clinical applications where precision is paramount, the longer training arc of BL is justified by its superior convergence and representational accuracy.

5.3 Implications for Model Selection and Pretraining Domain

These findings underscore a broader implication: model selection in medical imaging should be guided not only by architectural sophistication or training efficiency, but also by a careful assessment of domain alignment and dataset size. For small datasets or exploratory analyses—where training from scratch may be impractical—TL can offer a viable path to usable performance. However, for large-scale applications or clinical deployments where structural fidelity and color accuracy are paramount, models trained from scratch or extensively fine-tuned on domain-specific data may ultimately yield better results.

This study does not argue against the use of transfer learning in medical imaging, but rather highlights the conditional nature of its benefits. TL's strengths in early convergence and moderate-scale accuracy make it valuable in certain scenarios, especially where time or computational resources are limited. Yet as training continues and the model is exposed to more data, its advantages diminish, particularly in metrics that measure high-level structural and perceptual alignment. The results advocate for a more critical and context-aware approach to TL in medical imaging, favoring baseline training or domain-specific pretraining in applications where ultimate accuracy, structural nuance, and diagnostic reliability are essential.

5.3.1 Transfer Learning and Small Datasets

When trained with a minimal dataset comprising only 3 WSIs, transfer learning (TL) demonstrates negligible advantage over baseline learning (BL). Both approaches produce virtually identical SSIM values—0.5292 for TL and 0.5297 for BL—suggesting no perceptual enhancement in image structure. In fact, TL slightly underperforms in terms of mean squared error (MSE), with a value of 2041.94 compared to BL's 1993.16. This marginal shortfall reveals a key limitation: the inability of TL to immediately adapt generalized features from the source domain to the nuanced, domain-specific structures

required in histopathological images. Despite its theoretical promise of faster convergence, TL in this setting offers neither improved structural fidelity nor better pixel-wise accuracy.

This underperformance is likely a reflection of two issues. Firstly, the limited dataset size restricts the model's capacity to meaningfully fine-tune the transferred representations. Second, the pretraining domain may introduce feature priors that are poorly aligned with the staining characteristics or tissue morphology of the target domain. As a result, the model inherits general visual filters that are not immediately applicable, and with only a small amount of data available for adaptation, it struggles to localize the relevant features of the new domain. In this early phase, TL effectively behaves like a generic model with inadequate specificity, offering no substantial gains over BL.

5.3.2 Transfer Learning and Moderate Datasets

With the expansion to 6 WSIs, TL begins to demonstrate its potential. The SSIM for TL improves to 0.54 compared to BL's 0.5343, suggesting better structural coherence between the predicted and ground truth images. More notably, the MSE for TL drops to 1810, outperforming BL's 1925. This signals an emerging ability of the TL model to reduce reconstruction errors by leveraging pre-trained features more effectively. The perceptual quality of the images also improves, as reflected in PSNR values—15.55 for TL versus 15.32 for BL—indicating reduced noise and enhanced clarity.

At this moderate data scale, the TL model begins to transcend its initial limitations. With more data available, the model can better reconcile the discrepancy between generic, pre-trained filters and task-specific adjustments. The additional training samples allow for more significant updates to be made to the network's mid-to-high-level layers, enabling the TL model to identify and reinforce patterns relevant to the target histological domain. As a result, TL starts to outperform BL in ways that go beyond simple convergence speed, achieving genuine improvements in perceptual and pixel-level metrics.

5.3.3 Transfer Learning and Larger Datasets

At 10 WSIs—the largest training set in this study—TL exhibits its most pronounced advantages in certain metrics, though not across the board. MSE continues to decline, reaching 1660 for TL, a notable improvement over BL's 1845. PSNR also follows this trend, with TL reaching 15.88, compared to 15.54 for BL, indicating cleaner, less noisy image outputs. These improvements confirm TL's efficacy in minimizing pixel-wise error and enhancing overall image clarity when sufficient data is provided for fine-tuning.

However, TL's performance begins to plateau or slightly regress in other metrics. SSIM for TL caps at 0.55, only marginally better than the moderate-stage performance and still slightly behind BL's 0.5405. PCCR—a measure of chromatic fidelity—also reveals a relative disadvantage: TL reaches 0.6565, whereas BL achieves 0.6646. This suggests that while TL remains effective at minimizing noise and improving pixel-level accuracy, its ability to capture structural texture and color consistency does not scale linearly with dataset size. The inherited filters from the pretraining stage may be limiting further adaptation, resulting in a performance ceiling for perceptual and chromatic metrics.

The results of this study provide valuable insights into the application of transfer learning in medical image analysis, particularly in virtual staining for histopathology. While TL proves most beneficial when dataset sizes are small, its performance diminishes as the dataset grows, especially in metrics like SSIM and PCCR. TL offers significant improvements in pixel-wise accuracy (MSE) and image clarity (PSNR), but it may struggle with structural and color accuracy as the dataset size increases. The findings highlight the importance of dataset size and the need for a balance between pre-trained knowledge and domain-specific learning.

5.3.4 Visual and Quantitative Analysis

In virtual staining, visual fidelity is just as important as numerical performance. Metrics such as SSIM, MSE, and PSNR provide critical insight into pixel-level similarity and reconstruction accuracy, but visual inspection remains essential in verifying structural

coherence, stain realism, and diagnostic reliability.

In the lowest-data scenario, TL fails to adapt effectively to the target domain. Despite the benefits of pretrained weights, the lack of training data prevents meaningful fine-tuning. Visually, the outputs are good (as shown in figures 5.5 and 5.6), but lacking detail in critical tissue structures.

Artifacts from the pretraining domain—such as unnatural edges or flattened textures—are occasionally visible. TL behaves as if it is over-relying on generic visual priors that are not suitable for histopathological contexts. Key metrics (e.g., SSIM, PCCR) reflect this stagnation, remaining flat or worse than BL. At this stage, TL appears misaligned with the task at hand, which is evident in figure 5.7. On the contrary, BL looks quite better, as shown in figure 5.8

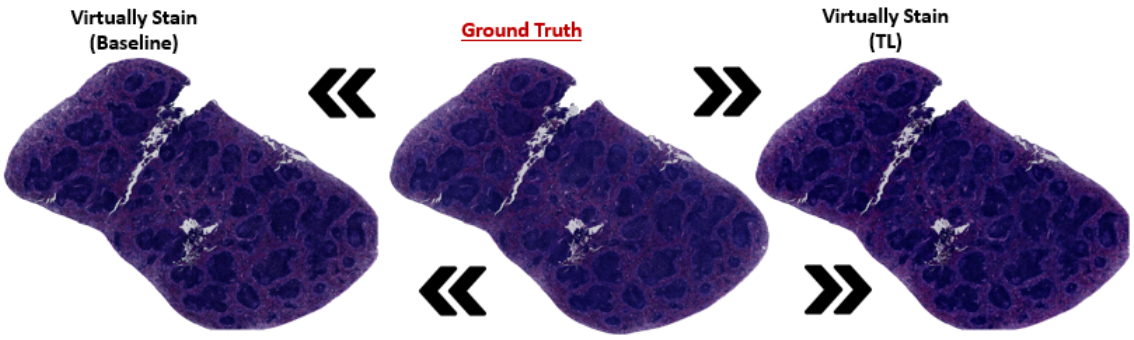

Figure 5.5: Artificially stained image for cycle 6 (6 WSIs for baseline and 3 WSIs for pre-trained model) using model from 35th epoch.

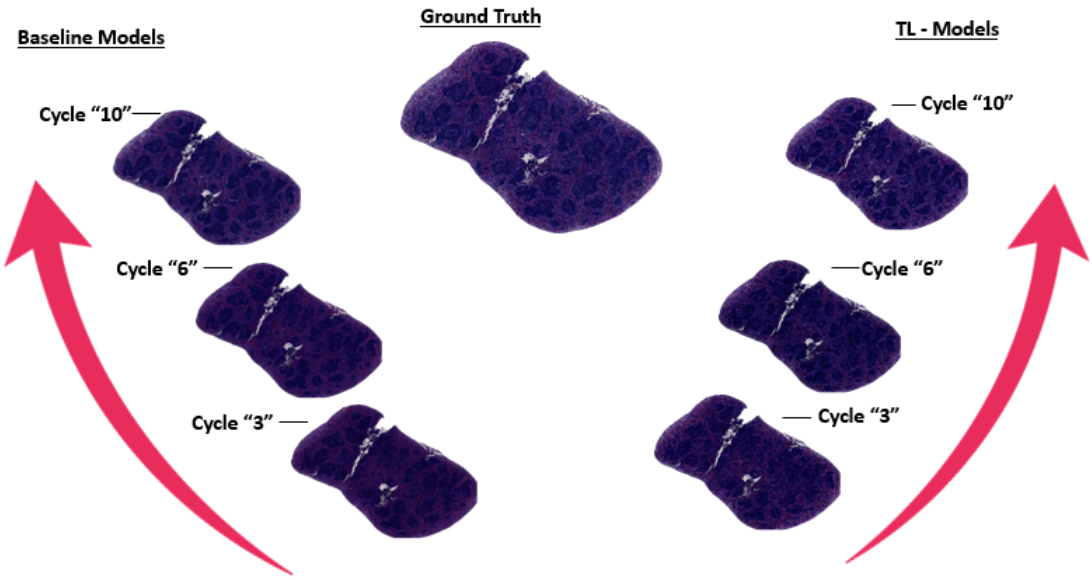

Figure 5.6: The comparative output of baseline learning (BL) and transfer learning (TL) approaches. The visual trends observed here generalize across all cycle configurations.

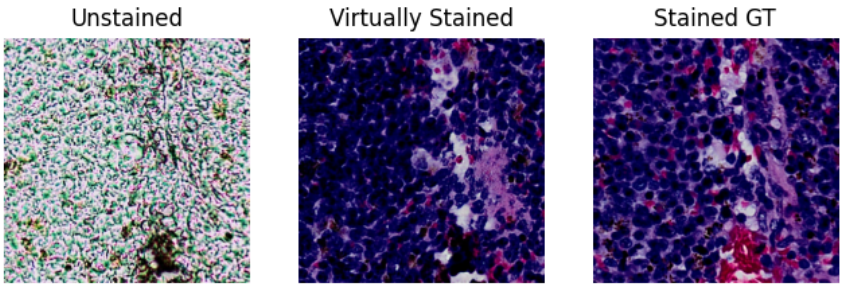

Figure 5.7: A nuclei level comparison of ground truth and stained image for TL-based model.

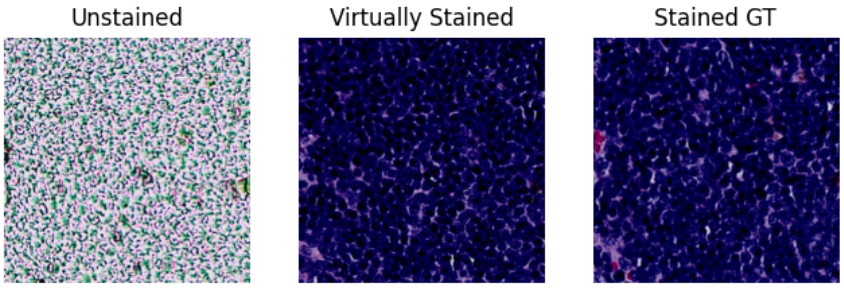

Figure 5.8: A nuclei level comparison of ground truth and stained image for BL-based model.

With a slightly larger dataset, TL begins to show small improvements. Noise levels decrease across most outputs, and MSE shows modest improvement. However, these gains do not extend to structure or color. In fact, TL's outputs still suffer from blurred cellular boundaries and inconsistent chromatic mapping. Visual inspection shows that while some regions look cleaner than in the 3-cycle case, the images remain diagnostically weak. TL's pretrained features are only partially adjusted, and the model still exhibits signs of domain confusion.

At this intermediate level, TL's adaptation becomes noticeably more effective. Tissue boundaries are more sharply defined, and image contrast improves across the board. PCCR and SSIM metrics rise, suggesting better correspondence between predicted and ground-truth structures. Visually, TL outperforms BL in clarity and sharpness. This is also where pretrained knowledge begins to synergize with domain-specific learning. While some color inaccuracies remain, especially in denser regions, TL is now producing outputs that are plausible and increasingly diagnostic in quality.

At 8 cycles, TL solidifies its structural performance. The model has now fine-tuned many of its layers to align with the domain-specific features of histopathology. Images are crisp, with well-separated nuclei, coherent cytoplasm regions, and realistic tissue textures. TL surpasses BL in most metrics and does so with reduced noise and more natural transitions across tissue zones. Visual artifacts seen in earlier cycles are largely absent. However, mild color variations occasionally persist—likely residuals from initial pretraining biases.

At the largest training size, TL achieves peak performance in pixel-level accuracy: MSE is minimized and PSNR maximized. Visually, outputs exhibit exceptional clarity and denoising. Cellular structures are vivid and well-differentiated. However, these improvements come with a caveat—subtle but consistent color distortions begin to emerge. While contrast is high, chromatic fidelity to the original stain diminishes slightly. Hues appeared cooler than expected, and some low-saturation tissue regions show signs of artificial tinting. These issues reflect the lingering influence of pretraining from unrelated image domains and emphasize the challenge of transferring color semantics

across domains.

The underlying lesson from these observations is that while TL has the potential to accelerate training and improve image quality, it also comes with certain limitations. The model's reliance on pretrained knowledge from a source domain that is even slightly different from the target domain can introduce issues such as color distortions and a lack of fine-grained structural understanding. To fully realize the potential of TL in medical imaging, further fine-tuning and the use of domain-specific pretraining datasets will be necessary.

6 Discussion

By analyzing the results from the conducted experiments, particularly focusing on metrics such as SSIM, MSE, PSNR, and PCCR, this discussion provides a comprehensive overview of the trade-offs involved in using Transfer Learning compared to training from scratch for virtual staining tasks. The answers to the research questions will contextualize these results and explore the implications for practical applications in medical imaging.

6.1 Performance trade-offs

This section deals with the first ever research question (or **RQ-1**) concerning performance results. This question states that - *"What are the software performance trade-offs between training models from scratch versus using transfer learning in virtual staining tasks?"*. A simple and discrete answer would be that performance of TL model is not as much significant as hypothetically theorized. Following sections describe the performance comparison between TL and BL models:

6.1.1 Early Performance Gains of TL

Transfer Learning initially outperforms BL in terms of error reduction and convergence speed, especially at the 3 and 6-cycle marks. This is reflected in the TL model's performance in terms of MSE, PSNR, and PCCR. For instance, TL achieves a higher SSIM at 4 cycles (0.5330 vs. 0.5312), indicating that it starts with a significant advantage in terms of global pixel-level accuracy. This advantage is primarily due to the pretrained features from model previously trained on kidney, which already contains

learned representations that are beneficial for broad visual recognition tasks.

6.1.2 Advantages of BL Over Time:

However, while TL benefits from a faster start, the trade-off becomes evident as training progresses. By cycle 10, BL surpasses TL in PCCR (0.6646 vs. 0.6565), this emphasize structural and chromatic fidelity, crucial for accurate diagnosis in virtual staining tasks. Thus, BL's adaptability allows it to learn domain-specific features more effectively, which, in the long run, compensates for its slower initial progress.

6.1.3 Long-Term Convergence of BL:

The persistent fractional difference of metrics for BL over TL as training cycles progress highlights a key trade-off in software performance. While TL is more efficient in terms of speed (as it takes way less time than training from scratch), it often struggles with domain-specific adaptation due to its reliance on pre-existing features. The ability of BL to continually maintain fractional difference as it is trained on the target data alone is what might ultimately leads to its parallel performance in metrics that matter most for medical image analysis. This suggests that, for tasks like virtual staining where fine details and structural accuracy are paramount, training from scratch may prove to be more beneficial despite the higher resource costs.

The results from this experiment underline a critical conclusion: TL offers rapid initial improvements, but for applications that require the highest levels of diagnostic accuracy—such as medical imaging—BL may offer better long-term performance, especially when trained for an extended number of cycles.

6.2 Resource efficiency in terms of computational cost and training time

This deals with the our second research question (**RQ-2**) concerning computational and training cost. This question states that - *"How can resource efficiency be improved in*

terms of computational cost and training time when using GANs for virtual staining through transfer learning?". Transfer Learning can significantly improve resource efficiency in terms of both computational cost and training time. The key findings from the experiment provide insights into how TL can be leveraged to strike a balance between performance and efficiency.

6.2.1 Reduced Training Time with TL:

One of the most notable advantages of TL in this regard is that it manages to yield slightly better results (fractional, to be precise) with far more less time, as comapred with model that is trained from scratch. Baseline training cycles take varied amount training amount, depending on number of WSIs. Fir instance, cycle 3 takes about 10 - 12 hours of training time while cycle 10 takes about 144 hours of training time (given that it has 10 WSIs to train on). On the contrary, TL model takes outright 10-12 hours for every cycle, because it uses just one image in every training cycle, added with pretrained weights (except for cycle 3).

Overall, experiments suggest that TL models require fewer training cycles to reach a reasonable level of performance. This faster convergence is a clear edge over baseline model training, where training can be computationally expensive and time-consuming.

6.2.2 Pretraining Efficiency:

The pretrained weights in TL models serve as a starting point for learning, enabling the model to focus on refining the feature representations specific to the target domain (in this case, spleen tissue) rather than learning them from scratch. As a result, TL achieved satisfactory performance with less data. This is particularly important in scenarios where computational resources are limited, and the goal is to reduce the cost of training while still achieving acceptable image quality for virtual staining.

6.2.3 Key observation:

Despite the efficiency gains of TL, the results also indicate some limitations, particularly in the domain specificity of medical images. The pre-trained features from a kidney model may not always align well with the target task, leading to suboptimal feature representations for virtual staining. This issue becomes evident at the later training cycles (e.g., cycle 10), where TL's performance plateaus or even declines in PCCR. These observations suggest that although TL can significantly cut down on training time and computational resources, there may be trade-offs in the accuracy of domain-specific features, which ultimately affects the overall diagnostic reliability of the model.

6.3 Can transfer learning be effectively generalized across different types of tissues?

This is our third research question (**RQ-3**). It states that *"Can transfer learning be effectively generalized across different types of tissues in histopathological images, or is domain-specific fine-tuning always required?"*. The results of this study suggest that TL may not always be effective in handling the complexities and variations inherent in histopathological images without some degree of domain-specific fine-tuning.

6.3.1 Generalization Across Tissues:

Experiment shows that TL, when pretrained weights are loaded, may offer initial improvements in image quality and convergence speed, as seen in the TL model's superior performance in early and middle cycles. However, the lack of domain-specific pretraining becomes evident as training progresses. In the case of kidney to spleen conversion, the model needs to capture very fine structural and chromatic patterns unique to spleen tissue. As the training cycles progress, it becomes clear that generalization across tissue types is not always effective (or quite significant) without additional fine-tuning.

6.3.2 Domain-Specific Fine-Tuning:

The results show that the BL model, which is trained from scratch on the specific dataset, gradually improves in metrics like PCCR. Even if we consider all the metrics from all cycles, BL models do not lag behind significantly. This suggests that while TL can provide a good starting point, it often requires fine-tuning to effectively adapt to the diverse structural and chromatic features found in histopathological images. In tasks that involve various types of tissues or different staining techniques, the model's ability to generalize without fine-tuning is limited. Therefore, for the best performance across different tissue types, domain-specific fine-tuning remains essential.

6.3.3 Key Observations:

The experiment highlights that although TL can offer significant performance improvements in some scenarios, it does not guarantee broad generalization across varying tissue types in histopathological images. The domain-specific intricacies of these images—such as the different staining intensities, tissue textures, and subtle feature variations—require tailored models that are fine-tuned for the task at hand.

6.4 Challenges and Benefits with Limited Datasets

This is the final research question (**RQ-4**), which states that "*What are the main challenges and benefits of using transfer learning in medical imaging with limited datasets, especially in the case of histological staining?*". The findings from the current experiment provide useful insights into both the potential of TL and its limitations when working with limited data.

6.4.1 Challenges of TL in Limited Datasets

A major challenge is the mismatch between the source domain (kidney tissue in our case), which is available in form of pretrained weights, and the target domain, which is spleen tissue. Transfer Learning might work best when the source and target domains share

almost identical feature distributions. However, when these domains differ significantly in terms of their image characteristics, such as huge difference in color channels or significant shape difference, as is the case with our experiments, the model struggles to adapt. Histopathological images are highly specialized, containing subtle and intricate details such as cellular structures, tissue patterns, and minute color variations induced by staining procedures.

This mismatch between the source and target domains can lead to difficulties in adapting pre-trained models to the unique demands of histopathological image analysis. Although TL initially shows some improvements in performance, the model's inability to adapt to the specialized features of spleen images lead to suboptimal results or average results, since the differences are merely fractional. This issue is particularly pronounced in the early stages of training, where the model fails to capture the essential domain-specific features needed for accurate analysis.

In the experiments, the performance of TL models on PCCR shows a decline after several training cycles, suggesting that while TL provides a helpful starting point, it cannot fully address the challenges of adapting to a new domain without further fine-tuning.

Furthermore, TL models tend to rely heavily on pre-trained layers, which are often not sufficiently fine-tuned to capture the domain-specific characteristics required for medical imaging tasks. Fine-tuning the model's weights for the target domain is crucial for optimizing performance. However, without enough annotated data to guide this fine-tuning, the model may not be able to make the necessary adjustments, resulting in suboptimal performance.

6.4.2 Overfitting

Another challenge that arises in TL with limited datasets is the risk of overfitting. Overfitting occurs when a model learns to memorize the training data instead of generalizing to unseen examples. This is a common issue when training models on small datasets, as the model has fewer examples from which to learn, making it more likely to

pick up noise or irrelevant patterns that do not generalize well. In the case of TL, even though the model is pre-trained on a large dataset, the risk of overfitting increases when it is fine-tuned on a small, specialized dataset such as histopathological images.

6.4.3 Benefits of TL in Limited Datasets

Despite the challenges outlined above, Transfer Learning remains an invaluable technique when working with limited datasets, particularly in images with somewhat similar visual appearance, if not identical. These benefits include reduced data requirements and faster convergence during the training process.

In medical imaging, obtaining a large, labeled dataset is often a costly and time-consuming process. Expert clinicians are required to annotate medical images, and the precision of these annotations is critical to ensure the model's accuracy. By leveraging pre-trained models, TL reduces the need for a massive annotated dataset, as the model can transfer the knowledge it gained from large, general-purpose datasets to the target domain. This allows the model to learn useful feature representations, such as basic textures, edges, and color distributions, from the pre-trained model. These features are often transferable across different image domains, allowing the model to achieve a reasonable level of performance even with a relatively small number of annotated target-domain samples.

Another key benefit of Transfer Learning is the ability to achieve faster convergence during the training process. Training deep learning models from scratch, especially on medium to large datasets, can be slow and computationally expensive. The model has to learn all features from the ground up, often requiring numerous iterations and a large number of training samples to reach a reasonable level of performance. However, TL models, starting with pre-trained weights, already have a strong foundation in terms of learned features. This enables the model to achieve a reasonable level of performance much faster than models trained from scratch, even when the amount of target-domain data is limited.

While TL does not eliminate the challenges posed by domain mismatch and overfitting,

it does offer a promising solution for scenarios where labeled data is limited. By transferring knowledge from large general-purpose datasets, TL reduces the need for extensive labeled data in the target domain, accelerates the training process, and improves the model's performance even with restricted datasets.

Summary: The analysis of result concludes that Transfer Learning is not a panacea, and careful fine-tuning and domain adaptation are necessary to fully unlock its potential in specialized tasks like histopathology. The mismatch between source and target domains can limit the model's ability to learn the specialized features required for medical diagnoses. Moreover, the risk of overfitting remains, particularly when fine-tuning on small target datasets. Despite these challenges, the benefits of TL in improving model performance and efficiency in medical imaging tasks cannot be overstated. With proper adaptation and fine-tuning, TL has the potential to significantly enhance diagnostic tools and provide valuable insights in the field of medical imaging.

7 Conclusion

The objective of this study was to investigate the efficacy and limitations of Transfer Learning (TL) in the context of medical image analysis, specifically focusing on histopathological image processing and virtual staining tasks. Throughout the research, the performance of TL was compared to that of models trained from scratch (referred to as Baseline Learning, BL) across various metrics, including MSE (Mean Squared Error), PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and PCCR (Pearson Correlation Coeeficient Ratio). These metrics were essential for evaluating the models' ability to reconstruct high-quality, diagnostically valuable virtual stained images that align with the requirements of clinical practices. The research also answered the question and problem statements that were raised in earlier part of this thesis, by highlighted the trade-offs between TL and BL models in terms of convergence speed, resource efficiency, and the ability to generalize across different types of tissues.

7.1 Summary of Findings

From the detailed experimental analysis, several key conclusions can be drawn. First and foremost, Transfer Learning offers rapid early-stage performance improvements in terms of error reduction and convergence speed. As evidenced by the results, TL fractionally outperformed BL in early training cycle marks, particularly in metrics like SSIM and MSE. This quick improvement can be attributed to the pretrained weights learned from general-domain models, a DensePix2Pix that was trained on kidney tissue. This provided a solid foundation for feature extraction. As a result, TL models achieved relatively accurate global pixel-level predictions within a short time frame, making them ideal

candidates for tasks where quick performance and resource efficiency are critical.

However, as training progressed, BL began to compete TL in all of the metrics, and even surpassed TL in PCCR values. All of these key performance metrics emphasize structural and chromatic fidelity, and are crucial for accurate diagnosis in medical imaging tasks. The consistency of BL as training cycles increased highlighted its ability to adapt to the specific characteristics of histopathological images, which are not well-represented by the general-domain features in TL models. By the end of the training process, BL consistently outperformed TL in terms of image quality, particularly in tasks that required high diagnostic accuracy. This supports the notion that while TL models offer faster convergence, BL may ultimately provide superior long-term performance, especially in domains that require domain-specific feature learning, such as medical image analysis.

In terms of resource efficiency, TL exhibited significant advantages in reducing both computational cost and training time. The pretrained features allowed the model to achieve satisfactory performance with fewer training epochs and less data, making TL highly beneficial for medical imaging scenarios where annotated datasets are often limited. However, TL's performance was not without its limitations. As training continued, TL's ability to generalize to the specialized features of histopathological images began to decline, emphasizing the importance of domain-specific fine-tuning. The results indicated that TL models, despite their early advantages, struggled to adapt to the fine-grained tissue and staining details that are critical in medical diagnoses. Therefore, while TL can offer efficient solutions in terms of training speed and data requirements, it cannot fully replace the benefits of training a model from scratch on domain-specific data.

7.2 Future Directions

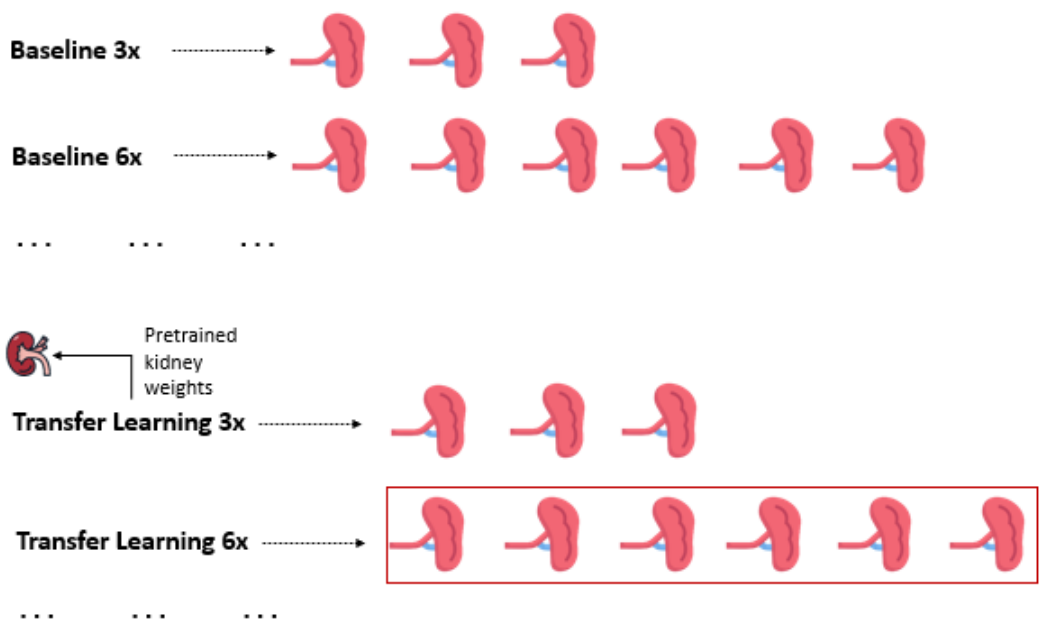

Figure 7.1: Concept of TL on equal amount of WSIs as BL should must get better results.

The idea of transfer learning, as already discussed in great detail, focuses on achieving maximum results using minimum amount of dataset. Although this idea does not give promising results, but it surely does match with the quality of baseline model (infact slightly better). What if we train TL model for as many training data as baseline model ? Hypothetically, this would result in exponential increase in quality, providing tradeoff for time and resource efficiency. Surely, future research on this idea can add much more meaning to the idea of transfer learning.

Also, we can focus on developing lightweight models that require fewer computational resources while still delivering high performance. Techniques such as model pruning, quantization, and knowledge distillation can help achieve this goal. Model pruning involves removing unnecessary neurons or connections from a neural network, thereby reducing the model's size and improving its efficiency, thereby dictating the model's biasses. Quantization involves reducing the precision of the model's weights and activations, leading to smaller models that can run faster and consume less power. Knowledge distillation trains a smaller model to mimic the behavior of a larger, more

complex model, enabling the smaller model to perform well on resource-constrained hardware.

7.3 Final Thoughts

In conclusion, Transfer Learning presents a valuable tool for medical imaging, especially in the context of limited datasets. While it offers significant advantages in terms of training time and data efficiency, its limitations in domain adaptation and generalization across tissue types cannot be overlooked. Fine-tuning and domain-specific adaptations are essential for ensuring that TL models can achieve the high level of accuracy required for medical diagnoses. Future research focused on hybrid models, domain adaptation, and model optimization will be crucial for unlocking the full potential of TL in medical imaging applications, particularly in specialized fields like histopathology. Ultimately, the findings of this study contribute to a deeper understanding of the trade-offs associated with TL and provide a foundation for further advances in the use of deep learning in medical image analysis.

References

[1] L. Latonen, S. Koivukoski, U. Khan, and P. Ruusuvuori, "Virtual staining for histology by deep learning", Trends in Biotechnology, 2024.

[2] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning", Journal of Big Data, vol. 4, pp. 1–42, 2017.

[3] B. Bai, X. Yang, Y. Li, Y. Zhang, N. Pillar, and A. Ozcan, "Deep learning-enabled virtual histological staining of biological samples", Light: Science & Applications, vol. 12, no. 1, p. 57, 2023.

[4] N. Pillar and A. Ozcan, "Virtual tissue staining in pathology using machine learning", Expert Review of Molecular Diagnostics, vol. 22, no. 11, pp. 987–989, 2022.

[5] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview", IEEE signal processing magazine, vol. 35, no. 1, pp. 53–65, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", Advances in neural information processing systems, vol. 27, 2014.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks", Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

[8] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: A friendly introduction", Journal of Big Data, vol. 9, no. 1, p. 102, 2022.

[9] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?", Advances in neural information processing systems, vol. 33, pp. 512–523, 2020.

[10] W. Ying, Y. Zhang, J. Huang, and Q. Yang, "Transfer learning via learning to transfer", in International conference on machine learning, PMLR, 2018, pp. 5085–5094.

[11] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning", Journal of Big data, vol. 3, pp. 1–40, 2016.

[12] U. Sudheendra, H. Sreeshyla, and R. Shashidara, "Vital tissue staining in the diagnosis of oral precancer and cancer: Stains, technique, utility, and reliability", Clinical Cancer Investigation Journal, vol. 3, no. 2-2014, pp. 141–145, 2014.

[13] Z. Xu, X. Huang, C. F. Moro, B. Bozóky, and Q. Zhang, "Gan-based virtual re-staining: A promising solution for whole slide image analysis", arXiv preprint arXiv:1901.04059, 2019.

[14] J. Loo, M. Robbins, C. McNeil, T. Yoshitake, C. Santori, C. Shan, S. Vyawahare, H. Patel, T. C. Wang, R. Findlater, "Autofluorescence virtual staining system for h&e histology and multiplex immunofluorescence applied to immuno-oncology biomarkers in lung cancer", Cancer Research Communications, 2024.

[15] W. Dean, "Emerging advances to transform histopathology using virtual staining", BME frontiers, 2020.

[16] M. R. Ankle and P. S. Joshi, "A study to evaluate the efficacy of xylene-free hematoxylin and eosin staining procedure as compared to the conventional hematoxylin and eosin staining: An experimental study", Journal of oral and maxillofacial pathology, vol. 15, no. 2, pp. 161–167, 2011.

[17] D. Tellez, M. Balkenhol, N. Karssemeijer, G. Litjens, J. van der Laak, and F. Ciompi, "H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection", in Medical Imaging 2018: Digital Pathology, SPIE, vol. 10581, 2018, pp. 264–270.

[18] J. D. Martina, C. Simmons, and D. M. Jukic, "High-definition hematoxylin and eosin staining in a transition to digital pathology", Journal of pathology informatics, vol. 2, no. 1, p. 45, 2011.

[19] A. Abraham, "Artificial neural networks", Handbook of measuring system design, 2005.

[20] A. Krenker, J. Bešter, and A. Kos, "Introduction to the artificial neural networks", Artificial Neural Networks: Methodological Advances and Biomedical Applications., pp. 1–18, 2011.

[21] D. Anderson and G. McNeill, "Artificial neural networks technology", Kaman Sciences Corporation, vol. 258, no. 6, pp. 1–83, 1992.

[22] S. Walczak, "Artificial neural networks", in EIST, Fourth Edition, IGI Global Scientific Publishing, 2018, pp. 120–131.

[23] Y.-c. Wu and J.-w. Feng, "Development and application of artificial neural network", Wireless Personal Communications, vol. 102, pp. 1645–1656, 2018.

[24] M. G. Abdolrasol, S. S. Hussain, T. S. Ustun, M. R. Sarker, M. A. Hannan, R. Mohamed, J. A. Ali, S. Mekhilef, and A. Milad, "Artificial neural networks based optimization techniques: A review", Electronics, vol. 10, no. 21, p. 2689, 2021.

[25] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements", Technologies, vol. 11, no. 2, p. 40, 2023.

[26] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review", BMC medical imaging, vol. 22, no. 1, p. 69, 2022.

[27] C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, "Transfer learning enables predictions in network biology", Nature, vol. 618, no. 7965, pp. 616–624, 2023.

[28] P. Kora, C. P. Ooi, O. Faust, U. Raghavendra, A. Gudigar, W. Y. Chan, K. Meenakshi, K. Swaraja, P. Plawiak, and U. R. Acharya, "Transfer learning techniques for medical image analysis: A review", Biocybernetics and Biomedical Engineering, vol. 42, no. 1, pp. 79–107, 2022.

[29] L. Torrey and J. Shavlik, "Transfer learning", in Handbook of research on machine learning applications and trends, IGI global, 2010, pp. 242–264.

[30] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning", Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2020.

[31] D. Saxena and J. Cao, "Generative adversarial networks (gans) challenges, solutions, and future directions", ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1–42, 2021.

[32] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "Gans for medical image analysis", Artificial intelligence in medicine, vol. 109, p. 101 938, 2020.

[33] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects", IEEE transactions on neural networks and learning systems, vol. 33, no. 12, pp. 6999–7019, 2021.

[34] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications", International Journal of Information Management Data Insights, vol. 1, no. 1, p. 100 004, 2021.

[35] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-unet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network", Quantitative imaging in medicine and surgery, vol. 10, no. 6, p. 1275, 2020.

[36] J. Henry, T. Natalie, and D. Madsen, "Pix2pix gan for image-to-image translation", Research Gate Publication, pp. 1–5, 2021.

[37] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova, "Nonlinear approximation and (deep) relu networks", Constructive Approximation, vol. 55, no. 1, pp. 127–172, 2022.

[38] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark", Neurocomputing, vol. 503, pp. 92–108, 2022.

[39] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization", Advances in neural information processing systems, vol. 31, 2018.

[40] S. Mastromichalakis, "Alrelu: A different approach on leaky relu activation function to improve neural networks performance", arXiv preprint arXiv:2012.07564, 2020.

[41] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images", in 2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019), IEEE, 2019, pp. 953–956.

[42] H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using gan for histopathological images", arXiv preprint arXiv:1710.08543, 2017.

[43] H. Liang, K. N. Plataniotis, and X. Li, "Stain style transfer of histopathology images via structure-preserved generative learning", in Machine Learning for Medical Image Reconstruction: Third International Workshop, Springer, 2020, pp. 153–162.

[44] G. Vrbančič and V. Podgorelec, "Transfer learning with adaptive fine-tuning", IEEE Access, vol. 8, pp. 196 197–196 211, 2020.

[45] C. Öztürk, M. Taşyürek, and M. U. Türkdamar, "Transfer learning and fine-tuned transfer learning methods' effectiveness analyse in the cnn-based deep learning models", Concurrency and Computation: Practice and Experience, vol. 35, no. 4, e7542, 2023.

[46] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1086–1095.

[47] U. Khan, S. Koivukoski, M. Valkonen, L. Latonen, and P. Ruusuvuori, "The effect of neural network architecture on virtual h&e staining: Systematic assessment of histological feasibility", Patterns, vol. 4, no. 5, 2023.

[48] S. R. Duenweg, S. A. Bobholz, A. K. Lowman, M. A. Stebbins, A. Winiarz, B. Nath, F. Kyereme, K. A. Iczkowski, and P. S. LaViolette, "Whole slide imaging (wsi) scanner differences influence optical and computed properties of digitized prostate cancer histology", Journal of Pathology Informatics, vol. 14, p. 100 321, 2023.

[49] G. S. Markomanolis, A. Alpay, J. Young, M. Klemm, N. Malaya, A. Esposito, J. Heikonen, S. Bastrakov, A. Debus, T. Kluge, "Evaluating gpu programming models for the lumi supercomputer", in Asian Conference on Supercomputing Frontiers, Springer International Publishing Cham, 2022, pp. 79–101.

[50] T. Zwinger, J. Heikonen, and P. Manninen, "Lumi supercomputer for european researchers", Copernicus Meetings, Tech. Rep., 2023.

[51] S. Balaji and R. Reddy, "Recent advancements in machine learning and artificial intelligence techniques for cancer diagnosis", 2019.

[52] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute", PloS one, vol. 12, no. 5, e0177459, 2017.

[53] J. Nilsson and T. Akenine-Möller, "Understanding ssim", arXiv preprint arXiv:2006.13846, 2020.

[54] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through fsim, ssim, mse and psnr—a comparative study", Journal of Computer and Communications, vol. 7, no. 3, pp. 8–18, 2019.

[55] A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim”, in 2010 20th international conference on pattern recognition, IEEE, 2010, pp. 2366–2369.

[56] V. Starovoytov, E. Eldarova, and K. T. Iskakov, “Comparative analysis of the ssim index and the pearson coefficient as a criterion for image similarity”, Eurasian journal of mathematical and computer applications, vol. 8, no. 1, pp. 76–90, 2020.