Karlsruhe Institute of Technology

Image-to-Image Translation for Medical Microscopy with Deep Neural Networks

Master's Thesis of

Philipp Marquardt

At the faculty of Computer Science
Institute for Anthropomatics and Robotics

Reviewer:	Prof. Dr.-Ing. Rainer Stiefelhagen
Second reviewer:	Prof. Dr.-Ing. Thomas Längle
Advisor:	Dr.-Ing. Simon Reiß

Duration: 15. April 2024 – 03. November 2024

Computer Vision for Human-Computer Interaction Research Group
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Title: Image-to-Image Translation for Medical Microscopy with Deep Neural Networks
Author: Philipp Marquardt

Philipp Marquardt

p.marquardt15@googlemail.com

Statement of Authorship

I hereby declare that this thesis is my own original work which I created without illegitimate help by others, that I have not used any other sources or resources than the ones indicated and that due acknowledgement is given where reference is made to the work of others.

, 03. November 2024



.....

(Philipp Marquardt)

Abstract

Current techniques for image-to-image translation in microscopy are largely based on generative adversarial networks. While diffusion-based models have recently gained traction, their application in histological image translation remains limited. This thesis seeks to address this gap by comparing state-of-the-art methods in medical microscopy image translation with latent diffusion approaches. A core contribution of this work is a novel latent refinement network based on the Swin Transformer architecture. In this approach, the conditioning image is first encoded using a variational autoencoder and then refined within the latent space using the Swin Transformer. This method highlights the critical role of a well-pretrained variational autoencoder in achieving effective and high-quality image translation. This method opens up future research possibilities, including few-shot image-to-image translation and tailored pretraining for medical imaging tasks. In addition, this thesis presents an in-depth overview of both recent and foundational image-to-image translation methods, examining their strengths and limitations.

To evaluate the proposed approach, tests were conducted on four distinct datasets, each presenting its own technical challenges. On datasets with precise annotations and limited samples, the model achieved a Structural Similarity Index Measure (SSIM) of 0.74, closely matching the state-of-the-art SSIM of 0.752 on a Brightfield Image-to-Image translation dataset. Additionally, the model demonstrated superior performance on larger datasets compared to other leading algorithms. It reached an SSIM of 0.613 on the Unstained to Stained dataset, surpassing Pix2Pix at 0.512. For the Breast Cancer Immunohistochemical Dataset, it achieved an SSIM of 0.496, compared to 0.477, and for the Multi IHC Stain Translation dataset, it scored 0.242, exceeding 0.2236. These findings underscore the model's effectiveness on larger datasets and highlight the potential of diffusion models over GAN-based methods for medical image translation tasks.

Zusammenfassung

Aktuelle Techniken zur Bild-zu-Bild-Übersetzung in der Mikroskopie basieren überwiegend auf generativen adversarialen Netzwerken (GANs). Obwohl diffusionsbasierte Modelle in letzter Zeit zunehmend an Bedeutung gewinnen, bleibt ihre Anwendung in der histologischen Bildübersetzung begrenzt. Diese Arbeit adressiert diese Forschungslücke, indem sie moderne Bildübersetzungsmethoden in der medizinischen Mikroskopie mit latenten Diffusionsansätzen vergleicht. Ein zentraler Beitrag dieser Arbeit ist ein neuartiges latentes Verfeinerungsnetzwerk, das auf der Swin-Transformer-Architektur basiert. In diesem Ansatz wird das bedingende Bild zunächst durch einen Variational Autoencoder (VAE) codiert und anschließend im latenten Raum unter Verwendung des Swin-Transformers weiter verfeinert. Diese Methode betont die entscheidende Bedeutung eines gut vortrainierten VAEs für eine effektive und qualitativ hochwertige Bildübersetzung. Diese Vorgehensweise eröffnet neue Perspektiven für zukünftige Forschungsarbeiten, wie etwa Few-Shot-Bild-zu-Bild-Übersetzung und spezialisiertes Pretraining für medizinische Bildgebungsaufgaben. Zudem bietet die Arbeit einen umfassenden Überblick über sowohl aktuelle als auch grundlegende Bild-zu-Bild-Übersetzungsmethoden und analysiert deren Stärken und Schwächen.

Zur Evaluierung des vorgeschlagenen Ansatzes wurden Experimente mit vier verschiedenen Datensätzen durchgeführt, die jeweils unterschiedliche technische Herausforderungen mit sich bringen. Auf Datensätzen mit präzisen Annotationen und begrenzter Probengröße erzielte das Modell einen Structural Similarity Index Measure (SSIM) von 0,74, was dem aktuellen Stand der Technik von 0,752 auf einem Brightfield-Bild-zu-Bild-Übersetzungsdatensatz nahekommt. Des Weiteren zeigte das Modell bei größeren Datensätzen eine überlegene Leistung im Vergleich zu führenden Algorithmen. So erreichte es auf einem Unstained-to-Stained-Datensatz einen SSIM von 0,613 und übertraf damit Pix2Pix mit 0,512. Beim Brustkrebs-Immunhistochemie-Datensatz erzielte es einen SSIM von 0,496 im Vergleich zu 0,477 und beim Multi-IHC-Färbeübersetzungsdatensatz einen Wert von 0,242 gegenüber 0,2236. Diese Ergebnisse unterstreichen die Effektivität des vorgeschlagenen Modells bei größeren Datensätzen und verdeutlichen das Potenzial diffusionsbasierter Methoden gegenüber GAN-basierten Ansätzen in der medizinischen Bildübersetzung.

Contents

1 Introduction and Outline **1**
- 1.1 Introduction . . . 1
- 1.2 Goal . . . 3
- 1.3 Outline . . . 3

2 Foundations **5**
- 2.1 Neural Networks . . . 5
  - 2.1.1 Neuron . . . 5
  - 2.1.2 Multilayer Perceptron . . . 6
  - 2.1.3 Activation Functions . . . 6
  - 2.1.4 Model Training . . . 9
- 2.2 Variational Autoencoders . . . 12
- 2.3 Diffusion Models . . . 13
  - 2.3.1 Forward Diffusion Process . . . 13
  - 2.3.2 Reverse Diffusion Process . . . 14
  - 2.3.3 DDPM and DDIM . . . 14
    - 2.3.3.1 Denoising Diffusion Probabilistic Models . . . 15
    - 2.3.3.2 Denoising Diffusion Implicit Models . . . 15
  - 2.3.4 Guided Diffusion . . . 16
    - 2.3.4.1 Classifier-Guided Diffusion . . . 16
    - 2.3.4.2 Classifier-Free Guidance . . . 16
  - 2.3.5 Latent Diffusion Models . . . 17
- 2.4 Attention . . . 18
  - 2.4.1 Mathematical Formulation . . . 18
  - 2.4.2 Self-Attention and Multi-Head Attention . . . 19
  - 2.4.3 Cross-Attention . . . 19
  - 2.4.4 Vision Transformer . . . 20
    - 2.4.4.1 Patch Embeddings . . . 20
    - 2.4.4.2 Transformer Encoder . . . 20
    - 2.4.4.3 Classification Head . . . 21
- 2.5 Optical Microscopy . . . 21
  - 2.5.1 Brightfield Microscopy . . . 22
  - 2.5.2 Differential Interference Contrast . . . 22
  - 2.5.3 Phase Contrast Imaging . . . 23
  - 2.5.4 Fluorescence Microscopy . . . 24

3 Task and Data Analysis **29**
3.1 Bright Field to Fluorescence . . . 29
3.1.1 Biological and Medical Motivations . . . 29
3.1.2 Technical Motivations . . . 30
3.1.3 Dataset Description . . . 30
3.2 Unstained to Stained . . . 31
3.2.1 Biological and Medical Motivations . . . 31
3.2.2 Technical Motivations . . . 32
3.2.3 Dataset Description . . . 32
3.3 Breast Cancer Immunohistochemical Dataset . . . 33
3.3.1 Biological and Medical Motivations . . . 33
3.3.2 Technical Motivations . . . 33
3.3.3 Dataset Description . . . 34
3.4 Multi IHC Stain Translation . . . 34
3.4.1 Biological and Medical Motivations . . . 34
3.4.2 Technical Motivations . . . 35
3.4.3 Dataset Description . . . 35

4 Related Work **37**
4.1 Traditional Approaches . . . 37
4.1.1 Histogram transformation . . . 37
4.1.2 Classical Machine Learning Techniques . . . 38
4.1.3 Color Deconvolution . . . 38
4.2 Deep Learning-Based Approaches . . . 39
4.2.1 Autoencoders . . . 39
4.2.2 Generative Adversarial Networks . . . 39
4.2.3 VQGAN . . . 40
4.2.4 Diffusion Based Image to Image translation . . . 40

5 Methods **43**
5.1 Competition Models . . . 43
5.1.1 Preprocessing . . . 43
5.1.2 Autoencoder . . . 44
5.1.3 Generative Adversial Neural Network . . . 45
5.1.4 Losses . . . 46
5.1.5 Dealing with missing organelles . . . 47
5.2 Hybrid Swin Conditional Refinement Latent Diffusion Model . . . 47
5.2.1 Preprocessing . . . 47
5.2.2 Variational Autoencoder Architecture . . . 48
5.2.2.1 Encoder . . . 48
5.2.2.2 Latent Space Representation . . . 48
5.2.2.3 Decoder . . . 48
5.2.2.4 Loss . . . 49
5.2.3 Conditional Module . . . 50
5.2.4 Spatial Cross-Attention . . . 51
5.2.5 Latent Diffusion Model . . . 53
5.2.6 Variations . . . 53

6 Evaluation **55**
6.1 Variational Autoencoder Finetuning 55
6.2 Brightfield to Fluorescence 58
6.3 Unstained to Stained 60
6.4 Breast Cancer Immunohistochemical Dataset 62
6.5 Multi IHC Stain Translation 64

7 Conclusion **67**
7.1 Summary 67
7.2 Future Work 68

Bibliography **69**

Appendix **79**

1. Introduction and Outline

This chapter provides an introduction to the topic of image-to-image translation in medical microscopy and explains why it is an important area of research. Additionally, the general goal of the thesis is presented, along with an overview of the outline and structure of the thesis.

1.1 Introduction

Medical diagnosis frequently relies on the examination of microscopic images, which are essential for detecting and diagnosing diseases. However, since many cells are nearly transparent, cell markers are often used to enhance contrast and reveal the underlying tissue structures and components. Over the past century, histological staining has become a critical tool in medicine, aiding both disease diagnosis and tissue examination [BG08]. Histological staining enables the visualization of different tissue structures when these stained samples are observed under a microscope [Tit09]. One of the most widely used stains is hematoxylin and eosin (H&E), which highlights the contrast between cell nuclei and the surrounding extracellular tissue [Tit05]. Additionally, immunohistochemical staining, which emphasizes collagen fibers and glycoproteins, is often employed in kidney tissue imaging [CCJ41]. These staining methods, by targeting specific molecular components, provide crucial insights into tissue architecture, cellular composition, and disease states. This information is essential for diagnosing conditions such as cancer, fibrosis, and autoimmune diseases.

While these staining techniques are invaluable, they are also labor-intensive and time-consuming, involving multiple steps such as tissue fixation, embedding, and sectioning. Furthermore, these chemical-based techniques require skilled histotechnologists, making them both costly and resource-intensive. In some cases, multiple different markers must be applied to a single tissue, a process known as multiplex staining [TNC+20], which is even more time-consuming and expensive. Even single-staining processes can take days or weeks to complete, due to the number of steps required. The complexity of traditional staining methods limits the speed at which diagnoses can be made, particularly in high-throughput clinical environments where rapid decision-making is essential. This is especially relevant

in cases like cancer biopsies, where a swift diagnosis can significantly impact treatment outcomes.

Recently, digital pathology has gained significant traction among practitioners [AJHVD12, ML16]. With digital pathology, high-throughput slide scanners are used to create digital images, which can then be analyzed using artificial intelligence. This approach enables faster, more affordable access to deep learning-based image transformations and diagnostics [WRJ+19, WLC+19]. One major innovation in this field is the development of virtual staining and stain-to-stain translation techniques. These methods can either completely eliminate the need for chemical-based staining or reduce the number of stains that must be applied to tissue samples [RWW+19]. By bypassing the physical staining process, virtual staining not only accelerates tissue analysis but also reduces the variability associated with manual staining procedures, improving the consistency and reliability of diagnostic results.

These virtual staining methods have already been applied to various tasks, such as label-free imaging modalities like autofluorescence [RWW+19], which is the first type of image dealt with in this thesis. Another area of application is stain-to-stain translation, where deep learning models have been employed to reduce the need for multiplex staining, allowing multiple types of stains to be virtually generated from a single tissue sample [D+20, dHZZ+21]. This innovation opens up new possibilities for simultaneously visualizing different molecular markers, which can improve diagnostic precision, particularly in complex diseases such as cancer where multiple biomarkers may need to be assessed in parallel. The advantage of generating multiple types of stains instantly is that different downstream analytical models can be used to support diagnosis, reducing the delay from tissue extraction to staining and diagnosis [LGZ+21].

Microscopy images can be obtained using various techniques, including brightfield microscopy [ODKD10], phase contrast microscopy [BS42], and differential interference contrast (DIC) microscopy [L+82], all of which play a significant role in this thesis. These images can be further processed to apply fluorescence markers, which help highlight different types of tissues on the sample slide. Each microscopy technique provides a unique view of the tissue's morphology, helping clinicians and researchers to gain comprehensive insights into cellular structures and interactions. Brightfield microscopy offers clear visualization of stained samples, while phase contrast and DIC microscopy enhance contrast in transparent samples without staining, making them essential for label-free imaging.

As mentioned, typical datasets can be divided into two categories: label-free imaging tasks, where the input consists of unstained tissue images directly captured under the microscope, and stain-to-stain image translation tasks. Although the nature of the tasks may differ, the deep learning models used for these tasks are generally similar. A key challenge in stain-to-stain translation tasks is the need to align tissue images with different stains, requiring precise registration to generate paired samples [MVH11]. Accurate image registration is crucial for ensuring that corresponding features in different stains are correctly aligned, especially when dealing with high-resolution images where even small misalignments can lead to incorrect interpretations.

There are several types of models used for image-to-image translation, which can be broadly categorized into autoencoders and generative models. Autoencoders, though not

generative models as they cannot sample new images and are deterministic, have been successfully applied to tasks such as brightfield-to-fluorescence translation through multi-task learning and adversarial losses [LOHJ21]. However, generative models, particularly those based on generative adversarial networks (GANs), are more commonly applied to image-to-image translation tasks. For aligned datasets, Pix2Pix [IZZE17], based on GANs, is one of the most widely used models for various medical image translation tasks. GANs have proven to be highly effective for tasks requiring realistic image synthesis, as their adversarial nature enables them to produce high-fidelity images that closely resemble the target distribution. More recently, diffusion-based models have emerged as a powerful tool for image generation and will be extensively covered in this thesis [HJA20, RBL+22]. These models leverage iterative denoising processes to generate high-quality images from noisy inputs, offering promising results for complex tasks such as generating diverse stains from a single tissue sample.

1.2 Goal

The goal of this thesis is to research existing methods for image-to-image translation in microscopy images and to identify their advantages and disadvantages. Additionally, the work aims to compile various datasets that capture a wide range of image-to-image translation tasks in microscopy and reflect the broad spectrum of challenges within this domain. Building on this foundation, a new method will be developed that demonstrates competitive performance on these datasets and aims to outperform existing methods.

To begin, participation in a global competition is planned, enabling direct comparison with other researchers and providing an opportunity to validate the effectiveness of the proposed approach. The method developed will explore the potential of latent diffusion models for image-to-image translation and seek to improve their performance to deliver higher-quality results.

1.3 Outline

The structure of this thesis follows a logical progression to introduce the reader to the foundational concepts, related work, datasets, methods, and results:

- **Foundations in Deep Learning and Microscopy:** This section provides an overview of the key concepts in deep learning and the relevant microscopy techniques used throughout the thesis. It ensures that the reader has a solid understanding of both the computational and biological aspects required to comprehend the techniques and datasets applied in later sections.
- **Datasets:** This section introduces the various datasets used in the thesis, explaining their biological motivations and technical challenges. A detailed statistical analysis of the datasets is presented to demonstrate their complexity and the technical issues they pose, including variability in staining and image alignment.
- **Related Work:** The related work chapter surveys the existing literature on staining normalization, stain-to-stain translation, and image generation. It covers both traditional and modern deep learning-based methods, discussing their evolution over recent decades. This section helps contextualize the research contributions made in this thesis by highlighting the limitations of previous approaches.

- **Methods and Contributions:** Here, the thesis describes the different methods applied, building on the foundations and related work. This section also introduces the final contributions of the thesis, focusing on the new approaches developed to address the challenges identified in the related work.
- **Results:** Both quantitative and qualitative results are presented, comparing baseline methods with the newly developed method. This section highlights the strengths and weaknesses of the proposed method, showing where it improves upon existing techniques and where limitations remain.

2. Foundations

This chapter gives a foundational introduction to deep learning, highlighting the core concepts utilized in this work. It provides an overview of deep learning principles and describes the training process of these systems. Additionally, a brief explanation of the self-attention mechanism is included, as it plays a crucial role in one of the primary architectures examined in this thesis.

2.1 Neural Networks

This section introduces artificial neural networks, which are the core of this work.

Artificial neural networks are statistical models drawing inspiration from biological neural networks. They attempt to mimic the functionality of brain neurons, which are interconnected via synapses and activate in response to specific inputs. As a fundamental component of this work, they warrant thorough examination.

2.1.1 Neuron

The neuron, or perceptron, forms a crucial element of neural networks[Ros58]. It serves as the basic unit of modern neural networks. In microscopy image processing, a neuron can be conceptualized as a function that processes input features, potentially representing pixel intensities or derived image characteristics.

Formally, a neuron implements the function:

$$f(x, w) = \phi\left(\sum_i x_i w_i + b\right) \tag{2.1}$$

Where x represents the input vector, w the weight vector, b the bias term, and ϕ the activation function. The activation function introduces non-linearity, allowing the network to model complex patterns in microscopy data.

2.1.2 Multilayer Perceptron

The multilayer perceptron [RHW86a] is a type of deep learning model comprising three or more layers of neurons: an input layer, k hidden layers, and an output layer. Each layer consists of n neurons, where n is a tunable hyperparameter. In image-to-image translation for microscopy, the input layer might receive one modality of microscopy image, while the output layer produces another.

An MLP defines a function $f : \mathbb{R}^i \to \mathbb{R}^o$, where i is the dimension of the input vector and o the dimension of the output vector. In matrix notation, a single hidden layer MLP can be expressed as:

$$f(x) = \phi_1(b^2 + W^2(\phi_2(b^1 + W^1 x))) \tag{2.2}$$

Here, ϕ_n denote activation functions, b^n bias vectors, W^n weight matrices, and x the input vector. As all neurons in each layer connect to all neurons in the subsequent layer, this architecture is termed a fully connected neural network.

2.1.3 Activation Functions

Activation functions play a crucial role in neural networks, especially in generative models used for microscopy image processing such as autoencoders, variational autoencoders, diffusion models, and image transformers. The choice of activation function can significantly impact the model's performance and training dynamics. Here some activation functions extensivly used in this thesis are introduced.

Leaky ReLU

The Leaky ReLU [XLD+20] is defined as:

$$f(x) = \max(\alpha x, x) \tag{2.3}$$

where α is a small constant, typically 0.01.

Its corresponding derivative is:

$$f'(x) = \begin{cases} \alpha, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases} \tag{2.4}$$

Leaky ReLU addresses the "dying ReLU" [LSSK19] problem by allowing a small, non-zero gradient when the unit is not active. This can be particularly useful in autoencoders and VAEs, where maintaining gradient flow through all parts of the network is crucial for effective training.

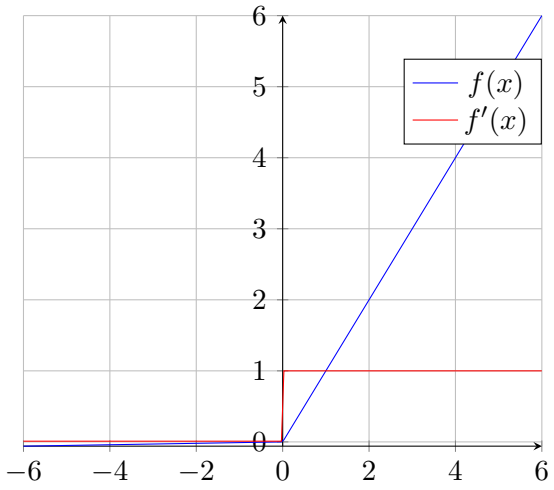


Figure 2.1: Graph of the Leaky ReLU activation function (blue) with its derivative (red).

GELU (Gaussian Error Linear Unit)

The GELU [HG16] activation is defined as:

$$\text{GELU}(x) = x \cdot \Phi(x) \tag{2.5}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

GELU has gained popularity in transformer architectures, including those used for image processing tasks. It provides a smooth approximation of ReLU and has been shown to perform well in self-attention mechanisms, making it suitable for image transformer models.

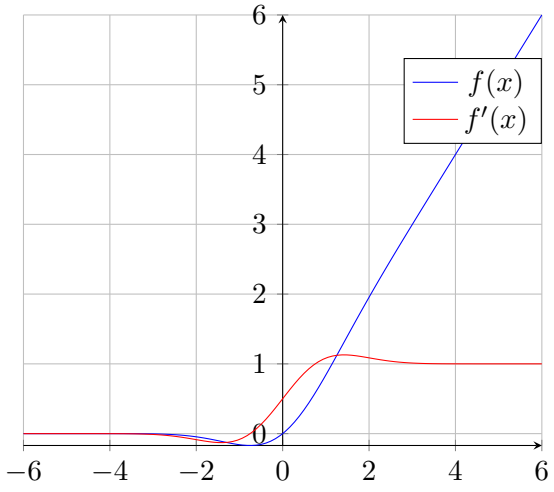


Figure 2.2: Graph of the GELU activation function (blue) with its derivative (red).

Swish

The Swish [RZL17] activation function is defined as:

$$f(x) = x \cdot \sigma(\beta x) \tag{2.6}$$

where σ is the sigmoid function and β is a learnable parameter or a constant (usually set to 1).

Swish has shown promising results especially in diffusion models. Its non-monotonic nature allows it to carry both positive and negative information, which can be beneficial in complex image generation tasks.

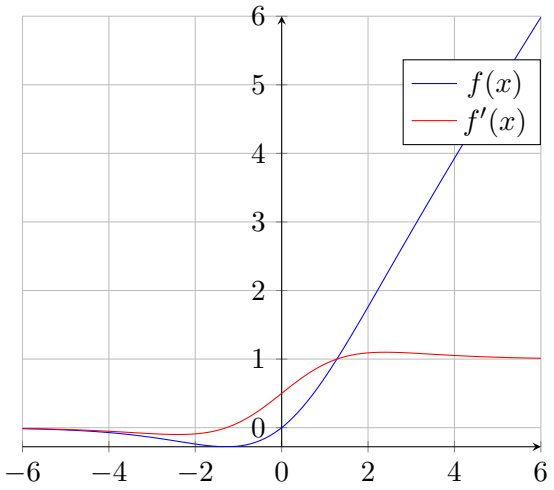


Figure 2.3: Graph of the Swish activation function (blue) with its derivative (red) for $\beta = 1$.

Mish

Mish [Mis20] is defined as:

$$f(x) = x \cdot \tanh(\ln(1 + e^x)) \tag{2.7}$$

Mish is a self-regularizing non-monotonic activation function that has shown good performance in various deep learning tasks, including those involving image processing. Its smooth nature can help in gradient flow, making it suitable for deep autoencoders and generative models.

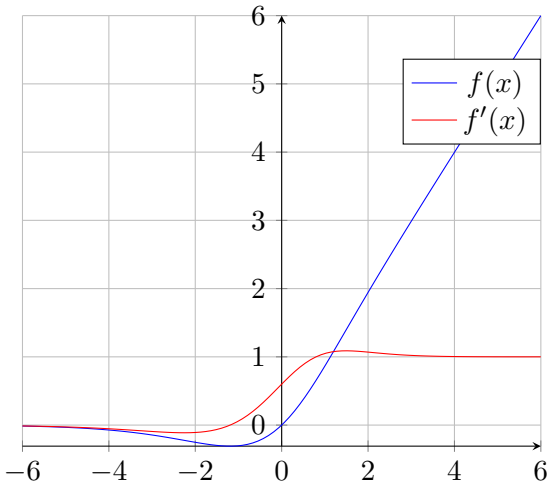


Figure 2.4: Graph of the Mish activation function (blue) with its derivative (red).

The choice of activation function in these advanced architectures often depends on the specific requirements of the task and the characteristics of the microscopy data being processed. For instance, Leaky ReLU or Mish might be preferred in the encoder and decoder parts of autoencoders and VAEs to ensure good gradient flow. GELU is often used in transformer-based models due to its compatibility with self-attention mechanisms. In diffusion models, Swish activations are commonly employed in the underlying neural networks that model the noise prediction or denoising process.

It's worth noting that the effectiveness of these activation functions can vary depending on the specific architecture, dataset, and task at hand. Therefore, empirical evaluation and careful consideration of the model's requirements are crucial when selecting activation functions.

2.1.4 Model Training

Backpropagation of Errors

Understanding the interplay of activation functions, their derivatives, and loss functions in training neural networks requires comprehension of backpropagation. Backpropgation [RHW86b] has become the standard for training various neural network architectures due to its efficiency in automatic differentiation of the loss function.

Backpropagation calculates the gradient of the loss function with respect to all model parameters:

$$\nabla L(y, \hat{y}) = \frac{\partial L(y, \hat{y})}{\partial \theta} \tag{2.8}$$

where L is the loss function, y is the true value, $\hat{y}$ is the predicted value, and θ represents all model parameters.

The algorithm's popularity stems from its simplicity and power, particularly its ability to train nonlinear neural networks with arbitrary connectivity, which is crucial for real-world problems that are inherently nonlinear.

Gradient Descent

Gradient descent [Rud16] is an iterative optimization process that adjusts model parameters to minimize error. Starting with randomly initialized parameters, vanilla gradient descent updates the model's parameters θ as follows:

$$\theta^{t+1} = \theta - \alpha \cdot \nabla L(\theta) \tag{2.9}$$

Here, α is the learning rate, a critical hyperparameter. If α is too high, the model may fail to converge or even diverge. Conversely, a very small α ensures convergence but at the cost of increased training time.

For convex loss surfaces, gradient descent guarantees convergence to the global minimum with an appropriate α. However, for non-convex surfaces, it may converge to a local minimum.

Vanilla gradient descent calculates the gradient over the entire training dataset before performing a single parameter update, which can be slow and memory-intensive for large datasets. It also precludes online training.

Stochastic Gradient Descent (SGD)

SGD [Rob51] addresses the limitations of vanilla gradient descent by updating parameters after each training sample:

$$\theta^{t+1} = \theta - \alpha \cdot \nabla L(\theta, y_i, \hat{y}_i) \tag{2.10}$$

This approach eliminates redundant computations, allows for online training, and is more suitable for large datasets. However, updating θ after each sample introduces fluctuations in the loss function.

SGD's ability to handle large datasets and perform online training makes it particularly relevant in modern machine learning contexts where new data is continuously available.

Mini-batch Gradient Descent

Mini-batch gradient descent strikes a balance between vanilla and stochastic gradient descent. It updates the model parameters θ using mini-batches of size $n \in \mathbb{N}$, where n is a hyperparameter known as the batch size. This approach reduces the variance of the loss function, potentially leading to smoother convergence. The update rule is:

$$\theta^{t+1} = \theta - \alpha \cdot \nabla L(\theta, y_{i:i+n}, y_{i:\hat{i+n}}) \tag{2.11}$$

The mini-batch size is primarily limited by available memory for storing gradients. There are techniques such as gradient accumulation [HSM17] that allow the use of large batch sizes without requiring additional memory. This is achieved by postponing the parameter update step until after several iterations, during which the gradients are accumulated and then scaled at the end. Additionally, there is theoretical background suggesting that smaller batch sizes can also be beneficial [LXT+18, ML18, QK20].

Adam (Adaptive Moment Estimation)

Adam [KB14], is a popular first-order gradient-based optimization algorithm. It combines advantages from Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp), introducing adaptive learning rates for each parameter.

The update rules for Adam are:

$$\begin{aligned} m_t &= \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \\ v_t &= \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot (g_t)^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \end{aligned} \tag{2.12}$$

The final update rule for Adam is:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{2.13}$$

Here, β_1, β_2, and ϵ are hyperparameters, often kept at their default values.

AdamW

AdamW [LH17] is a variant of Adam that implements decoupled weight decay regularization. It addresses the issue of L2 regularization not functioning as intended when used with adaptive gradient algorithms like Adam.

The key difference in AdamW is how weight decay is applied. Instead of incorporating it into the gradient computation, AdamW directly applies weight decay to the parameters:

$$\theta_{t+1} = \theta_t - \alpha \cdot (\frac{1}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t + \lambda \theta_t) \tag{2.14}$$

Where λ is the weight decay factor.

This modification allows for more effective regularization, often leading to better generalization in deep learning models. AdamW has shown particular success in training large-scale models and is frequently used in state-of-the-art architectures for tasks such as image-to-image translation in microscopy.

The choice between Adam and AdamW often depends on the specific requirements of the task and the characteristics of the dataset. In practice, AdamW is increasingly favored for its improved generalization capabilities, especially in complex deep learning models.

2.2 Variational Autoencoders

Variational Autoencoders [LBY+20] (VAEs) are a type of generative model. They can learn complex probability distributions over high dimensional data.

At its core, a VAE consists of two primary components: an encoder network $q_\phi(z|x)$, which approximates the intractable posterior distribution $p_\theta(z|x)$, and a decoder network $p_\theta(x|z)$, modeling the likelihood of the data given the latent variables. Both networks are parameterized by neural networks, enabling them to learn intricate, non-linear mappings between the data space and the latent space.

VAEs are trained by maximizing the variational lower bound (ELBO) on the log-likelihood of the data. This objective function is expressed as:

$$\mathcal{L}(\theta, \phi; x^{(i)}) = \mathbb{E}_{q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] - D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z))$$

Here, the first term represents the expected reconstruction log-likelihood, while the second term is the Kullback-Leibler divergence between the approximate posterior and the prior.

A key addition in VAEs is the reparameterization trick, which enables backpropagation through the sampling process. Instead of directly sampling from $q_\phi(z|x)$, we express z as a deterministic function of x and an auxiliary noise variable ϵ:

$$z = g_\phi(\epsilon, x) \quad \text{where} \quad \epsilon \sim p(\epsilon)$$

For Gaussian distributions, this is typically implemented as:

$$z = \mu + \sigma \odot \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, I)$$

The VAE training process involves stochastic gradient descent on minibatches of data. The procedure can be summarized in the following algorithm:

Algorithm 1 VAE Training Procedure

for each epoch **do**
  for each minibatch $\{x^{(i)}\}$ where $i = 1$ to M **do**
    for each datapoint $x^{(i)}$ in minibatch **do**
      Sample $\epsilon \sim p(\epsilon)$
      Compute $z = g_\phi(\epsilon, x^{(i)})$
      Compute ELBO $\mathcal{L}(\theta, \phi; x^{(i)})$
    end for
    Compute average ELBO over minibatch
    Compute gradients $\nabla_{\theta,\phi}\mathcal{L}(\theta, \phi)$
    Update parameters θ and ϕ using optimizer (e.g., Adam)
  end for
end for

In the common scenario where both the prior $p_\theta(z)$ and the approximate posterior $q_\phi(z|x^{(i)})$ are Gaussian, the KL-divergence term in the ELBO can be computed analytically:

$$-D_{\mathrm{KL}}(q_\phi(z|x^{(i)}) \,\|\, p_\theta(z)) = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\left((\sigma_j^{(i)})^2\right) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2\right)$$

where J is the dimensionality of z, and $\mu_j^{(i)}$ and $\sigma_j^{(i)}$ are the j-th elements of the mean and standard deviation vectors for datapoint i, respectively.

2.3 Diffusion Models

Diffusion models [HJA20] are a type of generative model that have gained significant attention in recent years because of its breakthroughs in image genereation [RBL+22]. Drawing inspiration from non-equilibrium thermodynamics, these models create a Markov chain of diffusion steps to progressively introduce random noise into data, then learn how to reverse this process to generate samples from noise. Unlike VAEs or flow models, diffusion models use a fixed procedure and work with high-dimensional latent variables, often matching the dimensionality of the original data.

The key idea behind diffusion models is to transform a complex data distribution into a simple, tractable distribution through a series of small perturbations, and then learn to invert this process. This approach allows for highly flexible and expressive generative modeling, capable of producing high-quality samples across various domains.

2.3.1 Forward Diffusion Process

The forward diffusion process forms the foundation of diffusion models. It describes a gradual transformation of data points into noise through a sequence of T steps. Given a data point x_0 sampled from a real data distribution $q(x_0)$, the forward diffusion process adds small amounts of Gaussian noise at each step, producing a sequence of increasingly noisy samples $x_1, x_2, ..., x_T$.

Formally, the forward process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

where β_t is the variance schedule that controls the step sizes, and I is the identity matrix. As t approaches T, x_t becomes increasingly noisy, eventually converging to an isotropic Gaussian distribution when $t = T$.

A notable property of this process is the ability to sample x_t at any arbitrary time step t in closed form using the reparameterization trick:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, and $\epsilon \sim \mathcal{N}(0, I)$.

2.3.2 Reverse Diffusion Process

The reverse diffusion process is the heart of the generative capability in diffusion models. It aims to reconstruct the original data by learning to denoise the samples, effectively reversing the forward process. Ideally, sampling from $p(x_{t-1}|x_t)$ would be preferred, but this is intractable as it requires knowledge of the entire dataset. Instead, a model $p_\theta(x_{t-1}|x_t)$ is learned to approximate these conditional probabilities.

The reverse process is defined as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

where μ_θ and Σ_θ are learned functions parameterized by a neural network with parameters θ.

To train this model, optimization of a variational lower bound on the negative log-likelihood is performed:

$$\mathcal{L} = \mathbb{E}_{q(x_{0:T})}\left[-\log p(x_T) + \sum_{t=1}^{T} \mathrm{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\right]$$

This objective can be further simplified and reparameterized to predict the noise ϵ added during the forward process:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right]$$

where ϵ_θ is the noise prediction network.

The training procedure for diffusion models can be summarized in the following algorithm:

Algorithm 2 Diffusion Model Training

for each iteration **do**
  Sample data x_0 from the dataset
  Sample a time step t uniformly from $\{1, ..., T\}$
  Sample noise $\epsilon \sim \mathcal{N}(0, I)$
  Compute noisy sample $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$
  Compute loss $\mathcal{L} = \|\epsilon - \epsilon_\theta(x_t, t)\|^2$
  Update θ to minimize $\mathcal{L}$ using gradient descent
end for

2.3.3 DDPM and DDIM

Two prominent diffusion model variants are Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM) [SME20]. While both are based on the same underlying principles, they differ in their sampling procedures and trade-offs between sample quality and generation speed.

2.3.3.1 Denoising Diffusion Probabilistic Models

DDPM, introduced by Ho et al., follows the standard diffusion process described above. It uses a Markovian forward process and learns to reverse this process step by step. The sampling procedure in DDPM is inherently stochastic, requiring many steps (typically 1000 or more) to generate high-quality samples.

Key characteristics of DDPM:

- Stochastic sampling process
- High sample quality
- Requires many sampling steps (slow generation)
- Strong theoretical grounding in probabilistic modeling

2.3.3.2 Denoising Diffusion Implicit Models

DDIM, proposed by Song et al., modifies the diffusion process to allow for faster sampling. It introduces a non-Markovian forward process that, when reversed, enables deterministic sampling and fewer generation steps.

Key characteristics of DDIM:

- Deterministic sampling process
- Comparable sample quality to DDPM
- Requires fewer sampling steps (faster generation)
- Allows for controlled generation through latent space manipulation

The primary difference lies in the forward process. While DDPM uses a Markovian process where each step only depends on the previous one, DDIM constructs a non-Markovian process that depends on all previous steps. This modification allows DDIM to take "shortcuts" during the reverse process, significantly reducing the number of required sampling steps while maintaining sample quality.

Mathematically, DDIM modifies the forward process as:

$$q_\sigma(x_t|x_0) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t - \sigma_t^2}\epsilon_0 + \sigma_t\epsilon_1$$

where σ_t is a new hyperparameter controlling the stochasticity of the process. When $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$, DDIM reduces to DDPM.

The choice between DDPM and DDIM depends on the specific requirements of the application. DDPM may be preferred when sample quality is paramount and generation speed is not a critical factor. Conversely, DDIM is advantageous in scenarios requiring faster generation or when latent space manipulation is desired.

2.3.4 Guided Diffusion

Guided diffusion techniques have improved diffusion models by allowing for more controlled and precise generation. These methods make it possible to include extra information or constraints during the generation process, leading to outputs that meet specific conditions or attributes. Two main approaches to guided diffusion are classifier-guided diffusion and classifier-free guidance.

2.3.4.1 Classifier-Guided Diffusion

Classifier-guided diffusion [NDR+21] uses a pre-trained classifier to steer the generation process towards desired conditions. This approach involves training a classifier $f_\phi(y|x_t, t)$ on noisy images x_t at various time steps t. During the reverse diffusion process, the gradients from this classifier are used to modify the noise prediction, guiding the generation towards the desired class or attribute.

The modified noise prediction can be expressed as:

$$\tilde{\epsilon}_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) - \sqrt{1 - \bar{\alpha}_t}\nabla_{x_t} \log f_\phi(y|x_t, t)$$

where $\epsilon_\theta(x_t, t)$ is the original noise prediction, and $\nabla_{x_t} \log f_\phi(y|x_t, t)$ represents the gradient of the log-probability of the desired class y with respect to the input x_t.

The classifier guidance strength can be controlled by introducing a scaling factor s:

$$\tilde{\epsilon}_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) - s \cdot \sqrt{1 - \bar{\alpha}_t}\nabla_{x_t} \log f_\phi(y|x_t, t)$$

This approach allows for strong conditioning and can be particularly effective for tasks such as class-conditional image generation or attribute-based editing. However, it requires training and maintaining an additional classifier model, which can increase computational overhead.

2.3.4.2 Classifier-Free Guidance

Classifier-free guidance [HS22], achieves conditional generation without the need for a separate classifier. This method combines a conditional model $\epsilon_\theta(x_t, t, y)$ and an unconditional model $\epsilon_\theta(x_t, t)$ within a single network. The key idea is to train the model to handle both conditional and unconditional generation, then interpolate between these outputs during inference.

The classifier-free guided noise prediction is defined as:

$$\tilde{\epsilon}_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + s \cdot (\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$$

where s is a guidance scale parameter that controls the strength of the conditioning. When $s = 0$, the generation process is equivalent to unconditional generation, while larger values of s result in stronger adherence to the conditioning information.

Algorithm 3 Classifier-Free Guided Diffusion Training

for each iteration **do**
  Sample data x_0 and condition y from the dataset
  Sample a time step t uniformly from $\{1, ..., T\}$
  Sample noise $\epsilon \sim \mathcal{N}(0, I)$
  Compute noisy sample $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$
  With probability p:
    Compute loss $\mathcal{L} = \|\epsilon - \epsilon_\theta(x_t, t, y)\|^2$
  Otherwise:
    Compute loss $\mathcal{L} = \|\epsilon - \epsilon_\theta(x_t, t)\|^2$
  Update θ to minimize $\mathcal{L}$ using gradient descent
end for

The training procedure for classifier-free guidance involves randomly dropping the conditioning information during training, allowing the model to learn both conditional and unconditional generation:

Classifier-free guidance has shown impressive results across various conditional generation tasks, including text-to-image synthesis and class-conditional image generation. Its simplicity and effectiveness, coupled with the ability to control the guidance strength at inference time, have made it a popular choice in recent diffusion model applications.

2.3.5 Latent Diffusion Models

Latent Diffusion Models (LDMs) make diffusion models more efficient by performing the diffusion process in a compressed latent space rather than in the full pixel space. This approach reduces computational demands while still producing high-quality images.

The LDM process has three main steps:

1. **Perceptual Compression:** An encoder $\mathcal{E}$ compresses the input image x into a latent vector $z = \mathcal{E}(x)$, reducing the data's dimensionality while keeping essential information.

2. **Latent Diffusion:** Diffusion occurs in this smaller latent space, following similar principles to pixel-space diffusion but at a lower computational cost.

3. **Decoding:** A decoder $\mathcal{D}$ reconstructs the image from the processed latent vector, $\hat{x} = \mathcal{D}(z)$.

The LDM objective function is defined as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(x),\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathcal{E}(x) + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2\right] + \mathcal{L}_{\text{rec}}$$

where $\mathcal{L}_{\text{rec}}$ represents a reconstruction loss for the autoencoder.

LDMs also allow for conditioning on various types of information through cross-attention mechanisms, enabling flexible image generation. This conditioning process is described by:

$$\tau_\theta(z_t, t, c) = \text{Attention}(\text{LayerNorm}(z_t), \psi(c))$$

where c is the conditioning information, and ψ is an encoder specific to the conditioning domain. An illustration of the latent diffusion process, based on [RBL+22], is shown in Figure 2.5.

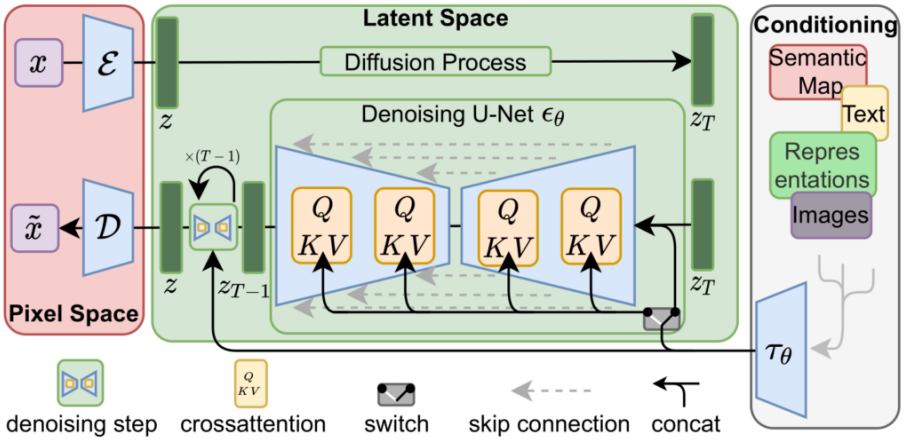


Figure 2.5: Latent diffusion model [RBL+22]

2.4 Attention

The attention mechanism is a method for selectively focusing on different parts of the input data to make predictions. In deep learning, it assigns importance to each part of the input, allowing the model to concentrate on relevant information while ignoring less important parts. This selective focus enables handling long-range dependencies, especially in tasks like image-to-image translation.

2.4.1 Mathematical Formulation

The attention mechanism uses three key components: queries (Q), keys (K), and values (V), typically derived from the input using learned weight matrices. If the input is represented as $X \in \mathbb{R}^{n \times d}$ (where n is the number of input elements and d is the feature dimension), the transformations to Q, K, and V are performed as follows:

$$Q = XW^Q \tag{2.15}$$
$$K = XW^K \tag{2.16}$$
$$V = XW^V \tag{2.17}$$

where W^Q, W^K, and W^V are learnable parameter matrices. The attention weights are computed by:

1. Calculating the similarity between queries and keys:

$$S = QK^T \tag{2.18}$$

2. Applying softmax to obtain normalized attention weights:

$$A = \text{softmax}\left(\frac{S}{\sqrt{d_k}}\right) \tag{2.19}$$

3. Computing the weighted sum of the values:

$$O = AV \tag{2.20}$$

Thus, the full attention function is represented by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.21}$$

2.4.2 Self-Attention and Multi-Head Attention

Self-attention uses the same input for queries, keys, and values, allowing each part of the input to interact with every other part. This is beneficial for tasks like image translation, where capturing spatial relationships is essential.

Multi-head attention enhances this by using multiple attention functions in parallel, each with its own set of learned weights:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{2.22}$$

Each head is computed as follows:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.23}$$

where W^O is a learnable parameter matrix.

2.4.3 Cross-Attention

Cross-attention extends the attention mechanism to operate on two different sets of inputs. Unlike self-attention, where queries, keys, and values come from the same source, in cross-attention, the queries are derived from one input while the keys and values are derived from another. This is essential in tasks that require alignment or comparison between different inputs, such as multi-modal processing.

In this thesis, cross-attention is crucial for image-to-image translation, particularly when using latent diffusion models. For guidance, latent diffusion models often cross attend [RBL+22] to the condition and therefore can generate output based on the condition.

2.4.4 Vision Transformer

The Vision Transformer (ViT) applies the Transformer architecture to images by treating an image as a sequence of smaller patches, instead of using convolutional filters like traditional CNNs. This allows ViT to capture long-range dependencies within an image as supposed to a convolutional based architecture.

2.4.4.1 Patch Embeddings

To process an image with a transformer, the image $x \in \mathbb{R}^{H \times W \times C}$ is divided into a grid of patches, where H and W are the height and width of the image, and C is the number of channels. Each patch is flattened into a vector, and these patches are treated as tokens, similar to words in a sentence in NLP tasks. Let the size of each patch be $P \times P$, so the image is reshaped into $N = \frac{HW}{P^2}$ patches. Each patch is flattened into a vector of size $P^2 \cdot C$ and linearly projected into a vector of size D (the latent dimension of the transformer):

$$x_p = \text{Reshape}(x) \quad \in \mathbb{R}^{N \times (P^2 \cdot C)} \tag{2.24}$$

Each patch is then passed through a learnable linear projection to obtain the patch embeddings:

$$z_0 = [x_{\text{class}}; x_p E] + E_{\text{pos}} \quad E \in \mathbb{R}^{(P^2 \cdot C) \times D},\ E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{2.25}$$

Here, x_{class} is a learnable "classification token" prepended to the patch sequence, similar to the [CLS] token in BERT [DCLT19], and E_{pos} represents the positional embeddings, which encode the relative positions of the patches in the image.

2.4.4.2 Transformer Encoder

The sequence of patch embeddings is then fed into a standard transformer encoder. The transformer encoder consists of alternating layers of multi-head self-attention (MSA) and multi-layer perceptrons, with layer normalization (LN) and residual connections applied at each layer. Let L be the number of layers in the transformer:

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \quad \forall l = 1, \ldots, L \tag{2.26}$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l \tag{2.27}$$

The multi-head self-attention mechanism allows each patch to attend to every other patch, capturing global dependencies in the image. The MLP block consists of two fully connected layers with a GELU non-linearity in between:

$$\text{MLP}(x) = \text{FC}_2(\text{GELU}(\text{FC}_1(x))) \tag{2.28}$$

The MSA block operates by computing the self-attention mechanism across the sequence of patches. For each attention head i, we compute queries, keys, and values as follows:

$$Q_i = zW_i^Q \tag{2.29}$$
$$K_i = zW_i^K \tag{2.30}$$
$$V_i = zW_i^V \tag{2.31}$$

where W_i^Q, W_i^K, and W_i^V are the learned weight matrices for the queries, keys, and values, respectively. The attention weights are computed as:

$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \tag{2.32}$$

The output for each attention head is the weighted sum of values:

$$\text{head}_i = A_i V_i \tag{2.33}$$

The outputs of all attention heads are concatenated and passed through a linear transformation to obtain the final output of the multi-head self-attention block:

$$\text{MSA}(z) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.34}$$

where W^O is a learned projection matrix, and h is the number of attention heads.

2.4.4.3 Classification Head

After passing through all L transformer layers, the final representation of the [CLS] token is used as the image representation. A classification head, typically a multi-layer perceptron (MLP), is then applied to the output of the [CLS] token to generate the final classification prediction:

$$y = \text{MLP}(\text{LN}(z_{\text{class}}^L)) \tag{2.35}$$

2.5 Optical Microscopy

Optical microscopy plays an important role in biological research and is widely applied across various disciplines within the life sciences.

In cell biology, it has been instrumental in advancing our understanding of cellular structures and functions. From Hooke's first observation of cells in 1665 [HMW15] to the sophisticated live-cell imaging techniques of today, microscopy has continuously enhanced our knowledge of cellular organization and dynamics. In microbiology, microscopy has enabled significant progress in the study of microorganisms [AJW+08]. Its ability to visualize bacteria, viruses, and other microorganisms has led to important discoveries in disease mechanisms and ecological systems. Microscopy remains essential for both research and

clinical diagnostics in this field. Developmental biology relies heavily on microscopy for observing real-time processes during embryonic development and cell differentiation [Gil16]. These observations have provided key insights into complex biological phenomena, from early embryogenesis to the formation of organs. Neuroscience has also greatly benefited from advances in high-resolution microscopy [DSW90]. This technology allows for detailed examination of neuronal structures and functions, deepening our understanding of brain activity at both cellular and network levels. In pathology, microscopy is fundamental to disease diagnosis, particularly in the field of oncology [AFR05]. The ability to examine tissue and cellular changes is often critical for accurate diagnosis and effective treatment planning. Molecular biology has been revolutionized by fluorescence microscopy [Lak06], which enables tracking specific molecules within cells.

2.5.1 Brightfield Microscopy

Brightfield microscopy is the most basic and widely used form of optical microscopy, operating on a simple principle: light passes through a specimen, creating contrast based on the sample's capacity to absorb or scatter light. In a brightfield microscope, light from a source—typically a halogen lamp or LED—passes through a condenser lens that focuses it onto the specimen. The transmitted light then travels through the sample and is collected by an objective lens above it. This process produces an image where the specimen appears darker against a bright background, hence the term "brightfield." A standard brightfield microscope setup includes several essential components: a light source, a condenser lens for focusing light, a stage to hold and position the sample, an objective lens to capture transmitted light, and either an eyepiece or a camera for viewing the image. These parts and the microscope is visualized in figure 2.6. Brightfield microscopy offers several advantages, including its simplicity, ease of use, and versatility in examining a broad range of samples, as well as its relatively low cost compared to other techniques. These factors have led to its extensive use across various scientific fields. It is particularly effective for observing stained biological samples, examining microorganisms, studying cell structures, and analyzing minerals and metals.

However, brightfield microscopy has limitations. It often provides low contrast for transparent or unstained samples, making certain specimens challenging to observe. Resolution is limited by the wavelength of visible light, and distinguishing structures at different depths within a sample can be difficult. To address these challenges, researchers frequently use stains to increase contrast in biological samples or employ oil immersion to improve resolution at high magnifications.

2.5.2 Differential Interference Contrast

Differential Interference Contrast (DIC) microscopy, also known as Nomarski Interference Contrast, is an optical technique that enhances contrast in unstained, transparent specimens [L$^+$82]. DIC has transformed the visualization of live biological samples and materials with subtle optical thickness variations.

DIC operates on the principle of interferometry, where a polarized light beam is split into two slightly offset beams that travel through the specimen along different paths. These beams are then recombined, producing interference that converts phase differences into

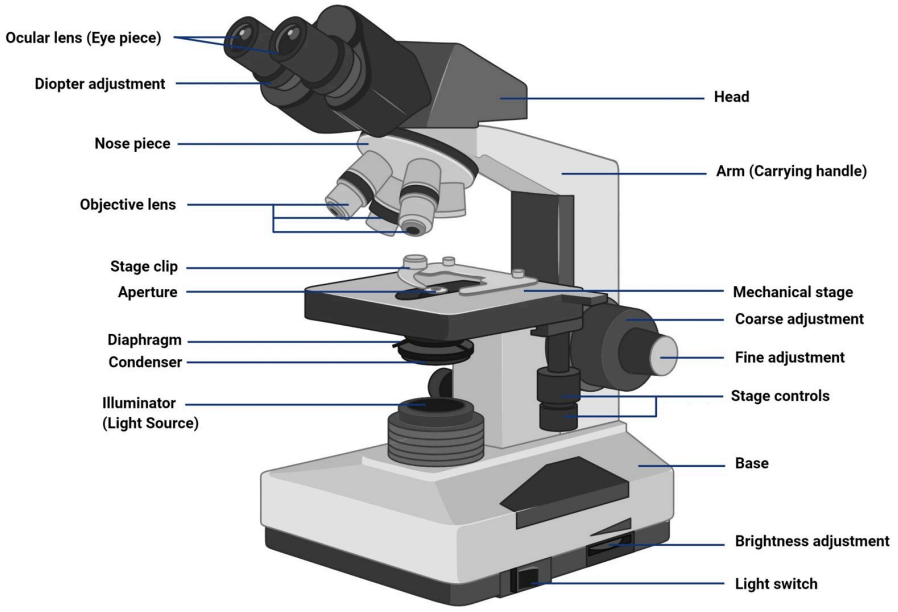


Figure 2.6: Visualization of a typical brightfield microscope [Mok24]

amplitude differences, resulting in a pseudo-3D image with high contrast. This enhanced imaging reveals details typically undetectable under brightfield microscopy. A typical DIC microscope includes several specialized components: a light source with a polarizer, a Nomarski prism for splitting the polarized light, the specimen, an objective lens, a second Nomarski prism to recombine the beams, and an analyzer positioned before the eyepiece or camera. DIC microscopy is particularly effective for live, unstained biological specimens, making it ideal for visualizing cellular structures, organelles, and the surface topography of various materials. It offers excellent resolution, optical sectioning capability, and high-contrast imaging without requiring stains, which has led to its widespread use in fields such as cell biology, embryology, and materials science. This technique enables visualization of features like cell edges, nuclei, and dynamic processes such as vesicle transport and cell division in live samples. In materials science, DIC effectively reveals surface structures and defects that may go undetected with other microscopy techniques.

However, DIC microscopy presents some challenges. It requires specialized, costly equipment and can be more complex to set up than brightfield microscopy, with optimal image quality depending on precise alignment of optical components. DIC images can be difficult to interpret, as the pseudo-3D effect does not directly represent true topography. Additionally, DIC is less effective with thick or highly absorbing specimens, and quantitative analysis can be challenging due to the non-linear relationship between optical path differences and image intensity. Despite these limitations, DIC remains a powerful method for high-contrast imaging of unstained, transparent specimens.

2.5.3 Phase Contrast Imaging

Phase Contrast Imaging (PCI) is an essential optical microscopy technique that has greatly advanced the study of live, unstained biological specimens [Zer42].

PCI works by transforming phase shifts in light passing through transparent samples into changes in light intensity, allowing otherwise invisible features to become visible. As light traverses a specimen, it undergoes phase shifts determined by the sample's thickness and refractive index. PCI utilizes these shifts to create contrast in the resulting image. A phase contrast microscope enhances a standard brightfield setup with two key components: a phase ring in the condenser and a phase plate in the objective. The phase ring generates a hollow cone of light to illuminate the specimen. Light passing directly through the specimen experiences a phase shift, while diffracted light does not. The phase plate then introduces an additional phase shift to the direct light. When these two light paths recombine, they interfere to produce intensity differences that create contrast in the image. PCI is especially valuable in cell biology, as it enables the study of live cells without staining, which can alter or harm cellular structures. It is frequently used to observe cell nuclei, organelles, and dynamic processes such as cell division and movement. One major advantage of PCI is its ability to provide high contrast for transparent specimens, which are nearly invisible under brightfield microscopy. It also supports live cell observation over extended periods with minimal preparation, typically requiring only that cells be placed in an appropriate medium on a microscope slide. A comparison between a brightfield image and a phase contrast image is shown in Figure 2.9, where the halo effect surrounding the cell edges is also clearly visible.

However, PCI has some drawbacks. The most notable is the presence of halo artifacts around specimen edges, which can obscure finer details. This halo effect is inherent to the PCI technique and, while it can be minimized, cannot be entirely eliminated. Additionally, PCI is best suited for thin, transparent specimens and is less effective for thicker samples. While PCI excels in qualitative imaging, it is less suited to quantitative analysis, as the relationship between specimen properties and image intensity is not linear, making precise measurements challenging.

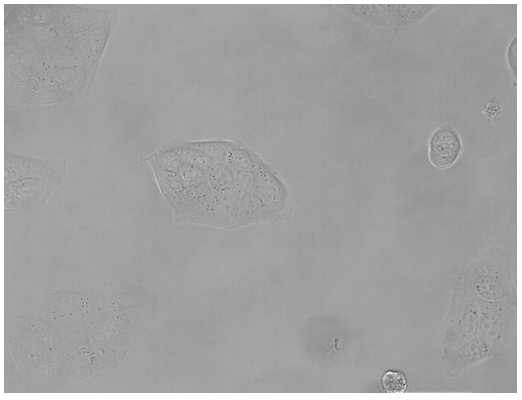


Figure 2.7: Brightfield Image



Figure 2.8: Phase Contrast Image

Figure 2.9: Comparison of the same sample under brightfield microsope and under a phase contrast microscope [Ock23]

2.5.4 Fluorescence Microscopy

Fluorescence microscopy allows researchers to visualize specific cellular structures and processes with high clarity and precision. This technique uses fluorescent molecules, or fluo-

rophores, to create high-contrast images of biological specimens. Fluorescence microscopy works by exciting fluorophores, which emit light when exposed to certain wavelengths. When a fluorophore absorbs light at its excitation wavelength, it reaches an excited state. As it returns to its ground state, it emits light at a longer wavelength, a process known as the Stokes shift, which is essential for producing fluorescence images.

A typical fluorescence microscope includes the following components:

- **Light Source**: Commonly a mercury or xenon lamp, LED, or laser.
- **Excitation Filter**: Selects the wavelength needed to excite the fluorophore.
- **Dichroic Mirror**: Reflects excitation light toward the specimen and transmits emitted light to the detector.
- **Emission Filter**: Isolates the light emitted by the fluorophore.
- **Objective Lens**: Focuses light on the specimen and collects the emitted fluorescence.
- **Detector**: Captures the fluorescent signal, typically using a camera or photomultiplier tube.

These components enable the selective visualization of fluorescently labeled structures against a dark background, resulting in high-contrast images. The basic principle is illustrated in Figure 2.10.

Fluorescent Markers for Cellular Structures

The versatility of fluorescence microscopy is largely due to the wide range of fluorescent markers available for labeling various cellular components. Here are some commonly used markers for key cellular structures that are employed in one of the datasets in this thesis:

Nucleus

DAPI [Kap95] is a widely used blue-fluorescent DNA stain that binds strongly to A-T rich regions in DNA. It permeates cell membranes, making it suitable for staining both live and fixed cells. Hoechst stains are another family of blue-fluorescent DNA stains with similar properties [LW75].

Mitochondria

MitoTracker dyes are commonly used for labeling mitochondria. These dyes passively diffuse across the plasma membrane and accumulate in active mitochondria. Variants include MitoTracker Green, which fluoresces in the lipid environment of mitochondria, and MitoTracker Red, which depends on mitochondrial membrane potential [PWP04].

Tubulin

Tubulin, a critical component of the cytoskeleton, can be visualized using fluorescently labeled antibodies against α- or β-tubulin in fixed cells. For live-cell imaging, fluorescent protein fusions (e.g., GFP-tubulin) or cell-permeable dyes like SiR-tubulin are used [KSV14].

Actin

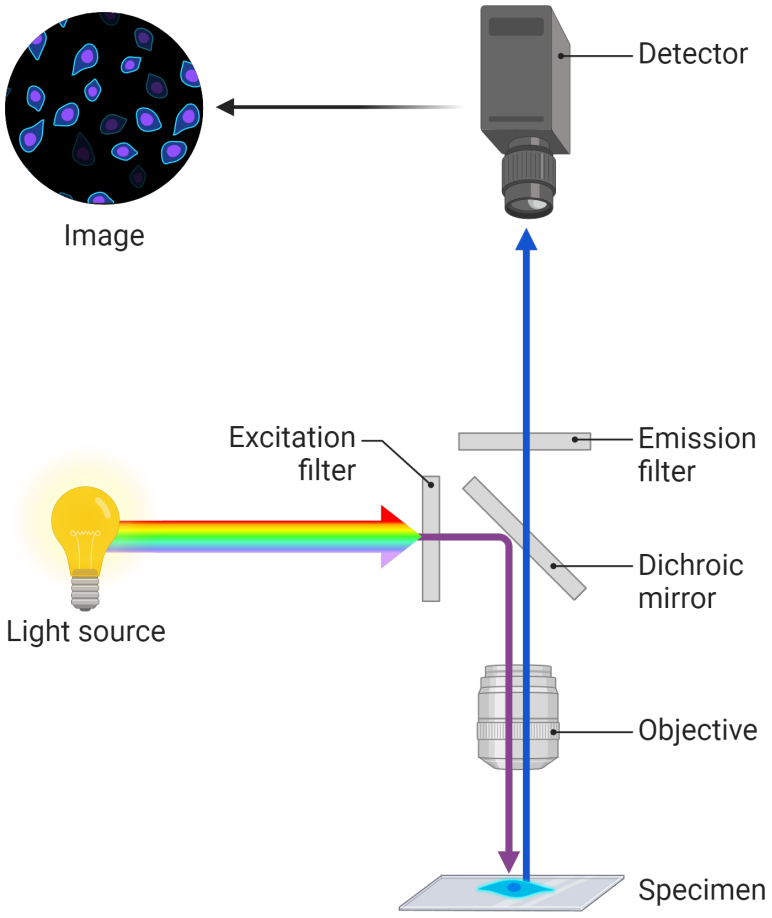


Figure 2.10: Visualization of a fluorescence microscope [Bio24]

Phalloidin, a toxin that binds to F-actin, is often conjugated to fluorescent dyes to visualize the actin cytoskeleton in fixed cells. For live-cell imaging, LifeAct fused to fluorescent proteins or SiR-actin can be employed [RCK+08, WDB+79].

Breast Cancer-Specific Markers

In addition to general cellular markers, specific stainings are important in breast cancer diagnosis and research:

- **HER2 (Human Epidermal Growth Factor Receptor 2)**: HER2 is a protein that promotes cell growth and is overexpressed in some breast cancers. Fluorescently labeled antibodies against HER2 help determine HER2 status, essential for treatment decisions. Positive HER2 staining appears as bright membrane fluorescence in cancer cells.
- **ER (Estrogen Receptor)**: ER is a nuclear receptor significant in hormone-responsive

breast cancers. Anti-ER antibodies conjugated to fluorophores are used to visualize ER expression, typically observed in the nuclei of cancer cells.

- **PR (Progesterone Receptor)**: Similar to ER, PR is a nuclear receptor important in hormone-responsive breast cancers. Fluorescently labeled anti-PR antibodies detect PR expression, with positive staining located in the nuclei of cancer cells.
- **Ki67**: Ki67 is a protein linked to cell proliferation, serving as a marker for rapidly dividing cells. Anti-Ki67 antibodies coupled with fluorophores assess the proliferation rate of cancer cells, with positive nuclear staining indicating the proliferation intensity and percentage of stained nuclei.

3. Task and Data Analysis

This master's thesis is structured around four key tasks, each examining unique aspects of image-to-image translation within the medical field. The complexity of these tasks stems from the variety of imaging modalities and the considerable differences among target images. Unlike standard image-to-image translation, which typically involves RGB color transformations, medical applications often work with grayscale images and require precise pixel-level accuracy. Consequently, specialized methods are essential for managing the diverse datasets commonly encountered in medical imaging.

This chapter introduces all datasets and its respective challenges.

3.1 Bright Field to Fluorescence

This dataset is the sole dataset provided for the competition and presents several challenges, including sparse outputs and limited data availability [FB24]. However, unlike other datasets used in this thesis, this data is pixel-perfect. The following sections introduce the biological and medical motivations, along with the technical challenges inherent to this dataset.

3.1.1 Biological and Medical Motivations

Fluorescence microscopy is a strong tool in biological and medical research, enabling the visualization of specific cellular components through the application of fluorescent markers. However, this technique necessitates manual biochemical labeling, which is both time-consuming and costly. Furthermore, the utilization of excitation light and fluorescent probes can perturb the cells under investigation, resulting in phototoxicity and photobleaching [WHT+23]. These effects constrain the feasibility of long-term imaging and can interfere with cellular processes due to their invasive nature.

In contrast, label-free transmitted light microscopy techniques such as bright-field, phase contrast, and Differential Interference Contrast are non-invasive, thereby reducing phototoxicity and preserving signal integrity. By facilitating the computational reconstruction of

fluorescence images from bright-field images, this research aims to bridge the gap between these two approaches. Reducing dependence on fluorescence microscopy addresses its associated limitations, providing a more efficient and less intrusive alternative for imaging.

3.1.2 Technical Motivations

The challenge in translating bright-field microscopy images to fluorescence images lies in the development of multi-output deep learning methods that operate on a single input. Specifically, the training data consists of images with missing channels and exhibits high variability across different acquisition conditions. This necessitates the creation of innovative architectures and loss functions tailored for sparse outputs. The ultimate goal is to provide biologists with a robust tool that performs reliably across various acquisition protocols, image sizes, cell lines, acquisition sites, modalities, and instruments.

To assess the generalizability of the proposed methods, data from one acquisition site will be excluded from the training dataset and reserved for final evaluation. This strategy tests the methods' ability to predict the optimal Z-focus [1] plane for any organelle, even under suboptimal acquisition conditions. Participants are encouraged to leverage data augmentation provided by the acquisitions, which includes large Z-stacks with predominantly out-of-focus planes.

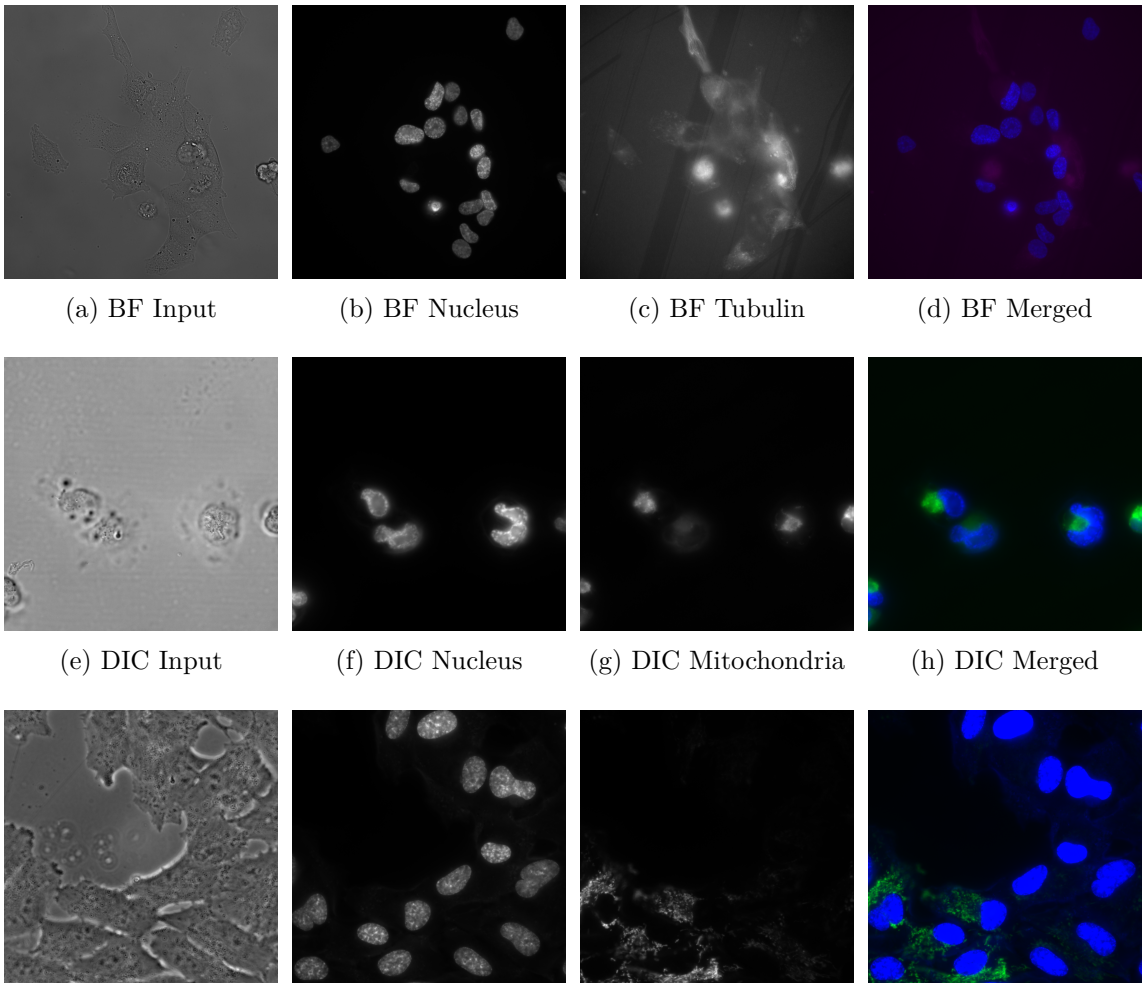
Performance will be evaluated using metrics that assess each of the four organelles and focus plane deviations and are introduced during the evaluation. Winners will be determined based on the best average performance across all metrics.

3.1.3 Dataset Description

The data for this task is sourced from the "Light My Cells: Bright Field to Fluorescence Imaging Challenge" organized by France-Bioimaging [FB24]. This challenge promotes the development of innovative deep-learning-based image-to-image translation methods for applications in biology and microscopy. Specifically, it focuses on predicting high-quality fluorescence images of various labeled organelles from label-free transmitted light images. The dataset presents extensive variability, including differences in Z-focus, multiple channels, acquisition sites, input modalities (such as Bright Field, Phase Contrast, and DIC), as well as variations in instruments, magnifications, cell types, and markers, all enabled by France-Bioimaging's robust infrastructure.

The main technical challenges include handling the significant variability in data, predicting multiple fluorescence channels from a single input, and managing sparse outputs where some channels may be absent in certain images. Additionally, ensuring the generalizability and robustness of models across various acquisition conditions is essential.

[1] A Z-Stack consists of multiple images that represent the depth dimension of an image



(a) BF Input (b) BF Nucleus (c) BF Tubulin (d) BF Merged

(e) DIC Input (f) DIC Nucleus (g) DIC Mitochondria (h) DIC Merged

(i) PC Input (j) PC Nucleus (k) PC Mitochondria (l) PC Merged

Figure 3.1: Comparison of BF, DIC, and PC imaging modalities. Each row shows the original image, nucleus, mitochondria/tubulin, and merged channels for each modality. [FB24]

3.2 Unstained to Stained

This dataset [ARA+24] consists of unstained-to-stained image data, making it one of the simpler tasks in this thesis, as all structures are theoretically visible in the unstained tissue, though at low contrast. This low contrast presents a unique challenge, requiring careful enhancement to reveal structural details accurately. The following sections discuss the motivations and specific challenges related to this dataset.

3.2.1 Biological and Medical Motivations

In histopathology, medical diagnosis relies heavily on stained tissue samples, with hematoxylin and eosin (H&E) staining being the most widely used technique, accounting for approximately 80% of all histological staining procedures worldwide. Staining enhances

cellular features, enabling pathologists to identify and differentiate pathological characteristics. However, traditional H&E staining has drawbacks: it is time-consuming, uses hazardous chemicals like formalin, ethanol, and xylene, and permanently alters the tissue sample, making further analysis impossible. These limitations have driven research into virtual staining approaches that could offer similar diagnostic insights while preserving the original tissue.

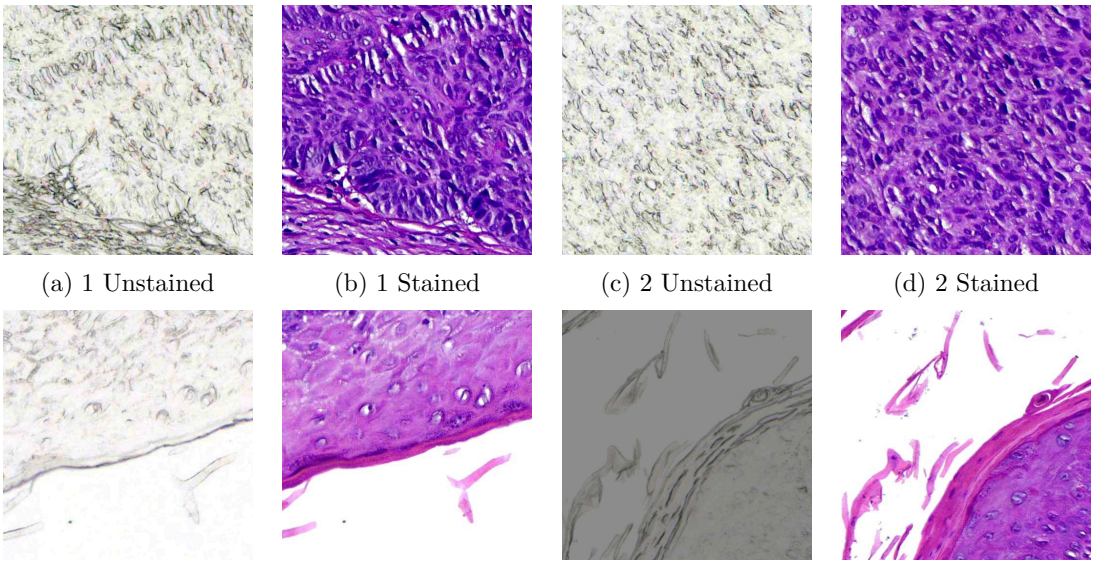
3.2.2 Technical Motivations

This dataset highlights several technical challenges common in virtual histological staining research. A major challenge is the physical deformation that tissue samples undergo during traditional staining, complicating the registration and alignment of unstained and stained image pairs. This inherent change in tissue morphology presents significant obstacles for supervised learning approaches, which rely on precise pixel alignment. Additionally, differences in staining protocols between laboratories lead to variations in appearance, making it challenging to establish a standardized ground truth. Addressing these technical constraints is essential for the development of robust computational pathology methods.

3.2.3 Dataset Description

The dataset [ARA$^+$24] consists of 56 pairs of unstained and H&E-stained skin tissue slides, which, when divided into individual samples, provide 92 matching tissue pairs. Images were captured using a standard brightfield microscope at 20× magnification. Each image was split into 512×512-pixel patches with a 256-pixel overlap, initially yielding over 40,000 paired patches. After a quality control process to remove patches with mostly background content, the final dataset contains 13,824 high-quality patches, with approximately 11,000 (80%) used for training and 2,824 (20%) reserved for testing.

Dataset preparation involved key preprocessing steps, including adaptive thresholding to isolate tissue areas and registration using Scale Invariant Feature Transformation (SIFT) to correct spatial misalignments between unstained and stained pairs. Traditional data augmentation was purposefully omitted to align with standard pathology practices, where tissue orientation is crucial for accurate diagnosis. All samples were collected with approval from the Advarra Institutional Review Board, following informed consent and adhering to relevant ethical guidelines and regulations.



(a) 1 Unstained (b) 1 Stained (c) 2 Unstained (d) 2 Stained

(e) 3 Unstained (f) 3 Stained (g) 4 Unstained (h) 4 Stained

Figure 3.2: Unstained and Stained Image Pairs [ARA+24]. Unstained images are the input and stained the desired output.

3.3 Breast Cancer Immunohistochemical Dataset

3.3.1 Biological and Medical Motivations

The diagnosis of breast cancer heavily relies on the microscopic analysis of tissue samples [BM21], where cellular structures and protein expressions are essential for guiding treatment decisions. While hematoxylin and eosin (H&E) staining serves as a foundational diagnostic method, evaluating human epidermal growth factor receptor 2 (HER2) expression through immunohistochemical (IHC) staining is critical for determining targeted therapies. HER2 expression is categorized into four levels (0, 1+, 2+, 3+), directly influencing treatment strategies and patient outcomes. However, the need for separate IHC staining adds extra cost and delays to the diagnostic process, potentially impacting the timely initiation of treatment.

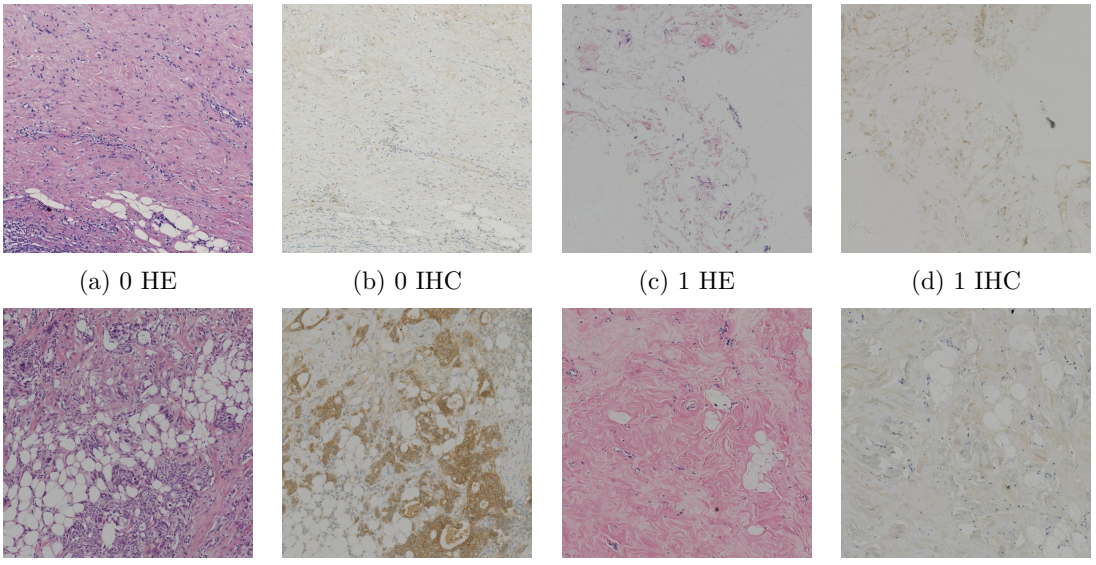
3.3.2 Technical Motivations

This dataset addresses key technical challenges in histopathological image analysis. A primary limitation is that identical tissue sections cannot undergo both H&E and IHC staining. As a result, consecutive tissue sections must be used, introducing structural differences between paired samples. These variations stem from two main sources: First, the sectioning process creates slight morphological differences between consecutive slices. Second, sample preparation steps, such as mounting and staining, can introduce deformations through stretching or compression. These challenges have traditionally hindered the development of methods for predicting IHC staining patterns directly from H&E images, making this dataset especially valuable for overcoming these limitations.

3.3.3 Dataset Description

The Breast Cancer Immunohistochemical (BCI) dataset [LZX+22] includes 4,870 paired image patches from 51 whole slide images (WSIs) of breast cancer tissue. These images were acquired with a Hamamatsu NanoZommer S60 scanner at a high resolution of 0.46 µm per pixel. Each image pair consists of an H&E-stained section and a corresponding IHC-stained section, specifically showing HER2 expression levels across all four categories: 240 pairs with IHC 0, 1,153 pairs with IHC 1+, 2,142 pairs with IHC 2+, and 1,335 pairs with IHC 3+.

To achieve optimal alignment between H&E and IHC images, a two-stage registration process was used. Initially, projection transformation was applied with manually selected corresponding points, followed by fine-tuned elastic registration using the elastix framework. Rigorous quality control ensured proper alignment and tissue integrity, resulting in a dataset of 1024×1024 pixel patches selected to exclude artifacts and poorly aligned regions, maintaining high data quality for computational analysis.



(a) 0 HE (b) 0 IHC (c) 1 HE (d) 1 IHC

(e) 2 HE (f) 2 IHC (g) 3 HE (h) 3 IHC

Figure 3.3: Comparison of HE and IHC Images for Cases 0-3 [BM21]. HE images are the input images and IHC the desired output images.

3.4 Multi IHC Stain Translation

3.4.1 Biological and Medical Motivations

Achieving pixel-perfect alignment between Hematoxylin and Eosin (H&E) and Immunohistochemical (IHC) stained images is often impractical due to the nature of tissue preparation and imaging processes. Nonetheless, the ability to translate H&E-stained images to IHC stains holds significant potential for breast cancer diagnosis, as detecting biomarkers like HER2, Ki67, ER, and PR is crucial for guiding treatment decisions. Developing robust techniques that accommodate misalignments and inconsistencies enhances the clinical applicability of virtual staining in diagnostic workflows.

3.4.2 Technical Motivations

The primary technical motivation in this task is to create models capable of accurately translating H&E-stained images to IHC stains, even without pixel-perfect alignment. Addressing issues arising from consecutive tissue slicing, morphological differences, imaging artifacts, and staining-induced degradation requires novel approaches. The Multi-IHC Stain Translation (MIST) dataset adds complexity by including multiple IHC stains, demanding models that can generalize across diverse biomarkers.

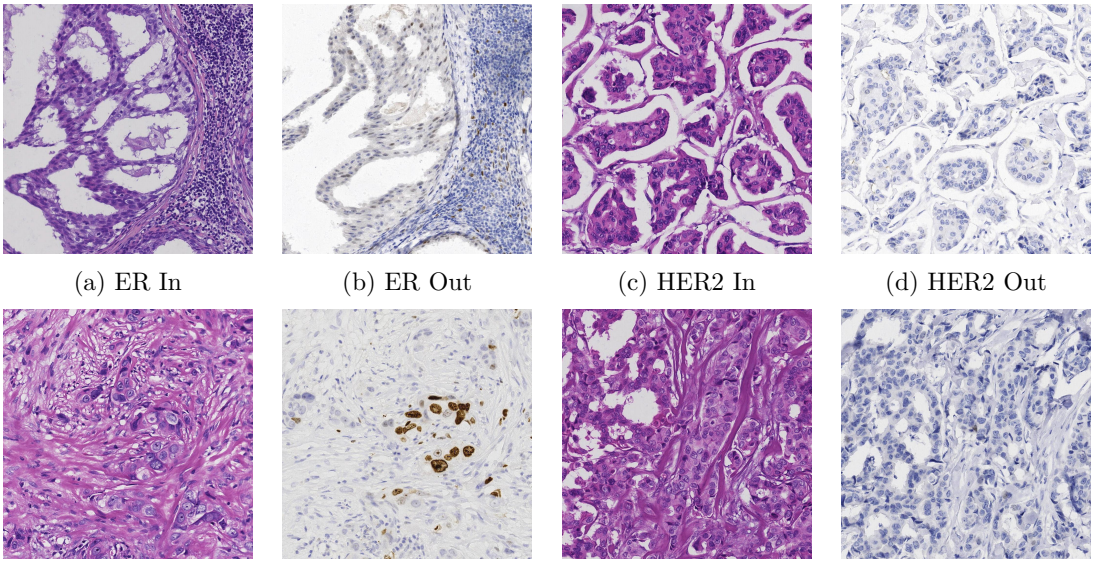
3.4.3 Dataset Description

The dataset used in this task, developed by Li et al. [LHCK23], is focused on translating H&E-stained images to IHC stains. A key challenge of this dataset is the inherent lack of pixel-perfect alignment, as each H&E-IHC image pair comes from consecutive tissue slices that have been individually stained and imaged. Misalignments are common due to variations in tissue morphology between slices, imaging artifacts, and potential staining-induced degradation like tissue tearing. These factors make direct pixel-to-pixel learning approaches more complex for H&E-to-IHC translation.

To address these challenges, Li et al. created the Multi-IHC Stain Translation (MIST) dataset [LHCK23], which includes H&E-IHC patches for four biomarkers critical in breast cancer diagnosis: HER2, Ki67, ER (Estrogen Receptor), and PR (Progesterone Receptor). This dataset is extensive and diverse, containing over 4,000 training patches and 1,000 testing patches for each stain, all extracted from whole-slide images (WSIs) of breast cancer tissue. Each patch is 1024×1024 pixels, providing high resolution and rich detail for the translation task.

Despite slice-to-slice inconsistencies, the patches retain crucial diagnostic information, allowing models to learn meaningful correlations between H&E morphological features and the molecular insights offered by IHC stains. This ability is especially relevant for breast cancer diagnosis, where identifying specific biomarkers such as HER2 or ER significantly informs treatment options.

The main technical challenges with this dataset involve managing inconsistencies between image pairs while enabling accurate translation from H&E to IHC stains. The dataset is designed to support the development of methods that can effectively learn from noisy and imperfectly aligned pairs, advancing robust image translation in the presence of variability.



(a) ER In (b) ER Out (c) HER2 In (d) HER2 Out

(e) Ki67 In (f) Ki67 Out (g) PR In (h) PR Out

Figure 3.4: Image pairs for ER, HER2, Ki67, and PR [LHCK23]. All "In" images are the input and the corresponding "Out" image is the desired output

4. Related Work

Digital staining has been an important topic for several decades, with numerous methods being introduced over time to improve the accuracy and efficiency of this task. Each of these methods has its own advantages and disadvantages, which serve as the foundation for the method developed in this work. Early approaches relied on simple statistical techniques to adjust the color of one staining to match another. However, with the advent of machine learning, and particularly deep learning, many new possibilities have come up to address this task more effectively. In the following sections, these methods will be categorized, and the most relevant approaches for this work will be presented and discussed.

4.1 Traditional Approaches

Early methodologies for image-to-image translation in medical imaging, such as stain normalization, predominantly utilized statistical techniques including histogram matching and color deconvolution [RAGS01]. These techniques were effective in standardizing color and intensity variations across images but were constrained by their inability to capture the complex spatial relationships inherent in biological tissues.

4.1.1 Histogram transformation

Histogram transformation methods play a significant role in standardizing color appearances across histopathological images by aligning them to a reference RGB color space. This reference color space is typically created from a set of images that represent a specific staining protocol or lab setting, ensuring that the resulting images are comparable in color distribution. Reinhard et al. [RAGS01] developed a notable method that facilitates color transfer by initially converting images to the $\ell\alpha\beta$ color space. This conversion allows for effective alignment of background colors between the source and target images, followed by a re-conversion to RGB space to achieve color normalization. Zheng et al. [ZJZ+18] suggested another approach that operates in HSV space, where normalization is achieved by finely adjusting global illumination and saturation, enhancing color consistency across samples.

Expanding upon Reinhard et.al [RAGS01], Wang et al. [WCW+07] introduced a linear transformation technique in the $\ell\alpha\beta$ color space to standardize color for applications in tumor segmentation and classification. This method uses a template image to guide color adjustments, ensuring consistency for automated analysis tasks. Similarly, Janowczyk et al. [JBM17] leveraged sparse AutoEncoders to cluster pixel data, with each cluster undergoing histogram alignment to match a reference color distribution, thereby improving the color standardization process.

In a further enhancement, Roy et al. [RLK19] refined Reinhard's transformation using fuzzy logic to better control color contrast and reduce variability, effectively addressing limitations in the conventional approach. This modified method supports a more consistent global color distribution by managing contrast and adjusting color channels in the ℓ and $\alpha\beta$ spaces. Vijh et al. [VSK21] presented an innovative approach that combines fuzzy logic with an advanced illuminant normalization technique, which corrects for variations in lighting conditions and uses modified spectral normalization. This method preserves the structural and histological details of the source image, enabling reliable stain normalization.

4.1.2 Classical Machine Learning Techniques

Beyond statistical methods, classical machine learning techniques have been employed for medical image translation tasks. Approaches such as structured random forests and k-nearest neighbors regression have been utilized to translate MRI images into CT images for PET attenuation correction [HGK+15, ZLL+16]. These methods function by learning mappings between modalities based on local image patches, using handcrafted features to represent tissue structures and textures. Although these techniques have demonstrated a degree of success, their reliance on manual feature engineering restricts their capacity to model the complex and high-dimensional nature of biological tissues. This limitation is particularly pronounced when dealing with the subtle variations present in microscopy images, where fine details are crucial for accurate interpretation and analysis.

4.1.3 Color Deconvolution

Whole slide images are stained in varied, independent processes, prompting the development of techniques that apply multiple transformations to achieve color consistency across images [KRTM14, BBL+16]. The complexity of these techniques lies in accurately classifying and isolating the contributions of each stain to the overall color composition. Color deconvolution is commonly used for this purpose, serving as a matrix decomposition method that separates light-absorbing stains and estimates color proportions in individual pixels by leveraging the spectral characteristics of each stain [RJ+01]. This approach follows a calibration-based method rooted in the Beer–Lambert law, which describes the relationship between light absorption and the concentration of an absorbing substance [Swi62].

Magee et al. [MTC+09] introduced a refined technique for multi-modal color deconvolution, incorporating linear normalization in the $\ell\alpha\beta$ color space along with stain-specific color adjustments. This method leverages a probabilistic prior and a Variational Bayesian Gaussian Mixture Model to achieve per-pixel stain transformations, addressing the limitations of linear normalization. Similarly, Khan et al. [KRTM14] proposed a method

based on nonlinear mapping, where pixels belonging to the same stain are grouped, and a staining vector is calculated by analyzing pixel distributions. This approach allows for a non-linear mapping that accurately transfers the color characteristics from different pixel types to a template image, ensuring uniformity.

Bejnordi et al. [BLT+15] introduced a hue-saturation-density (HSD) based approach for color standardization, transforming the RGB color space into chromatic and density components. This technique aligns the chromatic and density distributions of components like hematoxylin, eosin, and background in HSD space, with weighted adjustments applied to each stain before reconverting to RGB. This process ensures color standardization across histopathology images, enhancing consistency and reliability for subsequent analysis.

4.2 Deep Learning-Based Approaches

Deep learning has brought transformative advances across various fields, particularly in medicine. By learning from data, these methods can assist in disease detection, support physicians in diagnosis, and even enable autonomous diagnostic capabilities. In the domain of image-to-image translation, numerous methods and architectures have been developed over recent years, offering potential solutions for this task. This section introduces established methods in image-to-image translation as well as related architectures used in adjacent tasks, such as image generation, to provide a comprehensive overview of both current and emerging methodologies.

4.2.1 Autoencoders

Autoencoders function by encoding an input image into a latent space and subsequently reconstructing this latent space using a decoder. While they are less commonly employed than generative adversarial network-based approaches, they hold significant value, as demonstrated in this work. Numerous autoencoder architectures have been developed over the years. The simplest model is a U-Net-based autoencoder, which encodes the input image into a latent space using a series of convolutions and pooling operations, and then decodes it with transposed convolutions.

Lee et al. [LOK+18] introduced two autoencoders: the first generates a fluorescence channel from a brightfield image, while the second refines this output by combining it with the original brightfield image. In a subsequent work, Lee et al. apply a U-Net to encode a brightfield image into a latent space, using four different decoders to predict multiple organelles from this shared latent representation [LOHJ21].

4.2.2 Generative Adversarial Networks

Generative adversarial networks have been widely utilized for tasks in image generation and image-to-image translation [GPAM+20]. Unlike traditional autoencoders that rely primarily on reconstruction losses, GANs incorporate an additional discriminator network, which classifies patches of an input image as either real or generated, helping to enhance the realism of generated images. This architecture introduces two distinct loss functions: a generator loss and a discriminator loss. Training GANs presents a unique challenge, as both networks need to be carefully balanced to avoid instability, where one network overpowers the other, potentially leading to training collapse. Isola et al. [IZZE17] introduced

one of the earliest GAN-based architectures for image translation, Pix2Pix, which leverages a U-Net-based generator to convert input images to their target domain. To improve output quality, Pix2Pix utilizes a patch-based discriminator, which has proven effective for various tasks, including translating brightfield microscopy images to fluorescence images [CZMW+22]. In histopathology, Pix2Pix has also been successfully applied to normalize stain variations by translating between different stains [SC20]. However, Pix2Pix requires pixel-perfect paired images, which are often difficult to obtain in medical imaging due to the complex alignment required between modalities. To address this limitation, Zhu et al. [ZPIE17] proposed CycleGAN, an adaptation that enables image translation with unpaired datasets. This approach is especially valuable in medical contexts where only stain normalization is required, without introducing new structural details [SC20]. CycleGAN has broadened the applicability of GANs by making translation feasible for cases where paired data is challenging to acquire.

4.2.3 VQGAN

Vector Quantized Generative Adversarial Networks (VQGAN) [ERO21] are a type of autoencoder that uses quantization to improve output quality. Similar to a standard UNet-based autoencoder, VQGANs use an encoder to transform an image into its latent space. However, instead of decoding this latent space directly, VQGANs quantize it into discrete tokens using a learned codebook. A transformer then generates an enhanced latent representation by mapping these tokens to a new latent space, which is fed into a decoder to reconstruct the final image. Additionally, a discriminator is incorporated to further refine the quality of the generated images.

In the context of brightfield-to-fluorescence image-to-image translation, this method has been applied to generate multiple organelle types from a single brightfield input image, demonstrating its utility in detailed cellular structure prediction [LL24].

4.2.4 Diffusion Based Image to Image translation

In recent years, generative adversarial networks (GANs) have largely been supplanted by diffusion-based models for image generation and conditional image translation [HJA20, SME20, RBL+22]. Diffusion models provide several advantages over GANs, with one of the most significant being their simpler training process. Unlike GANs, which involve two networks with competing objectives, diffusion models typically operate with a single objective function, streamlining their training. The initial diffusion models, such as those described by Ho et al. [HJA20], performed pixel-based diffusion directly in the image space, making them resource-intensive due to the high computational demands. To address this, Rombach et al. [RBL+22] introduced latent diffusion, a technique that utilizes a variational autoencoder to encode the image and perform the diffusion process within a lower-dimensional latent space. This reduction in dimensionality significantly reduces computational requirements, making latent diffusion particularly useful for high-dimensional data, such as medical images, which this thesis also focuses on. Beyond image generation, diffusion models have shown success in other generative domains, such as music generation [HPW+23]. Recently, Peebles et al. [PX23] introduced Diffusion Transformers, which use a transformer architecture for the denoising process instead of the conventional U-Net.

As seen in large language models [DCLT19, RWC+19], transformers offer a scalable architecture, which also applies to diffusion models. However, due to the limited availability of medical data, transformer-based diffusion models were not explored in this thesis but represent an interesting direction for future research. However, for image-to-image translation, research utilizing diffusion-based models remains limited compared to generative adversarial networks [ZFT+24]. This thesis aims to address this gap by exploring diffusion models as an alternative approach for translation tasks.

5. Methods

The methods section is divided into two categories. Since the thesis involves participation in a competition, but the competition deadline was shortly after the thesis registration, several models were developed during the course of this work. These models, along with their respective preprocessing steps, differ between the competition dataset and the other datasets. Therefore, they are presented separately to account for these differences.

5.1 Competition Models

This section presents all the deep learning architectures, preprocessing steps, and loss functions used specifically for the competition.

5.1.1 Preprocessing

Since the dataset provides microscopy images at different z-levels, but the final validation only uses the best z-level, the same calculations for selecting the optimal z-level were performed in advance on the dataset. This approach offers several benefits, such as reducing computational costs by shrinking the dataset size and potentially decreasing overfitting. By excluding out-of-focus z-levels, the model is prevented from fitting to noisy data, ensuring it learns from clearer, more focused inputs. The algorithm for selecting the best focus image from a Z-stack [1], inspired by [SDN04], works by calculating the normalized variance of each image in the stack. The image with the highest normalized variance is selected as the best focus image. The normalized variance measures image sharpness by capturing the spread of pixel intensities, with higher variance indicating sharper, more focused images. Normalizing by the mean ensures the focus score is not affected by brightness differences. The steps of the algorithm are outlined below:

[1] A Z-Stack consists of multiple images that represent the depth dimension of an image

Algorithm 4 Selecting the Best Focus Image from a Z-stack

Input: Set of images in a Z-stack $\{I_1, I_2, ..., I_n\}$
Output: Image I_{best} with the best focus
Initialize $best_focus_score \leftarrow -1$
Initialize $I_{best} \leftarrow$ None
for each image I_k in the Z-stack **do**
    Convert I_k to grayscale: $I_k^{gray} = \text{convert_to_grayscale}(I_k)$
    Compute the variance of the grayscale image:

$$\sigma_k^2 = \frac{1}{M} \sum_{i=1}^{M} (I_k^{gray}(i) - \mu_k)^2$$

    where M is the total number of pixels in the image and μ_k is the mean pixel value of I_k^{gray}:

$$\mu_k = \frac{1}{M} \sum_{i=1}^{M} I_k^{gray}(i)$$

    Compute the normalized variance for image I_k:

$$\text{normalized_variance}(I_k) = \frac{\sigma_k^2}{\mu_k^2} \quad \text{if} \quad \mu_k \neq 0$$

    otherwise, set the normalized variance to 0.
    if the normalized variance of I_k is greater than $best_focus_score$ **then**
        Update $best_focus_score \leftarrow \text{normalized_variance}(I_k)$
        Update $I_{best} \leftarrow I_k$
    end if
end for
Return: I_{best}

This generates exactly one input-output pair for each microscopy image. Each individual image is then preprocessed by first scaling it to the range $[0, 1]$ by dividing the 16-bit grayscale image by 65535. After this, a percentile normalization is applied, where pixel values are adjusted based on the lower and upper percentiles (e.g., 2nd and 99.8th percentiles) of the image's intensity distribution. This normalization rescales the pixel values so that outliers (extremely low or high values) are suppressed, enhancing the contrast. This method is used to align with the competition's evaluation protocol.

5.1.2 Autoencoder

The architecture of the model is based on a UNet [RFB15], which has proven to be highly effective in medical image analysis due to its ability to capture both high-level and fine-grained spatial information through its symmetrical encoder-decoder structure. UNet architectures, as demonstrated by [IJK+21], are known for their strong performance across a wide range of medical imaging tasks, making them a reliable choice for this thesis.

The model uses an EfficientNet-B7 [TL05] encoder as its backbone for feature extraction. EfficientNet-B7 was chosen because it offers a good trade-off between performance and computational cost. It is capable of extracting detailed features while keeping resource usage reasonable.

Since the input images are normalized to lie within the range $[0, 1]$, the model uses a sigmoid activation function in the final layer to ensure that the output values also fall within this range. This is crucial for the model's tasks, such as image reconstruction, where the output must maintain consistency with the normalized input. The combination of the U-Net architecture with an EfficientNet-B7 encoder provides a strong foundation for tackling the image-to-image translation tasks central to this thesis. Two variations of this model are introduced, called **Autoencoder (Single Channel)** and **Autoencoder (All Channel)**, representing whether the model has one or four output channels.

5.1.3 Generative Adversial Neural Network

The Generative Adversarial Neural Network used in this competition is based on the Pix2Pix model [IZZE17]. Similar to the autoencoder described in 5.1.2, the Pix2Pix model utilizes an EfficientNet-B7 [TL05] encoder for the Generator. The model incorporates a patch-based discriminator with four convolutional downsampling layers, maintaining the structure from the original implementation [IZZE17]. The model's training process involves a composite weighted loss function that combines reconstruction and adversarial losses, as reflected in the code.

The total generator loss $\mathcal{L}_G$ can be formulated as:

$$\mathcal{L}_G = \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{pixel}}\mathcal{L}_{\text{L1}} + \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}},$$

where $\mathcal{L}_{\text{GAN}}$ represents the adversarial loss [IZZE17] computed using binary cross-entropy, $\mathcal{L}_{\text{L1}}$ is the L1 pixelwise reconstruction loss, $\mathcal{L}_{\text{MSE}}$ denotes the Mean Squared Error (MSE) loss, and $\mathcal{L}_{\text{SSIM}}$ refers to the Structural Similarity Index loss. The weights $\lambda_{\text{GAN}}, \lambda_{\text{pixel}}, \lambda_{\text{MSE}}, \lambda_{\text{SSIM}}$ are set to balance the contributions of each loss term. Specifically, these weights are chosen as $\lambda_{\text{GAN}} = 1$, $\lambda_{\text{pixel}} = 100$, $\lambda_{\text{MSE}} = 10$, and $\lambda_{\text{SSIM}} = 0.5$.

The discriminator loss $\mathcal{L}_D$ is defined to enhance the model's capability to differentiate between real and generated images and is expressed as:

$$\mathcal{L}_D = 0.5 \times (\mathcal{L}_{\text{real}} + \mathcal{L}_{\text{fake}}),$$

where $\mathcal{L}_{\text{real}}$ is the loss computed when the discriminator correctly classifies real input-target image pairs, and $\mathcal{L}_{\text{fake}}$ is the loss when it accurately identifies generated images as fake. Both $\mathcal{L}_{\text{real}}$ and $\mathcal{L}_{\text{fake}}$ are calculated using binary cross-entropy loss.

Overall, the final loss incorporates both the generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_G + \mathcal{L}_D.$$

Two variations of this model are introduced, called **Pix2Pix (Single Channel)** and **Pix2Pix (All Channels)**, representing whether the model has one or four output channels.

5.1.4 Losses

he loss functions used for the Autoencoder as well as the Generative Adversial Neural Network are described here. The losses consist of multiple components, each contributing to improving the image reconstruction quality by balancing pixel-wise accuracy and structural similarity.

Mean Squared Error Loss: The first loss function employed is the Mean Squared Error (MSE) loss, which measures the average squared difference between the predicted image I_g and the target image I_t. MSE is defined as:

$$L_{MSE}(I_g, I_t) = \frac{1}{N} \sum_{i=1}^{N} \left(I_g^{(i)} - I_t^{(i)} \right)^2$$

where N is the total number of pixels, and $I_g^{(i)}$ and $I_t^{(i)}$ are the pixel values of the generated and target images, respectively. This loss penalizes large differences between the images, encouraging pixel-wise accuracy. However, MSE is known to result in blurry reconstructions, especially in image generation tasks [MMI+22].

L1 Loss: The second pixel-wise loss used is the L1 loss, which measures the absolute differences between the predicted and target images. It is defined as:

$$L_{L1}(I_g, I_t) = \frac{1}{N} \sum_{i=1}^{N} \left| I_g^{(i)} - I_t^{(i)} \right|$$

L1 loss tends to produce sharper images compared to MSE loss because it penalizes outliers less severely. However, when used alone, it can still result in suboptimal visual quality, such as lack of fine texture details.

Structural Similarity Index Loss: To address the issue of blurry images produced by pixel-wise losses, the Structural Similarity Index (SSIM) is incorporated as a patch-wise loss function. SSIM provides a gradient signal that enhances image reconstruction quality by focusing on local patterns of pixel intensities.

SSIM compares the local structure of two images by using a sliding window of size w over the images. The SSIM between two images I_g and I_t is computed as:

$$SSIM(I_g, I_t) = \frac{(2\mu_g\mu_t + C_1)(2\sigma_{gt} + C_2)}{(\mu_g^2 + \mu_t^2 + C_1)(\sigma_g^2 + \sigma_t^2 + C_2)}$$

where μ_g and μ_t are the local means of the generated and target images, σ_g^2 and σ_t^2 are the local variances, and σ_{gt} is the covariance. C_1 and C_2 are constants to stabilize the division, typically set to $C_1 = (0.01)^2$ and $C_2 = (0.03)^2$. The SSIM loss is defined as:

$$L_{SSIM}(I_g, I_t) = 1 - SSIM(I_g, I_t)$$

This loss provides a patch-wise comparison of images, focusing on the structural similarity between the generated and target images, improving the perceptual quality of the reconstructions.

5.1.5 Dealing with missing organelles

The dataset has one major drawback: for each input image, only a few organelles are available. This complicates training, as typical supervised learning requires input-output pairs with all possible output images provided for each datapoint. To address this, only one organelle is used as the ground truth image for each training sample, allowing the model to maintain a shared embedding space in the encoder and decoder. The model, which generates a 4-channel output, is trained by selecting the corresponding output channel based on the provided organelle, and the loss is calculated only on that channel. This approach enables the model to effectively share both the encoder and decoder, instead of using separate decoders for each output channel, as done in other works [LOHJ21].

5.2 Hybrid Swin Conditional Refinement Latent Diffusion Model

This section introduces the method developed in this thesis, referred to as the Hybrid Swin Conditional Refinement Latent Diffusion Model. The name reflects the model's approach, which combines both direct and external conditioning, creating a hybrid conditioning framework. This conditioning is applied within the latent space of the encoded image, utilizing the same variational autoencoder for encoding and decoding the noisy images.

5.2.1 Preprocessing

Since latent diffusion models operate through variational autoencoders, requiring inputs normalized to the range $[-1, 1]$, the input images need to undergo normalization. For the brightfield-to-fluorescence task, however, an issue arises as the images are 16-bit grayscale, often utilizing only a small portion of the available intensity range. Normalizing these images to $[0, 1]$ directly would yield very low values across most pixels.

To address this, intensity rescaling is applied, which stretches or compresses image intensities to fit a specified range. This function maps the input image's intensity values from a specified range to a new range, here $[0, 1]$, preserving relative intensity differences and enhancing visibility.

In line with the competition's preprocessing steps, percentile normalization to the 2nd and 99.8th percentiles is applied. Finally, the grayscale image is converted to RGB format to match the pretrained variational autoencoder's RGB input requirement.

5.2.2 Variational Autoencoder Architecture

The Autoencoder is based on a U-Net [RFB15] architecture to generate a latent representation of the input image. The architecture consists of an encoder-decoder, as shown in 5.1.

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ be the input image with spatial resolution $H \times W$ and C channels. The goal is to map the image to a latent representation $\mathbf{z} \in \mathbb{R}^{H' \times W' \times D}$, where $H' \times W'$ is the downsampled resolution, and D is the number of latent channels.

5.2.2.1 Encoder

The encoder $E(\mathbf{x})$ maps the input image $\mathbf{x}$ to a latent representation $\mathbf{z}$. It performs a series of downsampling operations followed by residual blocks to extract features. Mathematically, the encoder can be expressed as:

$$\mathbf{z} = E(\mathbf{x}) = f_{\text{conv}}(f_{\text{res}}(\mathbf{x})),$$

where f_{conv} represents a sequence of convolutional layers, and f_{res} denotes the application of residual blocks with skip connections. The downsampling is applied at multiple stages to reduce the spatial resolution progressively, while maintaining the feature complexity using channels defined by the `ch_mult` parameter, resulting in $\mathbf{z} \in \mathbb{R}^{H' \times W' \times 2D}$, where $2D$ accounts for the double latent channels due to the `double_z` setting.

At the final stage, the encoder applies a convolutional layer to transform the output to $\mathbf{z} \in \mathbb{R}^{H' \times W' \times 2z_channels}$. This is split into mean and variance terms for the latent Gaussian distribution in the case of variational autoencoders, leading to the posterior distribution $q(\mathbf{z}|\mathbf{x})$. This results in a stochastic latent encoding:

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu, \sigma^2),$$

where μ and σ are predicted by the encoder.

5.2.2.2 Latent Space Representation

The latent vector $\mathbf{z}$ is reparameterized during training using the reparameterization trick [KSW15] to allow backpropagation through the stochastic node:

$$\mathbf{z} = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

After reparameterization, a learned convolutional layer is applied to transform $2z_channels$ to $2 \times$ `embed_dim`.

5.2.2.3 Decoder

The decoder $D(\mathbf{z})$ reconstructs the original input from the latent representation. It performs a series of upsampling operations, increasing the spatial resolution back to $H \times W$, while decreasing the number of channels. The decoder architecture can be expressed as:

$$\hat{\mathbf{x}} = D(\mathbf{z}) = f_{\text{upsample}}(f_{\text{res}}(\mathbf{z})),$$

where f_{upsample} denotes the upsampling operation, and f_{res} represents the residual blocks applied at each stage. The final layer outputs the reconstructed image $\hat{\mathbf{x}} \in \mathbb{R}^{H \times W \times C}$, using a convolutional layer followed by an optional tanh activation function.

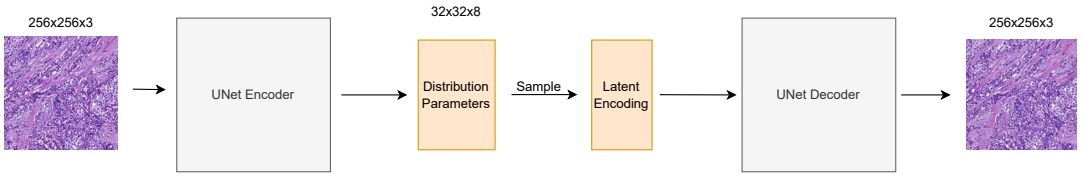


Figure 5.1: Variational Autoencoder for the Latent Diffusion model. The encoder produces the mean and the variance of a gaussian distribution from which a latent encoding is sampled. This sample is then reconstructed through the decoder.

5.2.2.4 Loss

The autoencoder loss function combines serveral different loss functions: pixel-wise loss, perceptual loss, KL divergence, and adversarial loss via a discriminator. These components work together to balance reconstruction accuracy, regularization, and adversarial learning. The total loss function $\mathcal{L}$ consists of three main parts: the reconstruction loss $\mathcal{L}_{\text{rec}}$, the Kullback-Leibler divergence (KL loss) [KL51] $\mathcal{L}_{\text{KL}}$, and the adversarial loss $\mathcal{L}_{\text{GAN}}$, weighted by an adaptive discriminator weight d_{weight}.

The reconstruction loss is a combination of pixel-wise loss and perceptual loss. The pixel-wise loss is computed as the absolute difference between the input x and the reconstruction $\hat{x}$:

$$\mathcal{L}_{\text{pixel}}(x, \hat{x}) = \|x - \hat{x}\|_1.$$

In addition, the perceptual loss, computed using a pre-trained perceptual similarity model (LPIPS) [ZIE+18], enhances the perceptual quality of the reconstructions. The total reconstruction loss is given by:

$$\mathcal{L}_{\text{rec}}(x, \hat{x}) = \mathcal{L}_{\text{pixel}}(x, \hat{x}) + \lambda_{\text{perc}} \cdot \mathcal{L}_{\text{perceptual}}(x, \hat{x}),$$

where λ_{perc} is a weighting factor that controls the influence of the perceptual loss.

The KL divergence loss measures how closely the posterior distribution $q(z|x)$ from the variational autoencoder approximates the prior distribution $p(z)$, typically a Gaussian. This is expressed as:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q(z|x) \parallel p(z)).$$

The KL loss encourages the posterior to match the prior, acting as a regularizer that helps prevent overfitting.

The adversarial loss encourages the generator to produce realistic reconstructions by using a discriminator D. The generator loss is based on the discriminator's predictions for the fake images $\hat{x}$ and is expressed as:

$$\mathcal{L}_{\text{GAN}} = -\mathbb{E}_{\hat{x}}[D(\hat{x})],$$

where $D(\hat{x})$ represents the discriminator's confidence that $\hat{x}$ is a real image. The discriminator itself is trained using a standard adversarial loss, which in this case can be either a hinge loss or a vanilla loss, depending on the configuration:

$$\mathcal{L}_{\text{disc}} = \max(0, 1 - D(x)) + \max(0, 1 + D(\hat{x})),$$

when using the hinge loss. This ensures that the discriminator correctly distinguishes between real and fake images.

Finally, the total generator loss combines the reconstruction loss, the KL loss, and the adversarial loss. It is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} + d_{\text{weight}} \cdot \lambda_{\text{disc}} \cdot \mathcal{L}_{\text{GAN}},$$

where λ_{KL} and λ_{disc} are weighting factors for the KL divergence and adversarial losses, respectively. The adaptive discriminator weight d_{weight} adjusts based on the gradient magnitudes of the reconstruction and adversarial losses, ensuring a balance between these objectives during training.

The total discriminator loss for its own update step is:

$$\mathcal{L}_{\text{disc_total}} = \lambda_{\text{disc}} \cdot \mathcal{L}_{\text{disc}}.$$

This completes the loss formulation, combining pixel-level accuracy, perceptual quality, regularization via KL divergence, and adversarial realism into a single comprehensive loss function for training generative models.

5.2.3 Conditional Module

The conditional module in this work is built upon the Swin Transformer [LLC+21] architecture, which extracts features at multiple hierarchical levels, rather than applying self-attention across a single spatial dimension. This multi-scale feature extraction allows the model to capture finer details and aligns more closely with the spatial structure of a UNet. The conditioning module receives the encoded output from the variational autoencoder, effectively reducing the spatial size of all inputs to $width/8$ and $height/8$. For this thesis, where input images are resized to 256x256, the input to the conditioning module becomes 32x32x4, with 4 representing the latent dimension of the VAE.

Swin Transformer is configured with a patch size of 2, instead of the common patch size of 4, to allow the model to extract even finer features, which are used as conditioning in the latent diffusion module. For this approach to function effectively, the variational autoencoder must be well-pretrained to create a meaningful latent space that the conditioning module can refine further.

A patch size smaller than 4 is typically not feasible due to hardware limitations if the input is too large. However, since this method operates in the latent space, where the spatial dimensions are significantly reduced, applying self-attention at a finer granularity—almost at the level of every token—becomes computationally feasible and is exploited in this method.

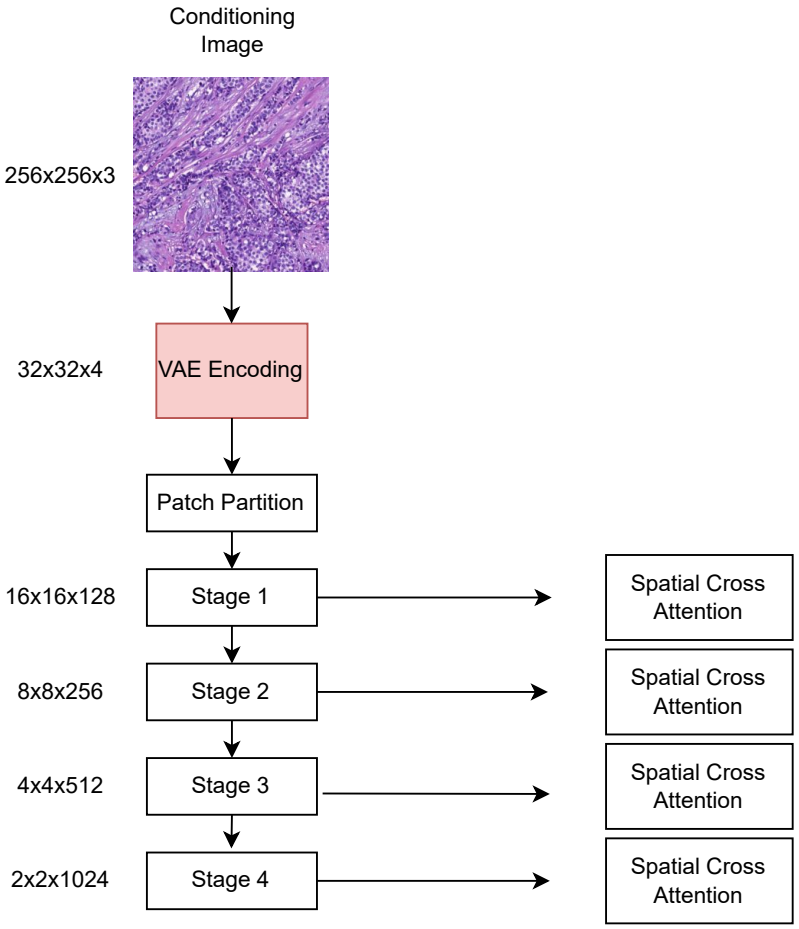


Figure 5.2: Conditioning Module of the method. The input to the Swin transformer is the encoded image after being put through the variational autoencoder. The different stages are then used as context for the Spatial Cross Attention.

5.2.4 Spatial Cross-Attention

Spatial cross-attention is an important component of the latent diffusion model, particularly in enabling the model to effectively attend to conditioning features during the denoising process. Mathematically, the cross-attention mechanism can be described as follows:

Let $X \in \mathbb{R}^{B \times C \times H \times W}$ be the input tensor representing the latent feature map, where B is the batch size, C is the number of channels, and $H \times W$ are the spatial dimensions. Similarly, let $C_f \in \mathbb{R}^{B \times C_c \times H_c \times W_c}$ represent the conditioning feature map obtained from the Swin Transformer, where C_c is the number of conditioning channels and $H_c \times W_c$ are the spatial dimensions of the conditioning features.

To compute spatial cross-attention, both the input latent feature map X and the conditioning feature map C_f are first reshaped into sequences of tokens:

$$X' = \text{rearrange}(X,\ 'BCHW \to B(HW)C')$$
$$C'_f = \text{rearrange}(C_f,\ 'BC_cH_cW_c \to B(H_cW_c)C'_c)$$

Here, $X' \in \mathbb{R}^{B\times(H\cdot W)\times C}$ and $C'_f \in \mathbb{R}^{B\times(H_c\cdot W_c)\times C_c}$ are tokenized representations of the input and conditioning feature maps, where each token represents a spatial location and its corresponding feature vector.

The core of the cross-attention mechanism involves computing attention between the latent tokens X' and the conditioning tokens C'_f. This is achieved by projecting X' and C'_f into query, key, and value spaces. Let $W_Q \in \mathbb{R}^{C\times d}, W_K \in \mathbb{R}^{C_c\times d}, W_V \in \mathbb{R}^{C_c\times d}$ be the learned projection matrices for the query, key, and value spaces, where d is the dimensionality of the attention head:

$$\begin{aligned} Q &= X'W_Q \quad \text{(Query)} \\ K &= C'_f W_K \quad \text{(Key)} \\ V &= C'_f W_V \quad \text{(Value)} \end{aligned} \tag{5.1}$$

The attention scores are computed as the scaled dot-product between the query and key matrices:

$$\text{Attention}(Q, K) = \frac{QK^T}{\sqrt{d}}$$

The resulting attention scores, $A \in \mathbb{R}^{B\times(H\cdot W)\times(H_c\cdot W_c)}$, represent the similarity between each latent token and each conditioning token. These scores are then normalized using a softmax function:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

Finally, the output of the cross-attention layer is computed as the weighted sum of the value tokens, where the weights are given by the attention scores:

$$\text{Output} = AV$$

This output is then reshaped back into the original spatial dimensions of X:

$$\text{Output}_{\text{final}} = \text{rearrange}(\text{Output},\ 'B(HW)C \to BCHW')$$

The cross-attention allows the model to selectively focus on relevant parts of the conditioning feature map C_f, guided by the input latent features X. This mechanism helps the

model refine the latent representations by incorporating important contextual information from the conditioning network, enabling better reconstruction of fine spatial details during the diffusion process.

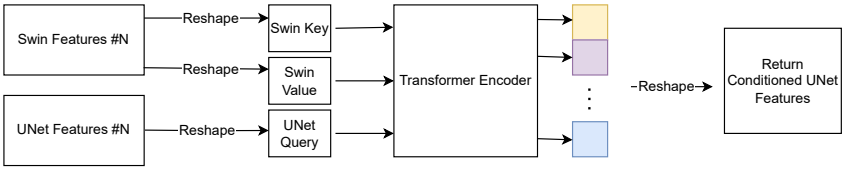


Figure 5.3: Spatial Cross Attention which takes in features from Swin and UNet and does cross attention to the Swin Features to condition the UNet features.

5.2.5 Latent Diffusion Model

The Latent Diffusion Model used in this work is based on an attention UNet [OSF+18] with ConvNeXt [LMW+22] building blocks. The model operates entirely in the latent space, where the input is first encoded into a compressed latent representation using a pretrained variational autoencoder as described in 5.2.2. This latent representation then serves as the input to the diffusion process.

At each stage of the UNet, spatial cross-attention blocks are applied. These blocks work by first passing the input through a ConvNeXt block and then performing cross-attention with the corresponding Swin features from the Conditioning Module. After that, the output of the block is used as a skip connection, and a downsampled copy is used as input for the next stage.

The decoder of the UNet essentially functions similarly to the encoder, with the difference being that the input is the upsampled output from the previous layer, concatenated with the skip connection from the corresponding encoder layer. This input is then passed through a ConvNeXt block, followed by cross-attention with the corresponding Swin features from the Conditioning Module.

The loss function for this model is the mean squared error:

$$\mathcal{L} = \text{MSE}(\boldsymbol{\epsilon}, \hat{\boldsymbol{\epsilon}})$$

where $\boldsymbol{\epsilon}$ is the true noise and $\hat{\boldsymbol{\epsilon}}$ is the predicted noise.

5.2.6 Variations

To show the effectiveness of the method, two variations of this model are introduced, called **Stable Diffusion Baseline** and **Stable Diffusion External Conditioning**. The Stable Diffusion Baseline generally follows the same architecture as described in 5.2.5, but with the exclusion of external conditioning. Therefore, conditioning in this method solely occurs due to concatenation. Stable Diffusion External Conditioning, on the other hand, also follows the same architecture as described in 5.2.5, but in this case, conditioning solely occurs through Swin Conditioning and not through concatenation. It is also important to note that these methods utilize the pretrained variational autoencoder from Stability AI [RBL+22] and not the finetuned version as this finetuning belongs to the own method.

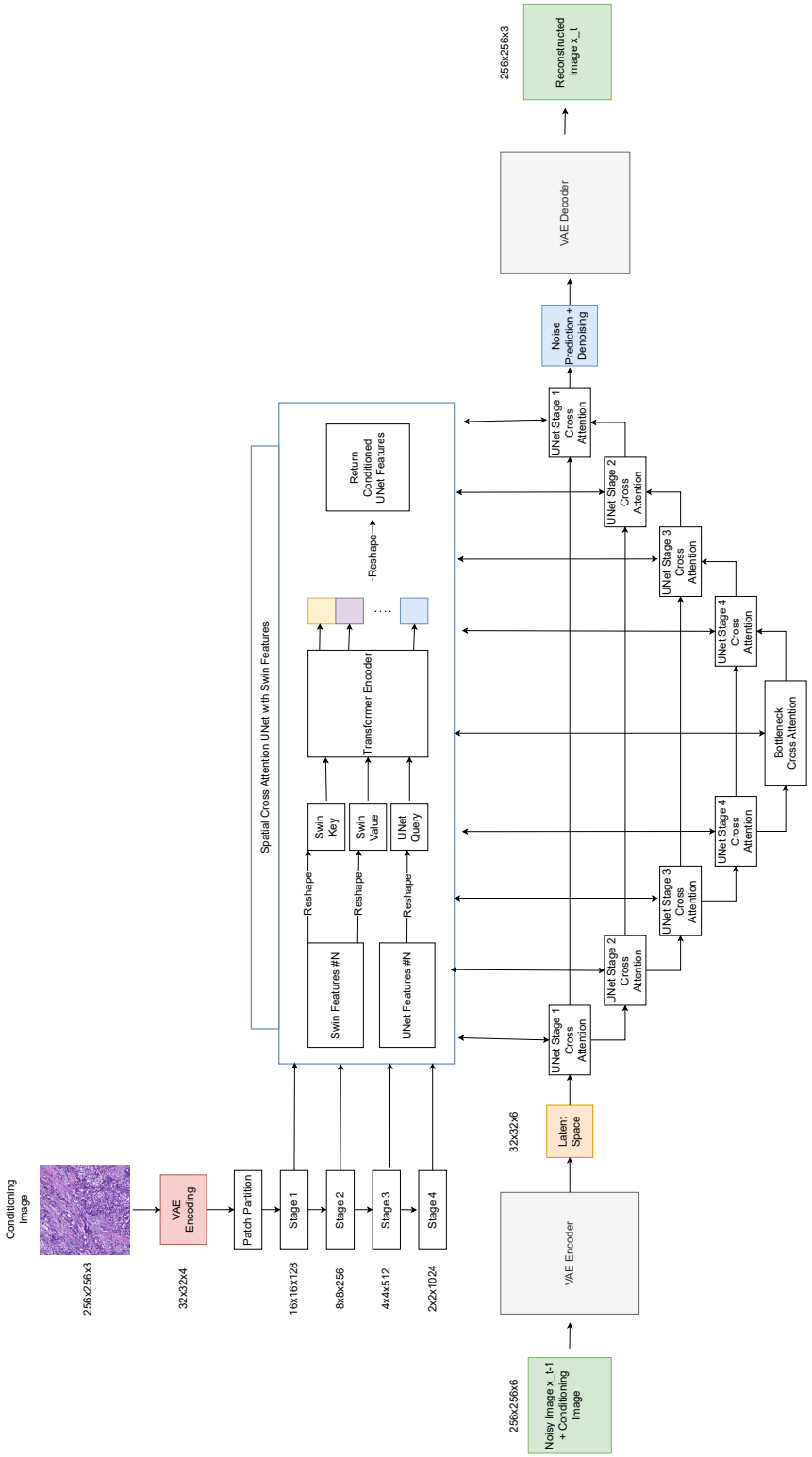


Figure 5.4: Whole Architecture of the developed method with the two inputs: The conditioning image from the top and the noisy image from timestep $x_t - 1$. The encoded noisy image and the encoded conditioning image is put through a UNet and at each stage the output of this stage cross attends to features from the Swin conditioning module. The output is the denoised image at timestep t.

6. Evaluation

In this section, the model's performance is evaluated across four different datasets, with the Brightfield-to-Fluorescence dataset serving as the competition dataset.

All training experiments were conducted on a single machine equipped with 2x 3090ti GPUs. AdamW was selected as the optimizer, and each model was trained with a batch size of 8. The learning rate was initialized at 1e-4 and decayed linearly to a minimum of 1e-6. All models were trained for up to 1000 epochs, with early stopping applied if no improvement in validation performance was observed for 100 consecutive epochs.

6.1 Variational Autoencoder Finetuning

The initial evaluation focuses on the variational autoencoder, fine-tuned for integration within the latent diffusion model. The autoencoder was trained separately on fluorescence and histological datasets, and the results for each dataset are presented individually.

Table 6.1: Metrics Comparison for Baseline and Finetuned Models across Different Channels

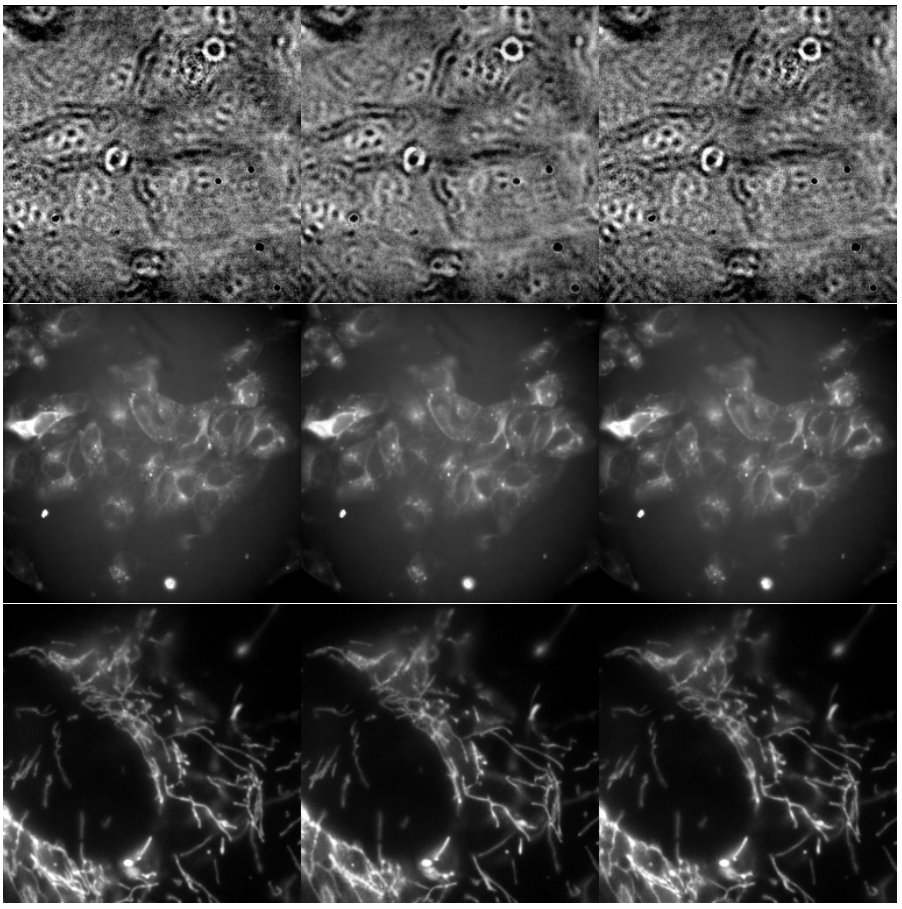
Model	**Nucleus**		**Mitochondria**		**Tubulin**		**Actin**		**BF**		**PC**		**DIC**	
	SSIM ↑	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑	MSE ↓
Baseline	**0.9669**	**0.0002**	0.9167	**0.0011**	0.8815	**0.0007**	0.9154	**0.0005**	**0.4475**	**0.0100**	0.7898	**0.0042**	0.7197	**0.0038**
Finetuned	0.9646	**0.0002**	**0.9275**	0.0011	**0.8818**	0.0008	**0.9154**	0.0005	0.4557	0.0114	**0.8059**	0.0045	**0.7322**	0.0039

In total, two different autoencoders were trained: one for the histological datasets and another for the fluorescence dataset. The finetuned autoencoder is based on the baseline model, which is the pretrained variational autoencoder from Stability AI's Stable Diffusion [RBL+22], as described in the methodology. As shown in Table 6.1, the finetuned version outperforms the baseline model in terms of SSIM across all metrics. However, it is noteworthy that for the Actin channel, no significant improvement in SSIM was observed, with a score of 0.8815 and a mean squared error of 0.0005. This small mean squared error is expected, given that fluorescence images often contain large black backgrounds

with only a small portion of the image containing an actual signal. Interestingly, the finetuned models sometimes shows a slightly higher mean squared error, despite qualitative evaluations showing no noticeable decrease in image quality. For the histological datasets, a total of three different models were trained. The baseline model, once again, is the pretrained VAE from Stability AI's Stable Diffusion [RBL+22]. The first finetuned model introduces the discriminator loss starting at 20,000 training steps, while the second model introduces the discriminator loss at 50,000 training steps. The model trained with the later-starting discriminator loss demonstrates the best performance across all datasets, achieving respective SSIM scores of 0.4487 and 0.5492. However, it performs slightly worse on the unstained-to-stained dataset with an SSIM of 0.8717. In qualitative analysis, the differences between the baseline and finetuned models are more pronounced. The baseline model tends to produce smoothed images across all datasets, particularly struggling with the reconstruction of fine details, such as small nuclei, as illustrated in Figure 6.2.

Table 6.2: Average SSIM Comparison for Different Models on Histological Datasets

	Multi IHC Stain Translation ER	**Breast Cancer Immunohistochemical Dataset**	**Unstained to Stained**
	SSIM ↑	SSIM ↑	SSIM ↑
Baseline Model	0.4226	0.5373	**0.8719**
Finetuned Model 1: $disc_start = 20000$	0.4391	0.5432	0.8648
Finetuned Model 2: $disc_start = 50001$	**0.4487**	**0.5492**	0.8717

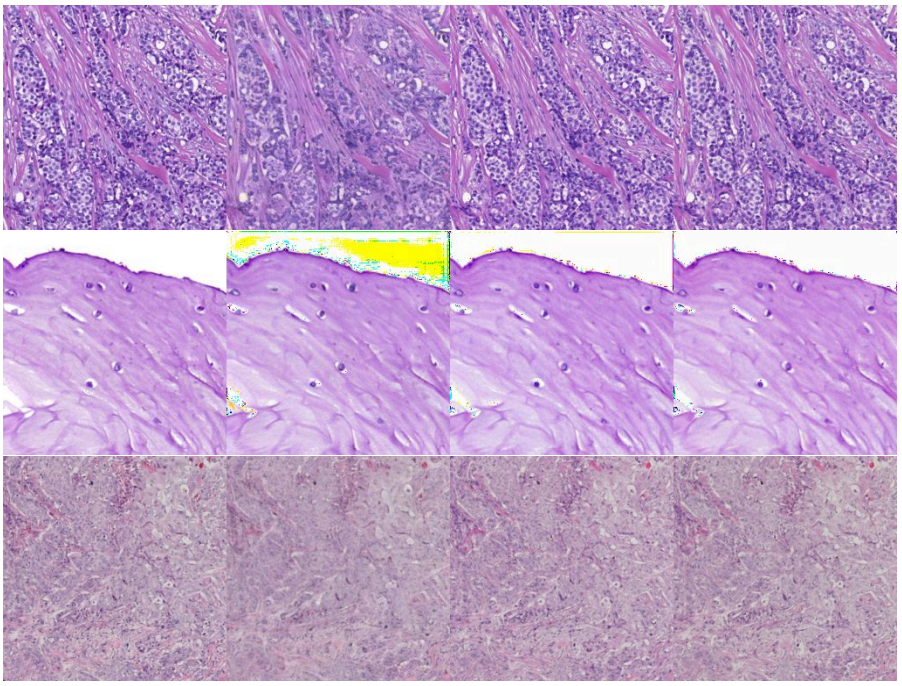


(a) Ground Truth

(b) Baseline: Stable Diffusion Pretrained Autoencoder [RBL+22]

(c) Finetuned Autoencoder

Figure 6.1: Results of applying autoencoder on the fluorescence data [FB24]. Each row shows results for ground truth, baseline, and the finetuned model of the variational autoencoder



(a) Ground Truth (b) Baseline: Stable Diffusion Pretrained Autoencoder [RBL+22] (c) Finetuned Model 1: *disc_start* = 20000 (d) Finetuned Model 2: *disc_start* = 50001

Figure 6.2: Results of applying autoencoder. Each row shows results for ground truth, baseline, and two finetuned models with different *disc_start* values. From left to right: ground truth, stable diffusion pretrained autoencoder, finetuning *disc_start* = 20000, and finetuning *disc_start* = 50001 [ARA+24].

6.2 Brightfield to Fluorescence

The results in 6.3 showcase the performance of the implemented models. Each of these models employs a weighted combination of L1 and L2 reconstruction losses, in addition to SSIM loss. For the Pix2Pix models, an additional GAN loss and discriminator loss are incorporated as described in 5.1.4. The all-channel models were designed to predict all four organelles simultaneously, whereas the single-channel models focus on predicting one organelle at a time, resulting in four distinct models—one for each organelle.

Table 6.3: Metrics Comparison for Different Models and Organelles

	Nucleus			**Mitochondria**			**Actin**			**Tubulin**			**Average**		
	SSIM ↑	MSE ↓	PCC ↑	SSIM ↑	MSE ↓	PCC ↑	SSIM ↑	MSE ↓	PCC ↑	SSIM ↑	MSE ↓	PCC ↑	SSIM ↑	MSE ↓	PCC ↑
Autoencoder (Single Channel)	0.820	**0.017**	0.713	0.786	**0.015**	0.672	**0.581**	0.025	0.612	0.788	**0.022**	0.697	0.744	**0.020**	0.674
Autoencoder (All Channel)	0.770	0.021	0.621	0.727	0.019	0.584	0.466	0.036	0.531	0.730	0.031	0.606	0.673	0.027	0.586
Variational Autoencoder (Single Channel)	0.730	0.035	0.705	0.685	0.020	0.665	0.440	0.022	0.565	0.688	0.044	0.710	0.636	0.030	0.661
Pix2Pix (Single Channel)	**0.838**	0.057	**0.820**	0.791	0.030	**0.780**	0.564	0.017	**0.617**	0.813	0.066	**0.844**	**0.752**	0.043	**0.765**
Pix2Pix (All Channel)	0.773	0.055	0.813	0.741	0.029	0.773	**0.581**	**0.016**	0.612	0.742	0.063	0.837	0.709	0.041	0.759
Double UNet (All Channel)	0.705	0.045	0.760	0.660	0.025	0.720	0.425	0.020	0.590	0.662	0.055	0.775	0.613	0.036	0.711
Own	0.765	0.020	0.633	**0.805**	**0.015**	0.733	**0.581**	0.025	0.612	**0.817**	**0.022**	0.697	0.740	0.021	0.669

6.3 demonstrates promising results for both the autoencoder and Pix2Pix models. The differences between the single-channel autoencoder and the single-channel Pix2Pix model

are minimal. However, with SSIM scores of 0.838, 0.791, 0.564, and 0.813, the Pix2Pix model outperforms the autoencoder in three of the four channels. The only organelle where the autoencoder achieves better performance is Actin, which has the least amount of training data. In terms of mean squared error (MSE), the differences between the models are again marginal, with no clear advantage for either model. The single-channel autoencoder performs best on average across all channels, with an MSE of 0.02. Although the Pix2Pix (Single Channel) model has a slightly higher average MSE of 0.043 compared to the autoencoder's 0.02, the Pix2Pix (Single Channel) model performed better during the first test iteration of the challenge and was therefore selected for submission. The results from the own method were obtained after the challenge. Given that this method is a latent diffusion model and the dataset is relatively small, it is expected that simpler architectures may perform similarly or even better. Especially in comparison to histological datasets, where new structures often need to be introduced that are not visible in the input image, the fluorescence dataset does not have such complexities. Therefore, even a simple autoencoder performs exceptionally well in this context.

When examining the results of the worldwide challenge, which had a total of 217 participants, the submitted Pix2Pix (Single Channel) model achieved an overall ranking of 7th place. The evaluation was based on standard metrics used in the field of image analysis. These include the Mean Absolute Error, which measures the absolute difference between predicted and ground truth images; the Structural Similarity Index Measure, which evaluates the perceived quality of the predicted image compared to the original based on luminance, contrast, and structure; the Pearson Correlation Coefficient, which assesses the linear correlation between the predicted and ground truth images; and the Euclidean and Cosine Distances for texture analysis, which compare the spatial distances between the original and predicted images. For organelles with more available training data, such as Nucleus and Mitochondria, the model performed even better, securing the 5th rank globally, and occasionally reaching 4th place, such as for the mean average error on the mitochondria channel. However, for channels with less available data, like Actin and Tubulin, the model ranked lower, with a 9th place for both organelles. This lower performance can partly be attributed to the fact that other publicly available datasets were allowed for use in the competition, and many competing models utilized ensembles of up to five models. Additionally, the Pix2Pix (Single Channel) model was trained on an image size of 448x448 due to time constraints and hardware limitations. Other methods, such as [LL24], reportedly used publicly available datasets like the Jump Cell Painting dataset [CCG+24, CAA+23], which contains an additional 58,237 tiles.

Table 6.4: Top 10 User Performances Across Different Channels from the competition. Row number 7 is the own submission

#	**Nucleus**					**Mitochondria**			**Actin**		**Tubulin**	
	SSIM (Pos)	MAE (Pos)	PCC (Pos)	E_dist (Pos)	C_dist (Pos)	MAE (Pos)	PCC (Pos)	SSIM (Pos)	PCC (Pos)	SSIM (Pos)	PCC (Pos)	SSIM (Pos)
1	0.7624 (1)	0.0612 (1)	0.7438 (4)	164.8783 (1)	0.1769 (4)	0.0840 (1)	0.5583 (5)	0.6793 (2)	0.6399 (3)	0.5883 (1)	0.6253 (3)	0.6945 (2)
2	0.7391 (3)	0.0725 (4)	0.7063 (5)	183.4105 (4)	0.1950 (5)	0.0926 (2)	0.5749 (4)	0.6798 (1)	0.5869 (5)	0.5270 (6)	0.6593 (1)	0.7083 (1)
3	0.7466 (2)	0.0699 (3)	0.7458 (3)	176.6736 (3)	0.1693 (3)	0.1032 (6)	0.5999 (3)	0.6537 (4)	0.5995 (4)	0.5460 (5)	0.5642 (5)	0.6774 (4)
4	0.6976 (6)	0.0697 (2)	0.7782 (1)	169.2790 (2)	0.1484 (1)	0.1034 (7)	0.6088 (2)	0.6287 (6)	0.6482 (2)	0.4946 (8)	0.5679 (4)	0.6431 (8)
5	0.6757 (7)	0.0872 (7)	0.7761 (2)	199.6779 (7)	0.1487 (2)	0.1198 (10)	0.6274 (1)	0.6247 (7)	0.6613 (1)	0.5543 (4)	0.6361 (2)	0.6819 (3)
6	0.7196 (4)	0.0798 (6)	0.6098 (7)	195.3898 (5)	0.2610 (7)	0.0958 (3)	0.4838 (7)	0.6557 (3)	0.5620 (6)	0.5548 (3)	0.5381 (6)	0.6699 (6)
7	0.7082 (5)	0.0791 (5)	0.6326 (6)	197.4403 (6)	0.2493 (6)	0.0972 (4)	0.4952 (6)	0.6426 (5)	0.2043 (9)	0.4404 (9)	0.4777 (7)	0.5543 (9)
8	0.5747 (10)	0.1035 (9)	0.3935 (10)	246.7677 (10)	0.4354 (10)	0.1023 (5)	0.3151 (10)	0.5922 (8)	0.5409 (7)	0.5766 (2)	0.4226 (8)	0.6458 (7)
9	0.6658 (8)	0.0891 (8)	0.5810 (8)	222.6335 (8)	0.2941 (8)	0.1106 (9)	0.4264 (8)	0.5913 (9)	0.1427 (10)	0.4138 (10)	0.4107 (9)	0.6729 (5)
10	0.5898 (9)	0.1036 (10)	0.5142 (9)	239.4568 (9)	0.3509 (9)	0.1036 (8)	0.4014 (9)	0.5592 (10)	0.5246 (8)	0.5061 (7)	0.3109 (10)	0.4840 (10)

Figure 6.3 presents the qualitative comparison of images generated by different models. The Pix2Pix model outputs appear somewhat smoothed, a commonly observed issue in models that rely heavily on reconstruction loss alone. This effect is due to the mean squared error averaging out errors, leading to less pronounced details. The Hybrid Swin Conditional Refinement Latent Diffusion Model generally yields satisfactory results across the sample images, though it encounters some challenges in reconstructing nuclei at lower magnification. This limitation may account for the slightly reduced SSIM score observed for the nucleus organelle. However, aside from this, the predictions produced by the Hybrid model are sharper and retain more detail than those from the Pix2Pix model, a quality that is particularly significant in the medical domain. An interesting observation, consistent with the autoencoder pretraining phase, is the occasional appearance of randomly colored pixels, the cause of which is yet to be determined. Despite this minor artifact, the results are highly encouraging, demonstrating strong potential for further application.

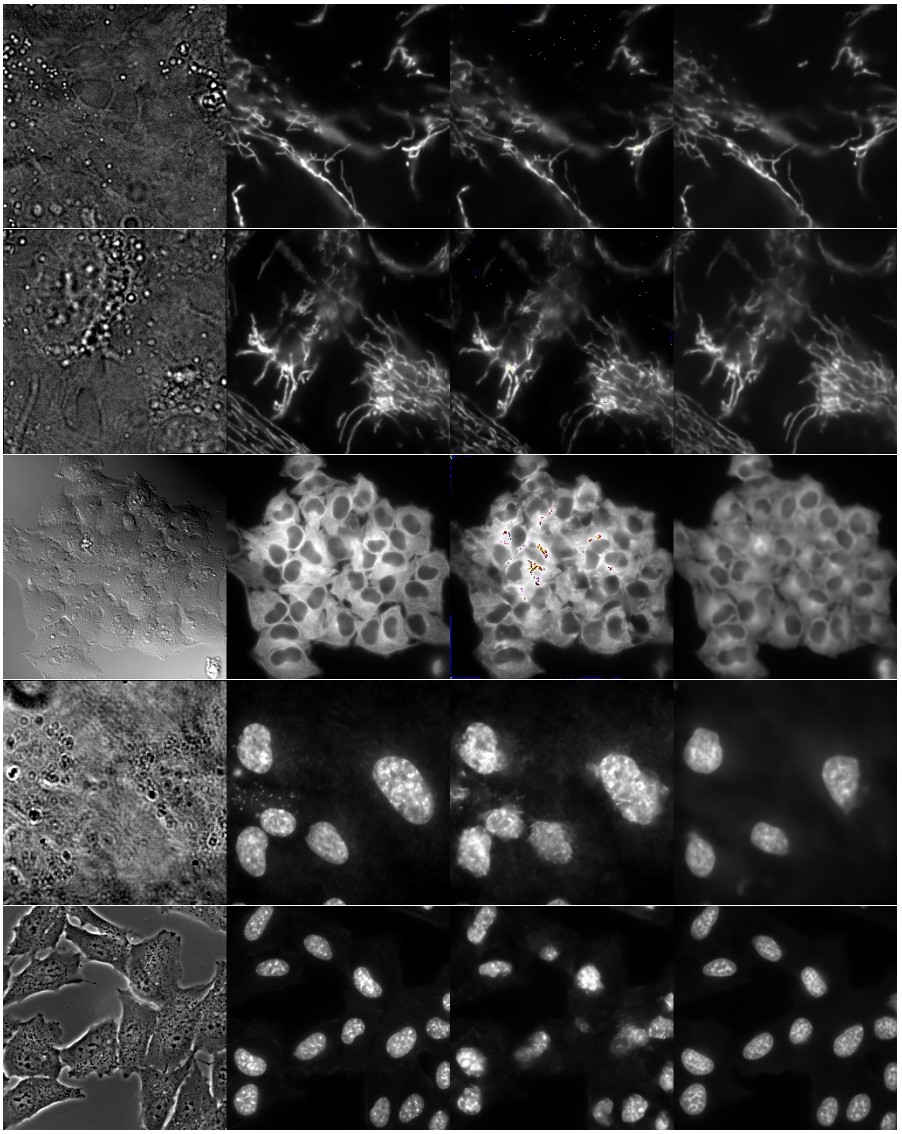
6.3 Unstained to Stained

Method	**SSIM**
Pix2Pix [IZZE17]	0.512
Latent Diffusion Baseline	0.565
Latent Diffusion Ext. Conditioning	0.584
Hybrid Swin Conditional Refinement Latent Diffusion	**0.613**

Table 6.5: SSIM comparison of different methods on the Unstained to Stained dataset

The evaluation on the unstained-to-stained dataset was conducted using SSIM as the primary metric because the labels are not entirely pixel-perfect due to the nature of the data acquisition process. The tissue images were registered on top of each other, introducing slight misalignment errors visible in the training data. Consequently, the Pix2Pix model [IZZE17], which relies on a reconstruction loss for the generator, is expected to perform poorly in this scenario. For the latent diffusion models, the baseline version utilizes direct conditioning by concatenating the channels of the conditioning network to the noisy input image as described in 5.2.6. The external conditioning model uses only external conditioning via the Swin Refinement module as described in 5.2.6.

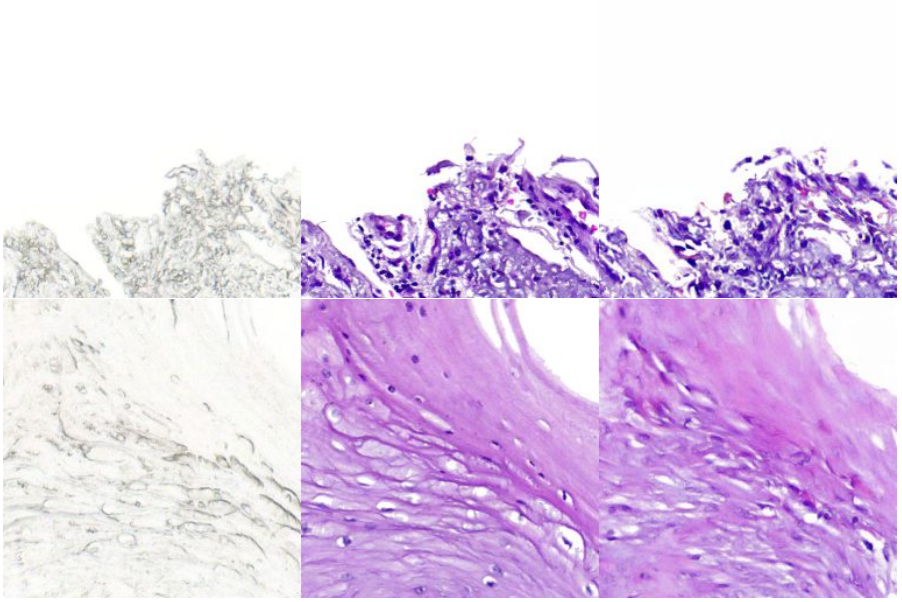
The own method achieves the highest SSIM of 0.613, as seen in 6.5, although the performance is relatively close to the other conditioning methods, which report SSIM values of 0.584 and 0.565, respectively. The only model that stands out negatively is the Pix2Pix model, with a notably lower SSIM of 0.512. Overall, these results highlight the superiority of latent diffusion models over generative adversarial networks, especially when working with data that is not pixel-perfect. This advantage can be attributed to the fact that the input to the VAE decoder is sampled, which helps mitigate the pixel discrepancies between the input and the ground truth images. When evaluating the own method qualitatively in figure 6.4, it becomes clear that, at first glance, the predictions appear very accurate. However, upon closer inspection of finer details, particularly around some intricate nuclei, the predictions are not entirely perfect. Despite this, the overall performance is impressive, demonstrating the model's strong capabilities on this dataset. The occasional omission of details can likely be attributed to the fact that not all structures are visible in the



(a) Input (b) Ground Truth (c) Own (d) Pix2Pix (Single Channel)

Figure 6.3: Results of the own developed method and the submitted Pix2Pix (Single Channel) model on the brightfield to fluorescence dataset [FB24].

unstained tissue, requiring the model to infer or "imagine" where certain structures might be.



(a) Input (b) Ground Truth (c) Model Prediction

Figure 6.4: Results of the own method on the unstained to stained dataset [ARA+24]

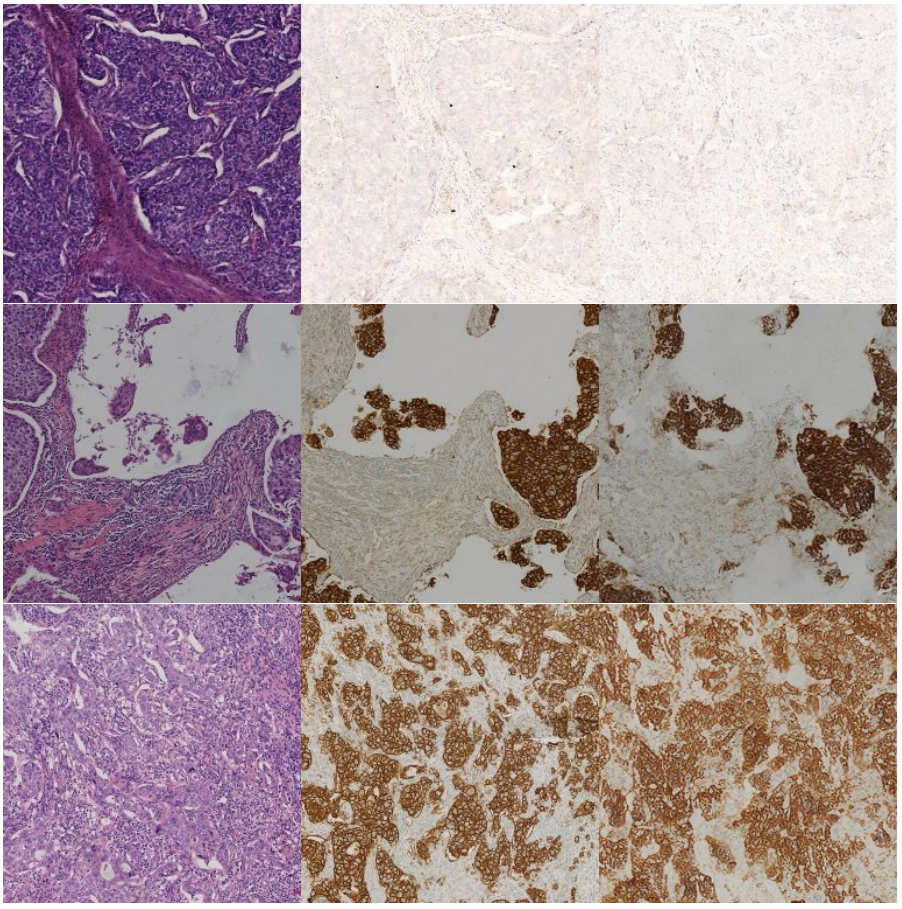
6.4 Breast Cancer Immunohistochemical Dataset

The evaluation on the third dataset was conducted similarly to the previous datasets, using SSIM as the primary metric to assess the quality of generated images. Due to inherent inconsistencies in the data acquisition process, which includes slight misalignments between input and target images, the generative models are challenged to provide pixel-perfect outputs. As such, generative models like CycleGAN and Pix2Pix, which rely heavily on pixel-level alignment for training and evaluation, are expected to encounter limitations in this dataset. The baseline models, including CycleGAN and Pix2Pix variations [BM21], show moderate performance, with SSIM scores ranging from 0.373 for CycleGAN to 0.471 for Pix2PixHD [BM21] as seen in 6.6. Pix2Pix (UNet generator) [BM21] achieves a marginally better SSIM of 0.419, but still struggles with the precision required for this dataset. The introduction of the ResNet generator [BM21] improves the performance slightly, bringing the SSIM to 0.440, while the Pyramid Pix2Pix [BM21] model further refines the reconstruction, achieving an SSIM of 0.477. CycleGAN, which performs well in tasks involving pure style transfer, shows limitations when generating new complex structures [BM21]. This shows that while CycleGAN excels in transferring stylistic features between domains, it falls short when structural accuracy and generation of new, intricate details are required. In contrast, Pix2Pix variants, which incorporate a supervised learning framework, perform better but still exhibit room for improvement. The Hybrid Swin Conditional Refinement Latent Diffusion Model appears to handle pixel misalignment better than the other approaches. This method achieves the highest SSIM score of 0.496, sur-

passing the performance of all baseline models. This suggests that diffusion-based models may be more robust when dealing with the inconsistencies posed by this dataset.

Method	**SSIM**
CycleGAN [LZX+22]	0.373
Pix2Pix (UNet generator) [LZX+22]	0.419
Pix2Pix (ResNet generator) [LZX+22]	0.440
Pix2PixHD [LZX+22]	0.471
Pyramid Pix2Pix [LZX+22]	0.477
Hybrid Swin Conditional Refinement Latent Diffusion	**0.496**

Table 6.6: SSIM comparison of different methods



(a) Input (b) Ground Truth (c) Own Model Prediction

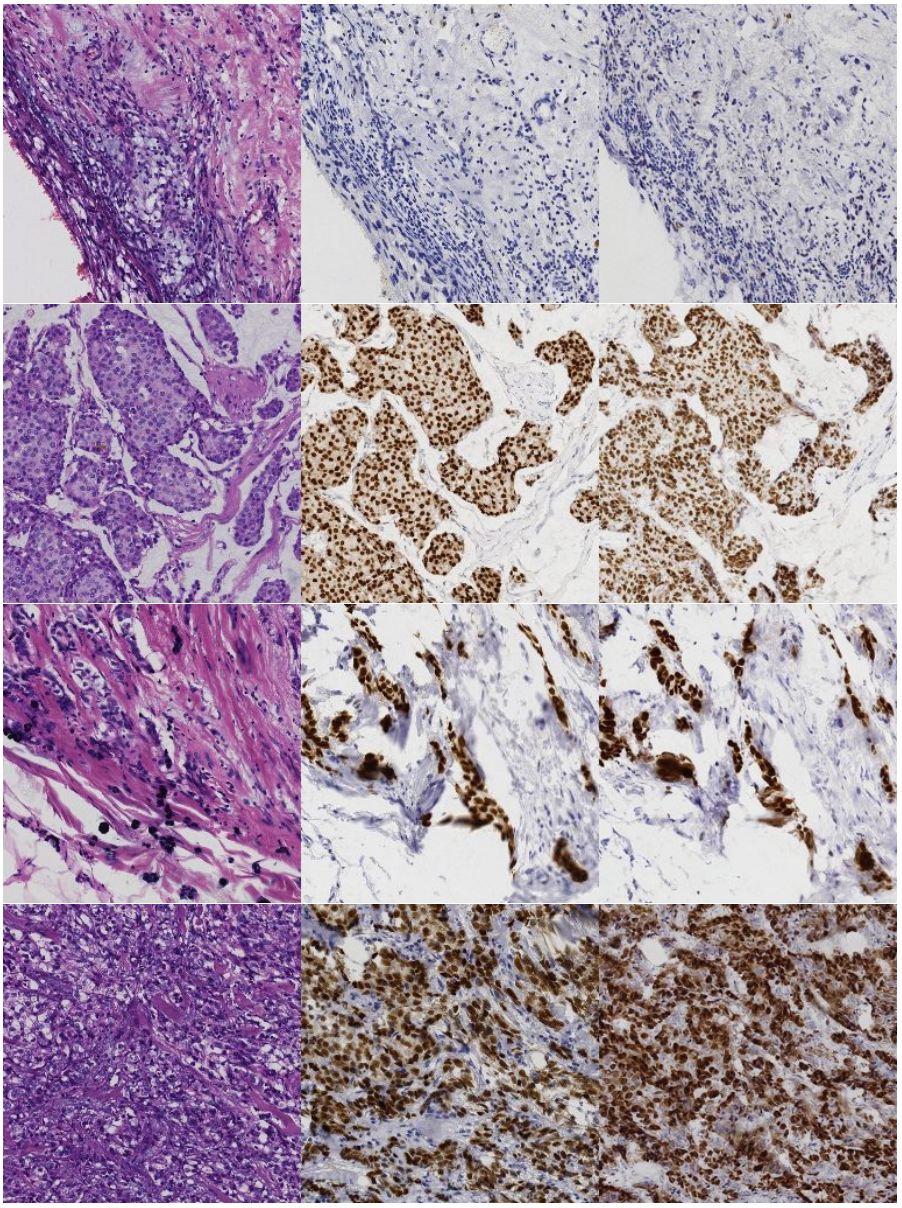
Figure 6.5: Results for different samples from the test set of the Breast Cancer Immunohistochemical dataset [BM21]. Each row contains the input image, the corresponding ground truth, and the model's prediction from left to right.

6.5 Multi IHC Stain Translation

The SSIM comparison of different methods, as shown in Table 6.7, highlights the performance differences across a variety of approaches. The evaluation was performed on a dataset where pixel-perfect accuracy is not guaranteed, and as such, these SSIM values provide useful insights into how each model handles imperfect data. The Pix2Pix model [IZZE17], as expected, performs the worst with an SSIM of 0.132. This is likely due to its reliance on pixel-wise reconstruction loss, which can be overly sensitive to small inaccuracies in the data. Given that the dataset contains inherent misalignments, Pix2Pix struggles to capture meaningful structures, resulting in poor performance. The Latent Diffusion Baseline Model, which directly conditions the noisy input image by concatenating the conditioning network's channels, shows an improved SSIM of 0.185. This demonstrates that diffusion-based models are better suited to handling imperfect data, as their use of sampling during the decoding process helps mitigate pixel-wise discrepancies. The Latent Diffusion Ext. Conditioning model performs even better, with an SSIM of 0.215. This model relies entirely on external conditioning via the Swin Refinement module, which likely helps the model focus on broader structures rather than pixel-level details, improving its ability to reconstruct meaningful features. The best-performing model is the Hybrid Swin Conditional Refinement Latent Diffusion Model, achieving an SSIM of 0.242. This model combines both direct and external conditioning, incorporating the strengths of both strategies to outperform the other methods. The small improvement over the other latent diffusion models shows the advantage of combining these conditioning techniques. Interestingly, the Adaptive Supervised PatchNCE model [LHCK23], with an SSIM of 0.2236, performs quite well compared to the diffusion models. This method, which focuses on learning patch-based representations, handles the non-pixel-perfect data reasonably well, but still falls short of the performance seen with the Hybrid Swin Conditional Refinement Latent Diffusion Model. Overall, the results indicate that latent diffusion models consistently outperform GAN-based models like Pix2Pix, especially when dealing with datasets that contain inherent pixel-level inaccuracies. The ability of diffusion models to sample during reconstruction helps them navigate inconsistencies in the data, resulting in higher SSIM scores across the board.

Method	**SSIM**
Adaptive Supervised PatchNCE [LHCK23]	0.2236
Pix2Pix [IZZE17]	0.132
Latent Diffusion Baseline	0.185
Latent Diffusion Ext. Conditioning	0.215
Hybrid Swin Conditional Refinement Latent Diffusion	**0.242**

Table 6.7: SSIM comparison of different methods on the Multi IHC Stain Translation dataset [LHCK23]



(a) Input (b) Ground Truth (c) Own Model Prediction

Figure 6.6: Results for different samples taken from the Multi IHC Stain Translation dataset. Each row contains the input image, the corresponding ground truth, and the own model's prediction from left to right [LHCK23].

7. Conclusion

This chapter summarizes the work conducted in this thesis and provides an outlook on potential directions for future research.

7.1 Summary

In this thesis, the background and origins of digital microscopy were introduced, and its advantages over classical microscopy were highlighted. Following this, the fundamentals of deep learning and microscopy were explained to provide a foundation for understanding the rest of the thesis. The various datasets used later in the work were then introduced, with explanations structured to convey the biological relevance, technical challenges, and general statistics of each dataset. It was also pointed out that the Brightfield to Fluorescence dataset is the competition dataset used for participation in the competition. The methods section was divided into two parts. The first part presented the models and approaches developed and evaluated for participation in the competition. The second part introduced the custom latent diffusion method that was evaluated across all datasets. In the evaluation section, the models were assessed on all datasets. First, the results of the fine-tuned variational autoencoder on the Brightfield to Fluorescence dataset were presented, followed by results on the other datasets. It was demonstrated that the fine-tuned autoencoder generally outperformed the baseline model, which often produced blurry samples. The first dataset evaluated was the Brightfield to Fluorescence dataset, where it was shown that even non-generative models, such as a simple autoencoder, performed well. This indicates that the input image generally contains sufficient information to predict the four organelles. There was only a slight difference between the Autoencoder and the Pix2Pix model, with the Pix2Pix model achieving an SSIM of 0.752 compared to 0.744 for the Autoencoder. The custom method showed competitive performance with an SSIM of 0.740 and an MSE of 0.021. Furthermore, the Pix2Pix (Single Channel) model achieved a global 7th place in the competition, which is a satisfactory result, especially since no additional external data were used for this thesis, while other participants utilized external data. The fact that the custom method did not achieve the best results here can be easily explained by the small dataset size. On the unstained-to-stained dataset, the strong performance of the custom method became apparent. Here, the latent diffusion models with

various conditioning methods significantly outperformed the Pix2Pix model. The custom method achieved an SSIM of 0.613 compared to 0.472 for the Pix2Pix model. This can be attributed to the fact that the Pix2Pix generator uses a pixel-based loss, making training very challenging on a dataset that is not entirely pixel-perfect. On the third dataset, the Breast Cancer Immunohistochemical Dataset, the custom method also outperformed various methods based on Pix2Pix. Although the difference was not as pronounced, the custom method achieved an SSIM of 0.496 compared to 0.477 for the next-best model, the Pyramid Pix2Pix. Similarly, on the final Multi IHC Stain Translation dataset, the custom model performed best with an SSIM of 0.242 compared to 0.2236 for the Adaptive Supervised PatchNCE model. The qualitative analysis also showed good prediction results across all tasks, with noticeably better performance on the last three datasets, which had more training data. Overall, it was demonstrated that latent diffusion models are very well-suited for image-to-image translation and, given sufficient data, can outperform existing methods.

7.2 Future Work

This thesis provides a solid foundation for further research in the field of image-to-image translation. One potential direction for future work would be to investigate whether pretraining the conditioning network directly in the latent space could lead to performance improvements.

Additionally, the variational autoencoder plays an important role in latent diffusion, making it essential to explore alternative architectures. For example, examining the impact of using a VQGAN [ERO21] could offer valuable insights. While the variational autoencoder in this thesis incorporates a discriminator, the latent representations provided by a VQGAN may capture more nuanced details. There is ongoing research in the development of autoencoders, with recent implementations often bypassing strict constraints on output range, such as the typical range of $[-1, 1]$, which could be easily enforced using a `tanh` activation function or simple clipping. Within the broader latent diffusion architecture, the autoencoder remains one of the most promising areas for further investigation, given its crucial role in representation quality and detail preservation.

Another promising path for future work is few-shot image-to-image translation, which would allow models to adapt to any staining type by leveraging only a few examples. This capability would significantly expand the flexibility and applicability of image translation models within medical imaging.

Additionally, exploring the replacement of the UNet with a diffusion transformer [PX23] presents an exciting direction, particularly for medical applications. The transformer's capacity to capture global features could provide an advantage in domains where contextual information is essential, making transformer-based denoising a potentially valuable alternative.

Bibliography

[AFR05] A. K. Abbas, N. Fausto, and S. L. Robbins, *Robbins and Cotran pathologic basis of disease*. Elsevier Saunders, 2005.

[AJHVD12] S. Al-Janabi, A. Huisman, and P. J. Van Diest, "Digital pathology: current status and future perspectives," *Histopathology*, vol. 61, no. 1, pp. 1–9, 2012.

[AJW+08] B. Alberts, A. Johnson, P. Walter, J. Lewis, M. Raff, and K. Roberts, "Molecular cell biology," *New York: Garland Science*, 2008.

[ARA+24] M. Z. Asaf, B. Rao, M. U. Akram, S. G. Khawaja, S. Khan, T. M. Truong, P. Sekhon, I. J. Khan, and M. S. Abbasi, "Dual contrastive learning based image-to-image translation of unstained skin tissue into virtually stained h&e images," *Scientific Reports*, vol. 14, no. 1, p. 2335, 2024.

[BBL+16] B. E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, and J. A. Van Der Laak, "Automated detection of dcis in whole-slide h&e stained breast histopathology images," *IEEE transactions on medical imaging*, vol. 35, no. 9, pp. 2141–2150, 2016.

[BG08] J. D. Bancroft and M. Gamble, *Theory and practice of histological techniques*. Elsevier health sciences, 2008.

[Bio24] A. Bioquest, "Cellular fluorescence microscopy troubleshooting and best practices," 2024, accessed: 2024-10-29. [Online]. Available: https://www.aatbio.com/resources/application-notes/cellular-fluorescence-microscopy-troubleshooting-best-practices

[BLT+15] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. Van Der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2015.

[BM21] R. Bose and C. X. Ma, "Breast cancer, her2 mutations, and overcoming drug resistance." *The New England journal of medicine*, vol. 385 13, pp. 1241–1243, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237608068

[BS42] C. Burch and J. Stock, "Phase-contrast microscopy," *Journal of Scientific Instruments*, vol. 19, no. 5, p. 71, 1942.

[CAA+23] S. N. Chandrasekaran, J. Ackerman, E. Alix, D. M. Ando, J. Arevalo, M. Bennion, N. Boisseau, A. Borowa, J. D. Boyd, L. Brino *et al.*, "Jump

cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations," *BioRxiv*, pp. 2023–03, 2023.

[CCG+24] S. N. Chandrasekaran, B. A. Cimini, A. Goodale, L. Miller, M. Kost-Alimova, N. Jamali, J. G. Doench, B. Fritchman, A. Skepner, M. Melanson *et al.*, "Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations," *Nature Methods*, pp. 1–8, 2024.

[CCJ41] A. H. Coons, H. J. Creech, and R. N. Jones, "Immunological properties of an antibody containing a fluorescent group." *Proceedings of the society for experimental biology and medicine*, vol. 47, no. 2, pp. 200–202, 1941.

[CZMW+22] J. O. Cross-Zamirski, E. Mouchet, G. Williams, C.-B. Schönlieb, R. Turkki, and Y. Wang, "Label-free prediction of cell painting from brightfield images," *Scientific reports*, vol. 12, no. 1, p. 10001, 2022.

[D+20] W. Dean *et al.*, "Emerging advances to transform histopathology using virtual staining," *BME frontiers*, 2020.

[DCLT19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[dHZZ+21] K. de Haan, Y. Zhang, J. E. Zuckerman, T. Liu, A. E. Sisk, M. F. Diaz, K.-Y. Jen, A. Nobori, S. Liou, S. Zhang *et al.*, "Deep learning-based transformation of h&e stained tissues into special stains," *Nature communications*, vol. 12, no. 1, pp. 1–13, 2021.

[DSW90] W. Denk, J. H. Strickler, and W. W. Webb, "Two-photon laser scanning fluorescence microscopy," *Science*, vol. 248, no. 4951, pp. 73–76, 1990.

[ERO21] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[FB24] France-Bioimaging, "Light my cells :bright field to fluorescence imaging challenge," 2024, accessed: 2024-10-30. [Online]. Available: https://lightmycells.grand-challenge.org/

[Gil16] S. F. Gilbert, "Ecological developmental biology: interpreting developmental signs," *Biosemiotics*, vol. 9, pp. 51–60, 2016.

[GPAM+20] 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.

[HG16] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[HGK+15] T. Huynh, Y. Gao, J. Kang, L. Wang, P. Zhang, J. Lian, and D. Shen, "Estimating ct image from mri data using structured random forest and

auto-context model," *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 174–183, 2015.

[HJA20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[HMW15] J. Hyllner, C. Mason, and I. Wilmut, "Cells: from robert hooke to cell therapy—a 350 year journey," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:33625406

[HPW+23] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank *et al.*, "Noise2music: Text-conditioned music generation with diffusion models," *arXiv preprint arXiv:2302.03917*, 2023.

[HS22] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[HSM17] J. R. Hermans, G. Spanakis, and R. Möckel, "Accumulated gradient normalization," in *Asian Conference on Machine Learning*. PMLR, 2017, pp. 439–454.

[IJK+21] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[IZZE17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[JBM17] A. Janowczyk, A. Basavanhally, and A. Madabhushi, "Stain normalization using sparse autoencoders (stanosa): application to digital pathology," *Computerized Medical Imaging and Graphics*, vol. 57, pp. 50–61, 2017.

[Kap95] J. Kapuscinski, "Dapi: a dna-specific fluorescent probe," *Biotechnic & histochemistry*, vol. 70, no. 5, pp. 220–233, 1995.

[KB14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106

[KL51] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[KRTM14] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[KSV14] N. Kaul, V. Soppina, and K. J. Verhey, "Effects of α-tubulin k40 acetylation and detyrosination on kinesin-1 motility in a purified system," *Biophysical journal*, vol. 106, no. 12, pp. 2636–2643, 2014.

[KSW15] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *Advances in neural information processing systems*, vol. 28, 2015.

[L+82] W. Lang *et al.*, "Nomarski differential interference-contrast microscopy," *ZEISS INFORMATION NO . 70*, 1982.

[Lak06] J. Lakowicz, "Principles of fluorescence spectroscopy," *University of Maryland School of Medicine Baltimore*, vol. 132, 2006.

[LBY+20] R. Lopez, P. Boyeau, N. Yosef, M. I. Jordan, and J. Regier, "Auto-encoding variational bayes," in *In Proceedings of the Second International Conference on Learning Representations, ICLR*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211146177

[LGZ+21] J. Li, J. Garfinkel, X. Zhang, D. Wu, Y. Zhang, K. De Haan, H. Wang, T. Liu, B. Bai, Y. Rivenson *et al.*, "Biopsy-free in vivo virtual histology of skin using deep learning," *Light: Science & Applications*, vol. 10, no. 1, p. 233, 2021.

[LH17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270

[LHCK23] F. Li, Z. Hu, W. Chen, and A. Kak, "Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 632–641.

[LL24] T. Le and E. Lundberg, "High-resolution in silico painting with generative models," *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2024/06/03/2024.05.31.596710

[LLC+21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[LMW+22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[LOHJ21] G. Lee, J.-W. Oh, N.-G. Her, and W.-K. Jeong, "Deephcs++: Bright-field to fluorescence microscopy image conversion using multi-task learning with adversarial losses for label-free high-content screening," *Medical image analysis*, vol. 70, p. 101995, 2021.

[LOK+18] G. Lee, J.-W. Oh, M.-S. Kang, N.-G. Her, M.-H. Kim, and W.-K. Jeong, "Deephcs: bright-field to fluorescence microscopy image conversion using deep learning for label-free high-content screening," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 335–343.

[LSSK19] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying relu and initialization: Theory and numerical examples," *arXiv preprint arXiv:1903.06733*, 2019.

[LW75] S. A. Latt and J. C. Wohlleb, "Optical studies of the interaction of 33258 hoechst with dna, chromatin, and metaphase chromosomes," *Chromosoma*, vol. 52, no. 4, pp. 297–316, 1975.

[LXT+18] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," *Advances in neural information processing systems*, vol. 31, 2018.

[LZX+22] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin, "Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1815–1824.

[Mis20] D. Misra, "Mish: A self regularized non-monotonic activation function," in *British Machine Vision Conference*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221113156

[ML16] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical image analysis*, vol. 33, pp. 170–175, 2016.

[ML18] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.

[MMI+22] A. Mustafa, A. Mikhailiuk, D. A. Iliescu, V. Babbar, and R. K. Mantiuk, "Training a task-specific image reconstruction loss," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2319–2328.

[Mok24] F. Mokobi, "Brightfield microscope: Principle, parts, applications," 2024, accessed: 2024-10-30. [Online]. Available: https://microbenotes.com/brightfield-microscope/

[MTC+09] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke, "Colour normalisation in digital histopathology images," in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, vol. 100. Daniel Elson London, 2009, pp. 100–111.

[MVH11] D. Mueller, D. Vossen, and B. Hulsken, "Real-time deformable registration of multi-modal whole slides for digital pathology," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 542–556, 2011.

[NDR+21] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335086

[Ock23] W. Ockenga, "Brightfield microscope: Principle, parts, applications," 2023, accessed: 2024-10-30. [Online]. Available: https://www.leica-microsystems.com/science-lab/microscopy-basics/phase-contrast-and-microscopy/

[ODKD10] O. Osibote, R. Dendere, S. Krishnan, and T. Douglas, "Automated focusing in bright-field microscopy for tuberculosis detection," *Journal of microscopy*, vol. 240, no. 2, pp. 155–163, 2010.

[OSF+18] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[PWP04] W. Pendergrass, N. Wolf, and M. Poot, "Efficacy of mitotracker green™ and cmxrosamine to measure changes in mitochondrial membrane potentials in living cells and tissues," *Cytometry Part A: the journal of the International Society for Analytical Cytology*, vol. 61, no. 2, pp. 162–169, 2004.

[PX23] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[QK20] X. Qian and D. Klabjan, "The impact of the mini-batch size on the variance of gradients in stochastic gradient descent," *arXiv preprint arXiv:2004.13146*, 2020.

[RAGS01] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[RBL+22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[RCK+08] J. Riedl, A. H. Crevenna, K. Kessenbrock, J. H. Yu, D. Neukirchen, M. Bista, F. Bradke, D. Jenne, T. A. Holak, Z. Werb *et al.*, "Lifeact: a versatile marker to visualize f-actin," *Nature methods*, vol. 5, no. 7, pp. 605–607, 2008.

[RFB15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, vol. abs/1505.04597, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:3719281

[RHW86a] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[RHW86b] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986. [Online]. Available: https://api.semanticscholar.org/CorpusID:205001834

[RJ+01] A. C. Ruifrok, D. A. Johnston *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.

[RLK19] S. Roy, S. Lal, and J. R. Kini, "Novel color normalization method for hematoxylin & eosin stained histopathology images," *IEEE Access*, vol. 7, pp. 28 982–28 998, 2019.

[Rob51] H. E. Robbins, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951. [Online]. Available: https://api.semanticscholar.org/CorpusID:16945044

[Ros58] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[Rud16] S. Ruder, "An overview of gradient descent optimization algorithms," *ArXiv*, vol. abs/1609.04747, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:17485266

[RWC+19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[RWW+19] Y. Rivenson, H. Wang, Z. Wei, K. de Haan, Y. Zhang, Y. Wu, H. Günaydın, J. E. Zuckerman, T. Chong, A. E. Sisk *et al.*, "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning," *Nature biomedical engineering*, vol. 3, no. 6, pp. 466–477, 2019.

[RZL17] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: a self-gated activation function," *arXiv: Neural and Evolutionary Computing*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:196158220

[SC20] P. Salehi and A. Chalechale, "Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis," in *2020 International conference on machine vision and image processing (MVIP)*. IEEE, 2020, pp. 1–7.

[SDN04] Y. Sun, S. Duthaler, and B. J. Nelson, "Autofocusing in computer microscopy: selecting the optimal focus algorithm," *Microscopy research and technique*, vol. 65, no. 3, pp. 139–149, 2004.

[SME20] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[Swi62] D. F. Swinehart, "The beer-lambert law," *Journal of chemical education*, vol. 39, no. 7, p. 333, 1962.

[Tit05] M. Titford, "The long history of hematoxylin," *Biotechnic & histochemistry*, vol. 80, no. 2, pp. 73–78, 2005.

[Tit09] M. Titford, "Progress in the development of microscopical techniques for diagnostic pathology," *Journal of histotechnology*, vol. 32, no. 1, pp. 9–19, 2009.

[TL05] M. Tan and Q. E. Le, "Rethinking model scaling for convolutional neural networks. arxiv 2019," *arXiv preprint arXiv:1905.11946*, 1905.

[TNC+20] W. C. C. Tan, S. N. Nerurkar, H. Y. Cai, H. H. M. Ng, D. Wu, Y. T. F. Wee, J. C. T. Lim, J. Yeong, and T. K. H. Lim, "Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy," *Cancer Communications*, vol. 40, no. 4, pp. 135–153, 2020.

[VSK21] S. Vijh, M. Saraswat, and S. Kumar, "A new complete color normalization method for h&e stained histopatholgical images," *Applied Intelligence*, vol. 51, no. 11, pp. 7735–7748, 2021.

[WCW+07] Y.-Y. Wang, S.-C. Chang, L.-W. Wu, S.-T. Tsai, and Y.-N. Sun, "A color-based approach for automated segmentation in tumor tissue classification," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 6576–6579.

[WDB+79] E. Wulf, A. Deboben, F. Bautz, H. Faulstich, and T. Wieland, "Fluorescent phallotoxin, a tool for the visualization of cellular actin." *Proceedings of the national academy of sciences*, vol. 76, no. 9, pp. 4498–4502, 1979.

[WHT+23] H. Wang, G. Han, H. Tang, R. Zhang, Z. Liu, Y. Sun, B. Liu, J. Geng, and Z. Zhang, "Synchronous photoactivation-imaging fluorophores break limitations of photobleaching and phototoxicity in live-cell microscopy." *Analytical chemistry*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264543153

[WLC+19] Y. Wu, Y. Luo, G. Chaudhari, Y. Rivenson, A. Calis, K. De Haan, and A. Ozcan, "Bright-field holography: cross-modality deep learning enables snapshot 3d imaging with bright-field contrast using a single hologram," *Light: Science & Applications*, vol. 8, no. 1, p. 25, 2019.

[WRJ+19] H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydın, L. A. Bentolila, C. Kural, and A. Ozcan, "Deep learning enables cross-modality super-resolution in fluorescence microscopy," *Nature methods*, vol. 16, no. 1, pp. 103–110, 2019.

[XLD+20] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, "Reluplex made more practical: Leaky relu," in *2020 IEEE Symposium on Computers and communications (ISCC)*. IEEE, 2020, pp. 1–7.

[Zer42] F. Zernike, "Phase contrast, a new method for the microscopic observation of transparent objects," *Physica*, vol. 9, no. 7, pp. 686–698, 1942.

[ZFT+24] I. Zingman, S. Frayle, I. Tankoyeu, S. Sukhanov, and F. Heinemann, "A comparative evaluation of image-to-image translation methods for stain transfer in histopathology," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1509–1525.

[ZIE+18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[ZJZ+18] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, and Y. Zhao, "Histopathological whole slide image analysis using context-based cbir," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1641–1652, 2018.

[ZLL+16] L. Zhong, L. Lin, Z. Lu, Y. Wu, Z. Lu, M. Huang, W. Yang, and Q. Feng, "Predict ct image from mri data using knn-regression with learned local descriptors," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 743–746.

[ZPIE17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

Appendix

```yaml
model:
  base_learning_rate: 4.5e-6
  target: ldm.models.autoencoder.AutoencoderKL
  params:
    monitor: "val/rec_loss"
    ckpt_path: C:/Users/philmarq/source/repos/latent-diffusion/logs
       /2024-10-18T00-50-54_autoencoder_kl_32x32x4/checkpoints/last.ckpt
    embed_dim: 4
    lossconfig:
      target: ldm.modules.losses.LPIPSWithDiscriminator
      params:
        disc_start: 25001
        kl_weight: 0.000001
        disc_weight: 0.5

    ddconfig:
      double_z: True
      z_channels: 4
      resolution: 256
      in_channels: 3
      out_ch: 3
      ch: 128
      ch_mult: [1, 2, 4, 4]  # num_down = len(ch_mult)-1
      num_res_blocks: 2
      attn_resolutions: [ ]
      dropout: 0.0

data:
  target: main.DataModuleFromConfig
  params:
    batch_size: 4
    wrap: True
    train:
      target: ldm.data.imagenet.CustomDataset
      params:
        roots:
```

```
          - "C:/Users/philmarq/Downloads/TrainingPhases/unstained/H&E␣
              stained"

    validation:
      target: ldm.data.imagenet.CustomDatasetValidation
      params:
        roots:
          - "C:/Users/philmarq/Downloads/TrainingPhases/BCI_dataset/
              BCI_dataset/HE/val/"

trainer:
  benchmark: True
  accumulate_grad_batches: 6
```