Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies

FIIT-182905-103157

Bc. Tomáš Tánczos

Augmentation of histopathology dataset by methods of generative neural networks

Master thesis

Thesis supervisor: prof. Ing. Vanda Benešová, PhD.

May 2025

Slovak University of Technology in Bratislava

Faculty of Informatics and Information Technologies

FIIT-182905-103157

Bc. Tomáš Tánczos

Augmentation of histopathology dataset by methods of generative neural networks

Master thesis

Study programme: Intelligent Software Systems

Study field: Computer Science

Institute: Institute of Computer Engineering and Applied Informatics

Thesis supervisor: prof. Ing. Vanda Benešová, PhD.

May 2025

Slovak University of Technology in Bratislava
Institute of Computer Engineering and Applied Informatics

Faculty of Informatics and Information Technologies
Academic year: 2024/2025
Reg. No.: FIIT-182905-103157

MASTER THESIS TOPIC

Student: **Bc. Tomáš Tánczos**
Student's ID: 103157
Study programme: Intelligent Software Systems
Study field: Computer Science
Thesis supervisor: prof. Ing. Vanda Benešová, PhD.
Head of department: Ing. Katarína Jelemenská, PhD.

Topic: **Augmentation of histopathology dataset by methods of generative neural networks**

Language of thesis: English

Specification of Assignment:

Analýza digitálnych histologických snímok je dôležitou súčasťou liečby pacienta a výskum metód založených na hlbokom učení pri analýze digitálnych histologických snímok je vysoko aktuálnou témou. Získanie veľkého počtu kvalitne anotovaných digitalizovaných histologických dát je však náročné a môže byt limitujúcim faktorom pre výskum nových prístupov s využitím hlbokých neurónových sietí. Ako jeden z možných prístupov riešenia je tvorba synteticky generovaných snímok pre rozšírenie súboru dát. Analyzujte súčasný stav poznania oblasti generovania syntetických histologických snímok, pričom sa hlavne zamerajte na riešenia metódami hlbokého učenia. Preskúmajte najnovšie trendy a architektúry neurónových sietí pre generovanie syntetických snímok. Porovnajte možnosti ich využitia v doméne histológie. Navrhnite vlastný systém hlbokého učenia na generovanie syntetických histologických snímok, ktorý bude možné využiť pre rozšírenie súboru dát reálnych snímok. Cieľom rozšírenia súboru dát je účinnejšie trénovanie a teda zlepšenie presnosti existujúceho diagnostického systému pre zvolenú aplikáciu spracovania histologických dát. Vlastné riešenie implementujte a vyhodnoťte na reálnych dátach so správne zvolenými kvantitatívnymi a kvalitatívnymi metrikami vyhodnocovania. Následne vaše výsledky porovnajte s už existujúcimi riešeniami.

Deadline for submission of Master thesis: 11. 05. 2025
Approval of assignment of Master thesis: 15. 04. 2025
Assignment of Master thesis approved by: prof. Ing. Vanda Benešová, PhD. – Study programme supervisor

Declaration of honor

I, Tomáš Tánczos, honestly declare that I prepared this work independently, based on consultations and using the mentioned literature. I used artificial intelligence (GenAI) tools for translation, text summarization and language editing. I used these tools exclusively to support the translation, shortening and formatting of the text, but not to generate original ideas or professional content.

in Bratislava, 11.05.2025

Tomáš Tánczos

Acknowledgement

I want to express my gratitude to my supervisor, prof. Ing. Vanda Benešová, PhD. for her professional approach and advice, thanks to which this work is completed. Also, I would like to thank my family and girlfriend for their support and patience during the writing of this thesis.

Annotation

Slovak University of Technology Bratislava
Faculty of Informatics and Information Technologies
Degree Course: Intelligent Software Systems

Author: Bc. Tomáš Tánczos
Master Thesis: Augmentation of histopathology dataset by methods of generative neural networks
Supervisor: prof. Ing. Vanda Benešová, PhD.
May 2025

Analyzing digital histopathological images is essential for medical diagnostics. The aim of this research is to augment histopathological datasets using generative neural networks and evaluate the impact of synthetic data. We review the current methods for synthetic image generation and, based on our research, prioritize experimenting with diffusion networks. Our solution combines image synthesis and inpainting into one model, differing only in the inference process. We conducted experiments using in-house histopathological image datasets of heart tissue, as previous research indicated a significant underrepresentation of the blood vessel class. Our experiment was executed on whole slide images of heart tissue and breast cancer, focusing on blood vessels. We synthesized images from pixel and latent space and compared their quality. The study assesses the quality of the generated images using quantitative metrics and visual analysis. Our results suggest that synthetic data can augment histopathological datasets, and the segmentation metrics indicate an improvement in the sensitivity of the segmentation models.

Anotácia

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Študijný program: Informatika

Autor: Bc. Tomáš Tánczos
Diplomová práca: Rozšírenie histopatologického súboru údajov metódami generatívnych neurónových sietí
Vedúci diplomového projektu: prof. Ing. Vanda Benešová, PhD.
Máj 2025

Analýza digitálnych histopatologických snímok je v lekárskej diagnostike kľúčová. Táto práca sa zameriava na rozšírenie súborov histopatologických údajov pomocou generatívnych neurónových sietí a vyhodnotenie vplyvu nových údajov na segmentačné modely. Skúmame súčasné metódy generovania syntetických obrazov a porovnáva ich s tými, ktoré najlepšie spĺňajú naše požiadavky. Na základe tohto hodnotenia sme sa rozhodli uprednostniť denoizujúce pravdepodobnostné modely difúzie pred generatívnymi adverznými sieťami. Vzhľadom na povahu procesov syntézy a inpaintingu obrazu naše riešenie kombinuje tieto dva procesy a využíva ich potenciál pri rozširovaní súboru údajov. Navrhované riešenie experimentuje na súboroch histopatologických obrazových údajov srdcového tkaniva, pretože počas predchádzajúceho výskumu sa ukázalo, že trieda ciev je výrazne nedostatočne zastúpená. Experimentovali sme so syntézou obrazu v pixelovom a latentnom priestore a tiež s veľkosťou latentného priestoru. V práci sa hodnotí kvalita vygenerovaných obrazov pomocou kvantitatívnych metrík a vizuálnej analýzy. Zistenia naznačujú, že generatívne modely môžu rozšíriť súbory histopatologických údajov a segmentačné metriky naznačujú zlepšenie citlivosti segmentačného modelu.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Digital histopathology is essential in patient diagnosis and treatment. Analyzing such digitized images is a time-consuming task, and the application of deep learning solutions to that is a timely topic. The main challenge in that field is collecting a significant amount of annotated histology data for the network's training.

Various methods make creating new datasets in this field challenging. The required data are not publicly available. However, they contain sensitive information about patient privacy, which complicates making them public. The second problem is that annotating such an image is also time-consuming and can be done only by an expert pathologist. Lastly, the amount of histology data is limited by the number of patients for a given disease. One potential solution is the creation of synthetic images to augment the dataset.

Our research focused on methods where generative networks are applied and analyzed. Finally, our solution examines various image synthesis approaches to create synthetic data with the goal of augmenting an existing dataset. Later, the augmented dataset was used as a training set for the segmentation model.

1.1 Motivation and goal

Our in-house heart tissue dataset (Section 6.2.1) contains a small amount of blood vessels annotation, which made it difficult to develop a segmentation solution in previous works [11], where the authors focused on the segmentation of various structures in heart tissue. The ultimate goal of the work is to improve the quantitative metrics for blood vessel segmentation by augmenting the dataset with synthetic data.

To achieve this goal, we prepared a two-step solution. In the first step, we will focus on image synthesis. The synthetic images will be created in two ways: fully synthetic images and partially edited existing ones. During the second step, we will systematically augment the existing dataset, and for every step of augmentation, we will train a segmentation process to examine the effect of new data on the models' behavior. We want to understand whether the augmentation effectively assists the model in learning a better representation of the features.

Chapter 2

Medical Imaging

Medical imaging allows us to examine human internal and external body parts closely. Its advantage is that, in most cases, it is possible without any invasion proceedings. They are one of the most important sources of information for healthcare workers since medical images make up 90% of every healthcare data. Imaging is often an important part of the patient's diagnosis, treatment, and surgery operations, where real-time imaging is utilized to help the process. The management of medical images is not simple because of the wide variety of medical modalities standards and patient privacies. Therefore, many image sources are scattered between hospitals and imaging centers, and there is a lack of centralized image data centers [5]. In medical modalities, we can differentiate computed tomography (CT), X-ray radiography, and digital pathology. This work will focus on digital histopathology, a subset of the pathology field [39].

2.1 Digital Histopathology

Histopathology is a specialized branch of pathology where the specialist visually examines the extracted tissue under the microscope. Usually, the tissue is placed on a glass slide and is inspected by an expert pathologist. With the advent of digitalization, tissue slides can now be scanned, creating whole slide images (WSI) that are stored in digital form. Figure 2.1 shows an example of a WSI.

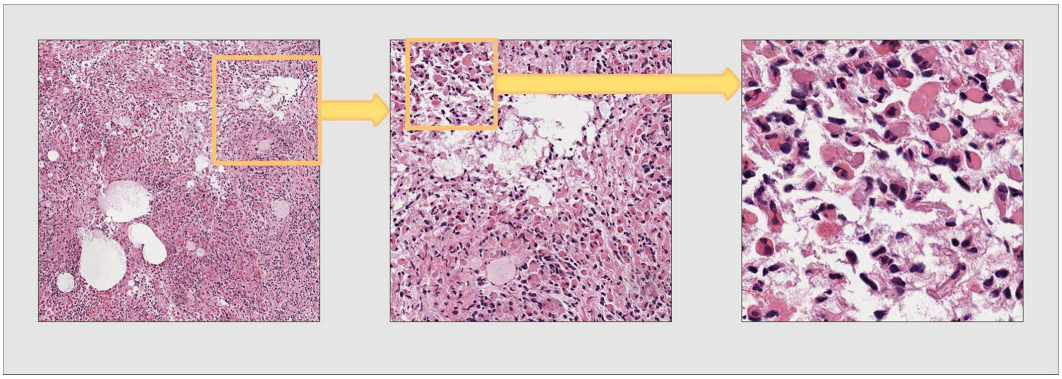

Figure 2.1: Sample of WSI image at various magnifications [5]

The WSI contains meaningful structural and pathological information, crucial in predicting a patient's future medical treatment. They are very detailed and have multiple levels of magnification. In [39], authors say that one WSI image on a magnification level of 40x can take around 10GB of disk space. This detailing allows experts to perform an in-depth image analysis, which was previously impossible, and can create more precise outcomes for the patient's condition [5].

2.2 Common Assignments in Digital Histology

Computer vision (CV) has many applications in medical imaging, like image reconstruction, enhancement, or registration. In the case of digital histology and WSI, the main task is usually the segmentation or classification of cells and bio-

logical structures. The union of these tasks is the semantic image segmentation, which target to segmenting and classifying the input part image such asarmacy image below. This is a crucial requirement in medical imaging, since we usually have small regions of interest (eg. tumors or lesion) as well as the background and it will usually bring about severe class-imbalance. A more complex task is an instance segmentation where We want to distinguish every instance in the given class. An example of such tasks is visible in Figure 2.2; in semantic segmentation, we want to find the chairs, but we want to count them during the instance segmentation.

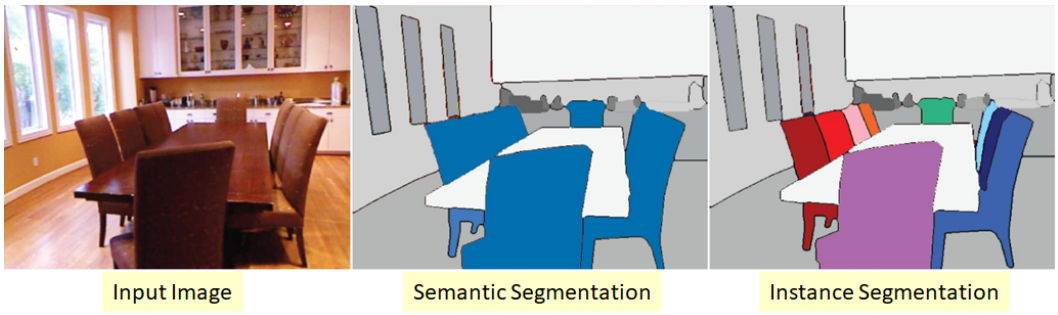

Figure 2.2: Comparing of semantic and instance segmentation [10]

2.3 The Role of Deep Learning

This work focuses on deep learning (DL) approaches and tasks they can perform, mainly in image synthesis. Thanks to the ability to simulate and learn complex patterns, it can perform its task accurately and comparably to an expert pathologist. One of the first applications of deep learning was detecting and segmenting a single nuclein in images. For example, it can distinguish healthy nuclein from cancer nucleic in an image of breast cancer [39]. Another complex task is disease grading, which is a crucial part of the analysis of cancer images. This grading system determines the severity of illness. In [8], a deep learning system was de-

veloped to grade prostate cancer images based on the Gleason grading standard. In these cases, the result is based on multiple variables, which depend on different aspects of the images [39]. These two approaches were just a small example of tasks that deep learning methods can execute. With the progress of development and growing computational resources, this set expands with more complex tasks. In recent years, a new trend has emerged - image synthesis. Thanks to various deep-learning architectures, high-quality image synthesis has become achievable [34]. This trend has also impacted histology images, leading to the publication of papers [14, 26, 38, 4, 2] on generating realistic samples.

2.4 Challenges in Digital Histopathology

Although many deep learning solutions are available in digital histopathology, we still face challenges in developing deep learning-based solutions, mainly because of the properties of WSIs. This section is aimed to introduce them.

The WSIs can have thousands of pixels in each dimension, but deep convolutional neural networks usually do not have input spatial dimensions in this size; if so, we need enormous computational power and a deeper network topology, making it even harder to train the model. Patching is used to address this problem in most cases. Patching divides the image into smaller tiles to fit them into the network. However, with this technique, we usually need to downsample the images, which leads to the loss of relevant information, and splitting the image can result in the loss of spatial information [5, 31].

The next problem arises when the neural networks require a large set of well-labeled training data, mainly the supervised ones. In our cases, the labeled training data are called annotations, and for the correct annotation, the expertise of a pathologist is required. The annotation process is complex and lengthy; therefore, the amount

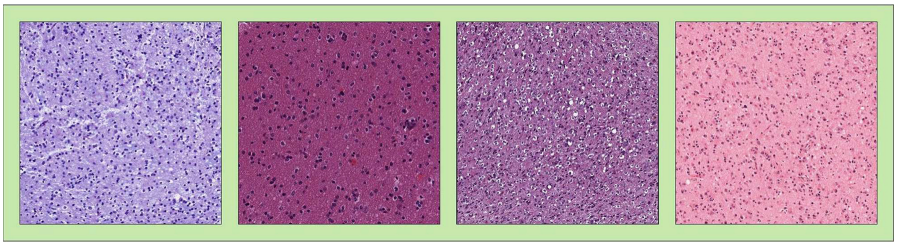

Figure 2.3: Color variation of stained brain tissue [5]

of well-annotated histopathological images can be one of the main bottlenecks in developing further solutions [5, 31].

The high variability in the nature and color of the images can also be challenging. Several biological structures have different patterns and different cell arrangements that have to be recognized by the computer on multiple levels of magnification. A usual step during the analysis of the sample is staining it with a stained reagent, which does not always result in the same colors. An example of this final color variation of a slice of brain tissue is in Figure 2.3. The result can also be affected by the type of medical scanning modalities, illuminance during scanning, and or tissue thickness. Color-connected issues are usually solved with a color normalization approach [5, 31].

Chapter 3

Computer Vision Using Deep Learning

Computer vision (CV) is a field that enables computers to interpret and process visual data, such as images and videos. The results of processing such data are decision-making or providing insight into tasks like object detection, image classification, or segmentation. Traditional CV tasks can be split into two main parts. The first part is feature extraction, which is done manually with algorithms like Scale-Invariant Feature Transform (SIFT) or Speeded Up Robust Feature (SURF)[1]. After this, the hand-crafted features are used to train machine learning models like Support Vector Machines, Decision Trees, or K-means clustering. Traditional methods often require extensive manual effort and expertise in feature extraction, which is being overcome by deep learning-based approaches. DL methods combine the two stages into a single process, with its most significant advantage being unsupervised feature extraction. With the application of these

[1]https://mikhail-kennerley.medium.com/a-comparison-of-sift-surf-and-orb-on-opencv-59119b9ec3d0

methods, there is no need for experts who know how to extract special features like edges, colors, and shapes from images since it all happens during the training of deep neural networks. The drawback of this approach is that the features that are learned by the network are not interpretable to humans [24].

3.1 Training of Deep Neural Networks

We must first understand the building blocks of the deep neural network (DNN) to understand how it trains. The most basic unit of a DNN is the single neuron, which is nothing other than a simple mathematical function that calculates the weighted sum of its input variables and adds bias to them. The Equation 3.1 describes this function, where x_i is the i^{th} input of the given neuron and w_i is the coressponding weight. Figure 3.1 illustrates how it could look like a simple multi-layer perceptron, with two hidden layers and one output neuron, also called output layer [1].

$$z = \sum_{i=1}^{n} (x_i \cdot w_i) + b \tag{3.1}$$

The main issue with a network like this is that it is just a composition of several linear functions, so it cannot learn non-linear representations. To address this issue, usually on the hidden layer, the neurons are wrapped into a non-linear activation function like ReLU (Equation 3.2) or LeakyReLU (Equation 3.3), but we can freely experiment and change the activation functions according to our goals. The task for the activation function is to manage, if the signal from a corresponding neuron should propagate to the next layer or not [1].

$$f(x) = \max(0, x) \tag{3.2}$$

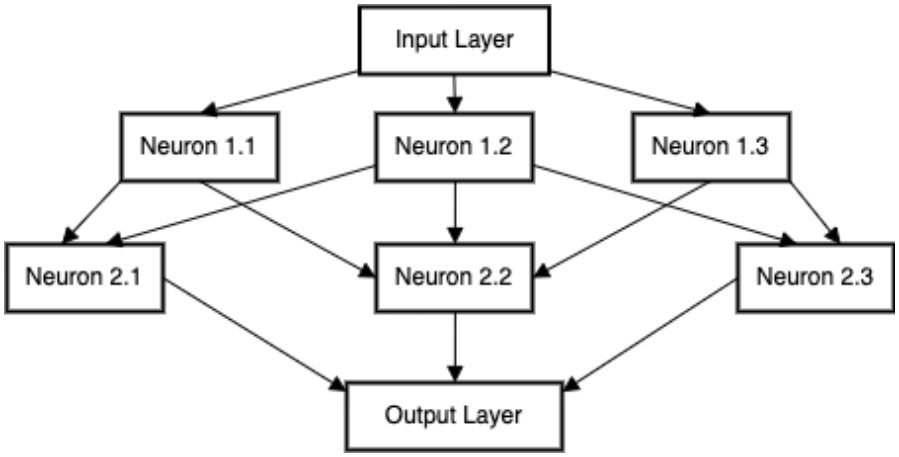

Figure 3.1: Architecture of a simple multi-layer perceptron

$$f(x) = \max(0.01x, x) \tag{3.3}$$

In most cases, the Sigmoid (Equation 3.4) function is used on the last layer of the DNN because of its ability to convert the input value into a range of $[0, 1]$, which can be interpreted as the probability of a given event in the task of binary classification.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.4}$$

In the context of training of DNN, we require three steps: the feed-forward process, when the input values propagate through the network, the backward phase, to calculate the gradient of each weight regarding the error of our cost function; Furthermore, the last step is updating weights and biases with a given learning rate. These steps aim to minimize the error of our cost function Θ, which can vary based on our goals [1].

3.2 Convolutional Neural Networks

Convolutional neural networks (CNN) are DNN types suited to work with data with strong spatial dependencies. Although CNN is usually used to process images or videos, it is also applicable to work with music, text, or sequential data if we consider it a special grid structure. His name is derived from the mathematical operation convolution, which is used inside of these networks. The Equation 3.5 presented the discrete convolution between two two-dimensional data; in our cases, I is our input signal, the image and K is the kernel. In convolutional layers, the kernel values are the equivalent of the weights from conventional DNN layers. The inspiration for CNN comes from the experiments on the visual cortex of cats [18], where the researchers discovered that the visual cortex works a layered principle, and different abstractions of spatial information are processed and recognized on different layers. In applying convolutional layers, this means that on the first layers, the kernels recognize simple shapes or edges, and the deeper layer can identify much more complicated structures, like faces.

$$(I * K)[x, y] = \sum_{i=-m}^{m} \sum_{j=-n}^{n} I[x - i, y - j]K[i, j] \tag{3.5}$$

A practical example of the convolution by kernel 3×3 on the image with one channel is visible in Figure 3.2. In the convolution, the spatial size of the image can be regulated by the padding and stride in the calculation. During the padding, we add values to the image's border; it can be zero or any other value that fits our case. Stride defines the number of pixels by how much to shift the kernel [20].

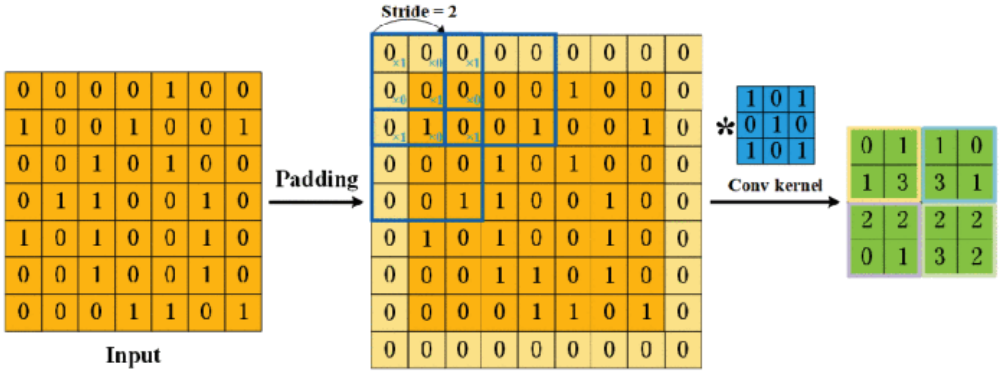

Figure 3.2: Convolution operation with kernel of size 3×3, padding $= 1$ and stride $= 2$ [20]

3.3 Image Segmentation

One of the tasks that has become a fundamental of CV and is used in robotic perception, augmented reality, and medical image analysis is segmentation of images. Image segmentation can be defined as assigning a categorical label to each pixel of an image (semantic segmentation) or assigning a distinct label to each object (instance segmentation). Many image segmentation methods have been developed over the years which employ different approaches such as thresholding, region-growing, and active contours. But, since images and objects in images are heterogeneous, these methods met limitation and need strong tuning to obtain effects. DNN-based approaches significantly outperform traditional techniques, enabling more robust performance. In the next section, we will discuss the DL approaches for image segmentation [22].

3.3.1 Segmentation Methods Using Deep Learning

Several approaches and architectures have also been developed for deep learning methods, like encoder-decoder or fully convolutional models. However, we want to discuss the U-Net architecture, which became state-of-the-art in image segmenta-

tion. This model was initially developed for medical image segmentation in 2015 by Ronneberger et al., but it has since been widely adopted in various domains that require image segmentation [30, 22]. The U-Net architecture stands of three main parts 3.3:

1. **Contracting Path:** The contracting path (encoder) captures contextual information. It is built from repeated applications of convolutional layers and max-pooling, which gradually reduce the image's spatial dimensions and preserve the most significant features. This part extracts high-level semantic features from the input image.

2. **Expanding Path:** The expanding path (decoder) reconstructs the segmentation map by up-sampling (typically by transposed convolution or by interpolation methods) feature maps. It combines features from the contracting path via skip connections to keep spatial details.

3. **Skip Connections:** Skip connections concat corresponding layers in the contracting and expanding paths. These connections mitigate the loss of spatial information and allow the model to learn the spatial details of the input, ensuring precise localization during reconstruction.

3.3.2 Evaluation metrics for segmentation

In semantic segmentation, quantitative evaluation must count both the spatial overlap of predicted and ground-truth regions and the accuracy of predictions. The Dice coefficient (Equation 3.6) measures the overlap between two masks. Precision (Equation 3.7) quantifies the proportion of correctly identified positive pixels, reflecting how specifically the model avoids false alarms. Recall (Equation 3.8) measures the fraction of actual positive pixels recovered, indicating the model's sensitivity to detect every region of interest. Under-segmentation can lead

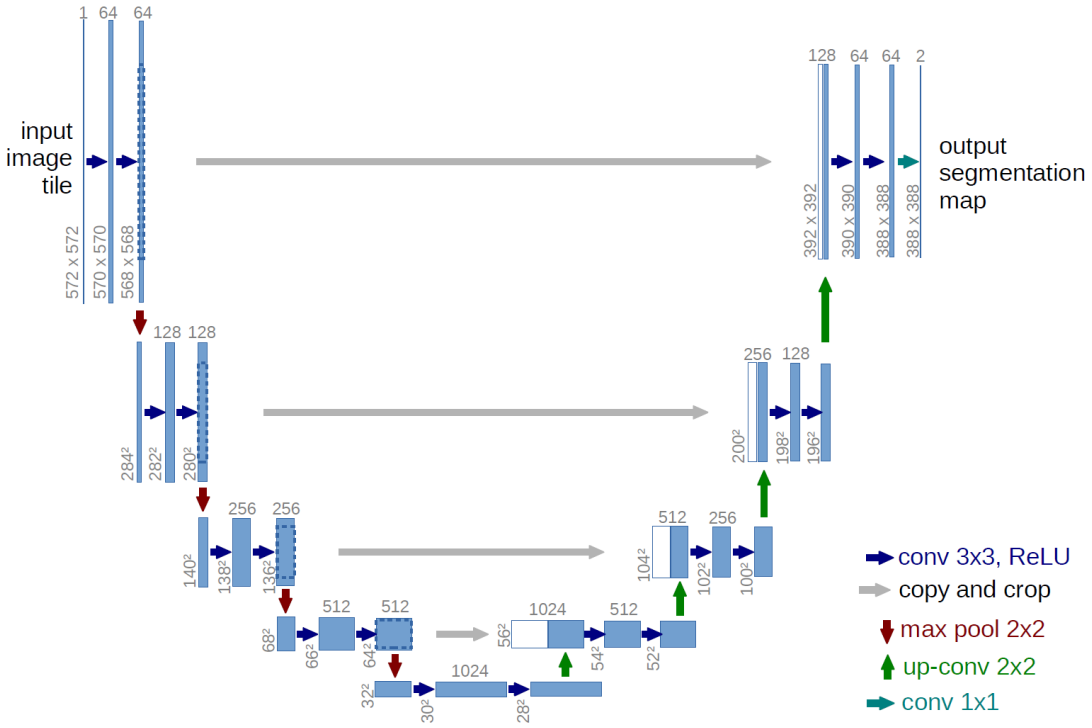

Figure 3.3: U-Net architecture [22]

to missed pathology in the medical domain, while over-segmentation can burden clinicians with incorrect findings. Therefore, reporting Dice alongside precision and recall provides a balanced assessment of both region agreement and diagnostic reliability.

$$\text{Dice} = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{3.6}$$

$$\text{Precision} = \frac{|A \cap B|}{|A|} \tag{3.7}$$

$$\text{Recall} = \frac{|A \cap B|}{|B|} \tag{3.8}$$

Chapter 4

Image Synthesis with Deep Learning

Image synthesis is a task that generates new images using computer algorithms or deep learning models. Three types of neural networks are mainly used for generation: generative adversarial networks (GAN), denoising diffusion probabilistic models (DDPM), and variational autoencoders (VAE). It is important to mention that generative models are distinguished from discriminative models in their operation. The discriminative ones try to estimate the boundary between the final classes, and the output of the discriminative model can be described mathematically as $P(Y \mid X)$, where Y is the target and X is the independent variable. The generative models try to model the distribution of particular classes, which we can express as an estimation of a distribution given by Equation 4.1 [12].

$$P(X \mid Y)P(Y) \tag{4.1}$$

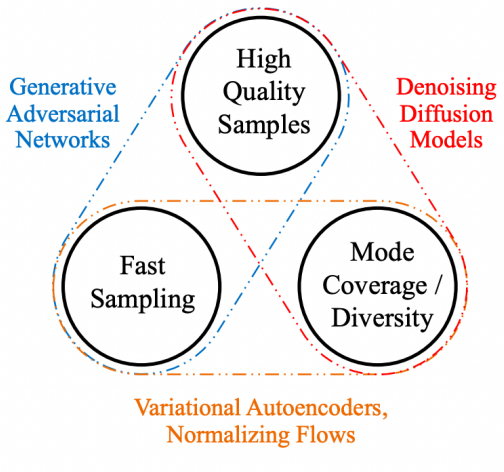

Figure 4.1: Generative learning trilemma by [36]

Image synthesis tasks have three key requirements: high-quality image, fast sampling, and diversity in generated images. These three models can not satisfy all requirements simultaneously but must depart from one to achieve the remaining two. Zhisheng Xiao et al. [36] called this phenomenon a generative learning trilemma.

The framework of generative adversarial networks, proposed by Goodfellow et al. (2014), is built from two separate models, G (generative model) and D (discriminative model). G tries to generate a realistic sample from a given data distribution. D is the second model, whose role is to determine if the given sample on its input is a real sample or came from G. The training goal is to maximize the probability that D will make a mistake. The training method for GANs is described in Figure 4.2. Generator G gets a random noise of z, from which it generates its sample $G(z)$. The discriminator gets on input x or $G(z)$, which he is supposed to distinguish. Subsequently, both are updated by the error from the loss function [33].

Variational autoencoders were introduced by Kingma and Welling in 2014. VAE

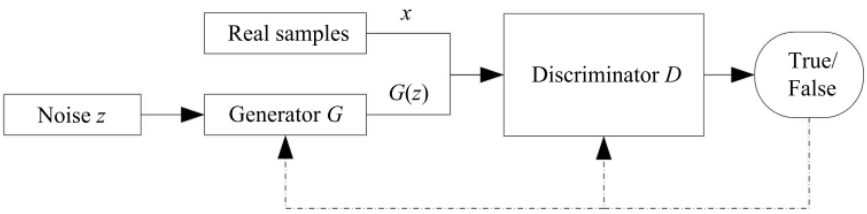

Figure 4.2: Training flow of GAN [33]

has a slight modification opposite to a classical autoencoder, allowing him to generate new samples from a given x. In VAE, the encoder part compresses the input data into a latent space, usually a Gaussian distribution. After that, the decoder takes a sample from compressed data and reconstructs it into a new sample $\hat{x}$. This approach is described in Figure 4.3 [19].

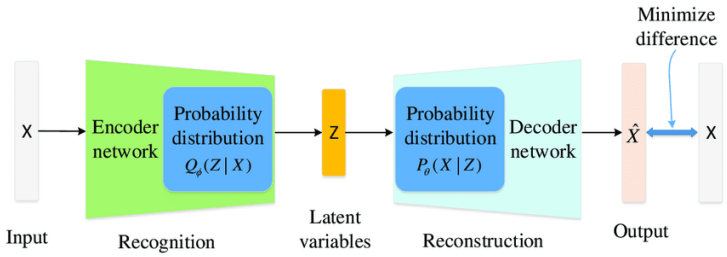

Figure 4.3: Architecture of variational autoencoder [37]

This work will focus on denoising diffusion probabilistic models that have already demonstrated their potential in producing diverse and high-quality synthesized images. Subsequent sections will discuss the possible application of DDPMs in various image synthesis tasks.

4.1 Denoising Diffusion Probabilistic Models

The denoising diffusion probabilistic model is a generative model proposed by Ho et al. [16]. Its core idea is built on the diffusion process, which we can divide into

forward and reverse diffusion; with these steps, we reduce the complex distribution of the input image, and we get an image of pure noise, and vice versa [16].

- **Forward Diffusion**: noise is systematically added to the image
- **Reverse Diffusion**: noise is systematically subtracted from the noisy image

The mentioned steps are presented in Figure 4.4., where X_0 is our input image, X_T is our noisy image after applying T step of noising. The function $q(x_t|x_{t-1})$ (Equation 4.2.) represents the Markov chain, where Gaussian noise is gradually added to the data, and $p_\theta(x_{t-1}|x_t)$ (Equation 4.3.) represents the conditional probability distribution for the denoising process.

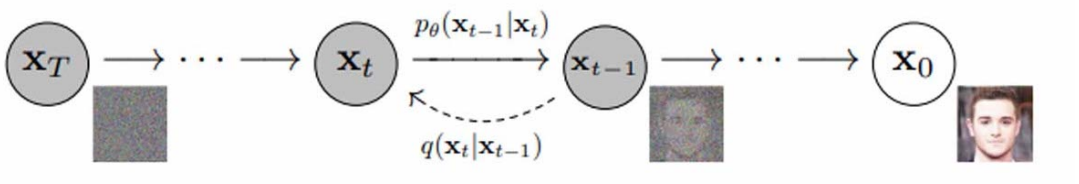

Figure 4.4: Noising and denoising process illustration [16]

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{4.2}$$

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4.3}$$

The solution is to learn how to simulate the backward process, after which the network can estimate the amount of noise in the image. It is achieved during the training process, where our model (generally a U-Net [30]) gets an image on the input with a defined noise added to it and estimates the added noise [16].

Algorithm 1 Pseudo-code for generating samples from a target distribution [16]

1: $x_T \sim \mathcal{N}(0, I)$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** x_0

At first glance, it might seem that a model generates a clear image in one step from pure noise by estimating the amount of noise, but this is inaccurate. The authors of DDPM [16] proposed an iterative sampling method (Algorithm 1.) to generate images. Where we start from a noise sampled from $\mathcal{N}(0, I)$, and for every iteration is subtracted part of the noise estimated by our model ϵ_θ is. Also, during every iteration, some noise z scaled by a scaling factor σ_t is added back to the image to improve diversity.

4.2 Latent Diffusion Models

Conventional DDPM networks require significant computational and time resources because they operate at the pixel level. As a solution, Rombach et al. [29] proposed the latent diffusion models (LDM) in 2022. The key point of LDMs is that the diffusion process is performed in a lower-dimensional latent space, significantly reducing the computational cost for training and inference time. By working in latent space, LDMs require fewer resources to generate high-resolution images (e.g., 512×512 or larger) and achieve faster inference times than pixel-based models. Another advantage of LDMs is their ability to include conditioning mechanisms, like text or image prompts, through cross-attention layers. These layers enable text-to-image synthesis, image inpainting, super-resolution, and style transfer applications.

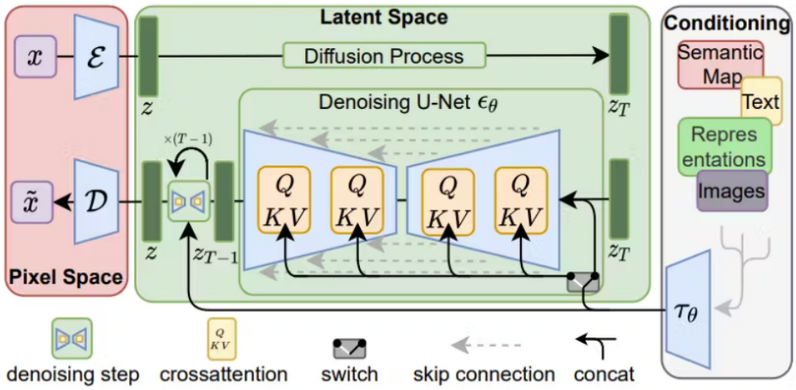

Figure 4.5: Latent diffusion model architecture [29]

Figure 4.5 shows the conceptual diagram of the LDM architecture. Here, $\mathscr{E}$ represents the encoder, and $\mathscr{D}$ is the decoder. In the middle of the figure, we see the diffusion process and the denoising step performed by ϵ_θ. The concatenations illustrate the conditioning mechanism using τ_θ. The training of such architecture can be split into two stages.

- Training of an autoencoder that maps the input image from pixel into latent space and back. The authors mentioned using a perceptual compression model with the following loss functions: Reconstruction Loss, Perceptual Loss, and KL-Divergence Loss.

- The second phase is similar to conventional DDPM training, with the key difference that LDM learns to estimate the noise in the latent space instead of the pixel space.

Finally, the LDM's inference is performed entirely in the latent space. After the final step of the denoising process, the latent representation is reconstructed back into the pixel space [29].

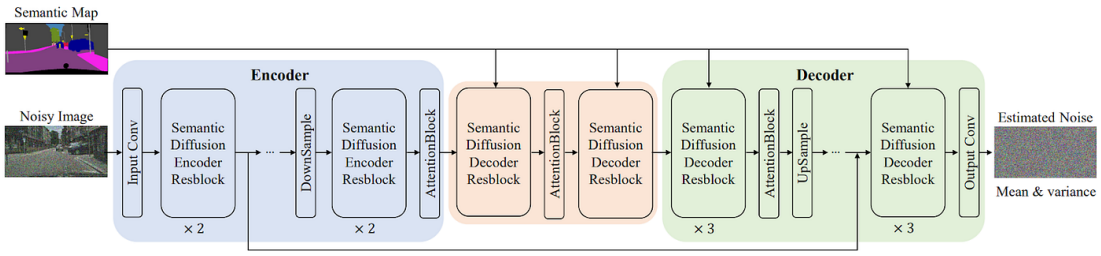

Figure 4.6: Architecture proposed by [35] for semantic image synthesis

4.3 Semantic Image Synthesis

Semantic image synthesis is the reverse equivalent of semantic image segmentation; we have a semantic mask for the task and want to generate a suitable image for the given mask. Synthetic data generated with semantic image synthesis could be used in supervised training because we obtain the particular annotation to the image sample. Wang et al. [35] bring a new and different approach to semantic image synthesis; till now, in most cases, the semantic mask was concatenated with the noisy image and directly passed to the network. The problem with this approach was that the semantic information could not be thoroughly exploited. In their approach, they embed the semantic map into the decoder part of the network with a multi-layer spatially-adaptive normalization operator (SPADE) [27]. Their architecture is shown in Figure 4.6 follows the state-of-the-art U-net shape and is built from multiple attention and residual blocks [13] using the SiLU activation function (Equation 4.4) and group normalization. SiLU [28] has the potential to overcome ReLU in a deeper model. These residual blocks are further modified to allow timestep embedding, and the decoder is extended with SPADE.

$$f(x) = x * sigmoid(x) \tag{4.4}$$

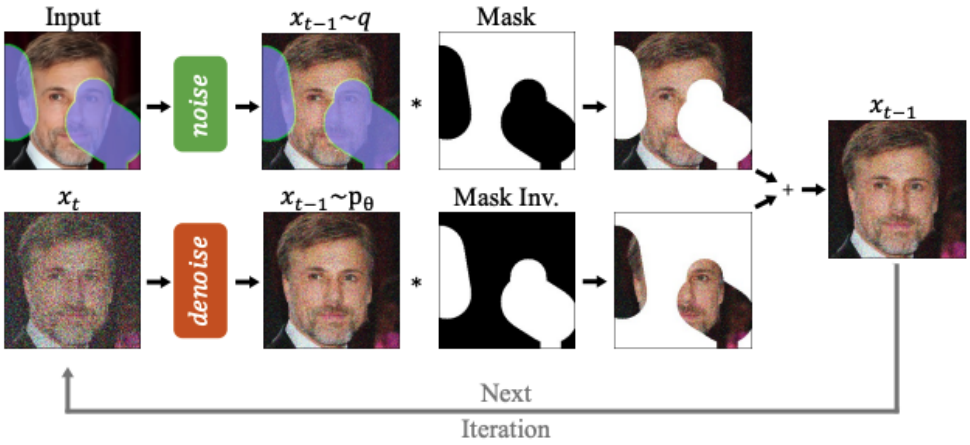

Figure 4.7: Inference stage of RePaint [21]

4.4 Image Inpainting

Image inpainting is a special case for image synthesis; inpainting aims to modify just a part of the image, for example, parts highlighted by the users, and the other details of the images stay unmodified. In the context of DDPM, Lugmayr, Andreas, et al. [21] proposed an image inpainting approach called RePaint. With this approach, they keep the part of the image without any changes and generate diverse samples for the masked region. Figure 4.7 shows how Lugmayr, Andreas, et al. modified the original DDPM sampling process. The upper row demonstrates that they sample and mask the noise version of the input image after $t-1$ steps. In the row below, the usual denoising step is visible with the difference of applying the inverse mask on the estimated image. As the last step, they add the two images to get x_{t-1}. The ability to apply this technique in histology images has already been tested by Mathias Öttl et al. [14]. They used the RePainting technique to remove the artifacts, like uneven illumination or accidental folding. Their work masked the corrupted region and let the model generate it as artifact-free. As mentioned earlier, a crucial part of image processing tasks such as segmentation or detection are well-annotated datasets.

4.5 Evaluation of Image Synthesis Models

Evaluating the model's performance in image synthesis is equally important as in other tasks like classification or segmentation. In image synthesis, we can evaluate our results qualitatively and quantitatively.

4.5.1 Qualitative evaluation

The central role in qualitative evaluation belongs to human visual perception. During the evaluation, the participants see a subset of synthetic images and evaluate their diversity and quality. The usual approaches in qualitative evaluations are:

- **Surveys:** Participants rank the realism, diversity, and quality of generated images based on predefined criteria.
- **Comparative Pairwise Analysis:** Images are presented in pairs, and participants select the "better" image based on criteria.
- **Visual Inspection for Artifacts:** Evaluators analyze images for visible flaws, such as noise, blurring, inconsistent textures, or unnatural boundaries.

This evaluation method is easy to deliver until our images are not from a specific domain and do not require any professional skills, like histopathological images. In that case, we must evaluate our results with domain experts for medical image synthesis. Another drawback of qualitative evaluation is its time cost, as manual reviews require significant effort for large datasets [17, 3].

4.5.2 Quantitative evaluation

The person's intentions and individual preferences can impact the qualitative evaluation, so measuring the model's performance quantitatively using a given set of

metrics is essential. One of these metrics is Fréchet Inception Distance (FID) proposed by Heusel et al. [15], which evaluates the distance of the extracted features (usually calculated by Inception Network, trained on the ImageNet dataset) distribution of real (r) and generated (g) images with calculation multivariate Gaussian (Equation 4.5) [6].

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \tag{4.5}$$

The second widely used metric is Kernel inception distance (KID) [7] improves FID by utilizing a polynomial kernel to measure the squared Maximum Mean Discrepancy between features of real and generated samples (Equation 4.6). This non-parametric test does not assume a Gaussian distribution, only that the kernel is a good similarity measure. It also requires fewer samples than FID; in both cases, the lower the value, the better the model [6].

$$\text{KID} = \text{MMD}(f_{real}, f_{synthetic})^2 \tag{4.6}$$

Chapter 5

Related Works

In this chapter, we will explore previous research on synthetic histology image generation. Our objective is to gain a better understanding of synthetic histology image generation. We will primarily examine studies that employed denoising diffusion probabilistic models for image synthesis and data augmentation.

5.1 Artifact Restoration and Large-Scale Synthesis in Histology Images

During the manipulation of tissue slices, before they are converted into Whole Slide Imaging (WSI) formats, these slices can suffer damage due to various factors, such as folding or uneven lighting. This damage is known as artifacts, and it can complicate the analysis of the images. Removing artifacts from WSI images is important in medical imaging. Several solutions based on Generative Adversarial Networks (GANs) designed to solve this issue, these methods may change the stain style, as they generate an entire image rather than just addressing the affected areas. To address this issue, Zhenqi He et al. [14] proposed the ArtiFusion

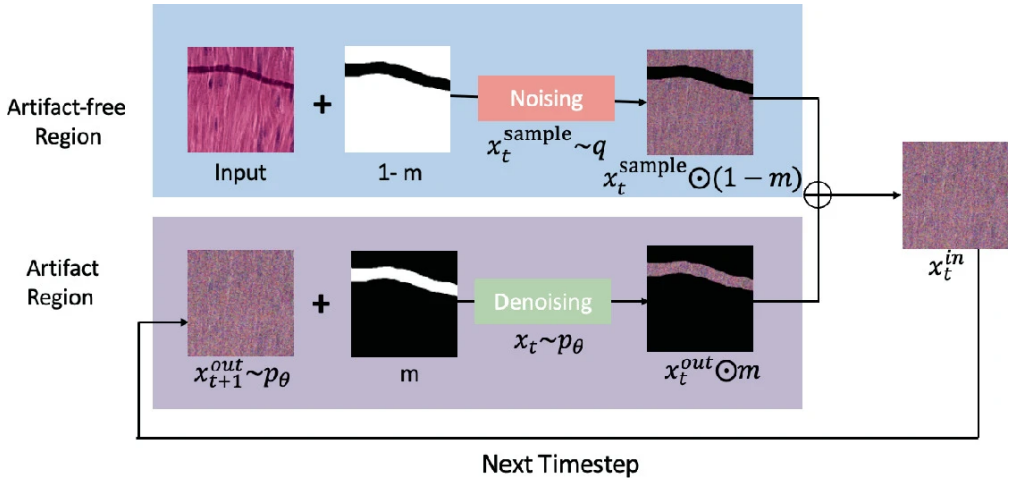

Figure 5.1: Inference stage of artifact restoration [14]

model using Denoising Diffusion Probabilistic Models (DDPM). A key aspect of their approach is that, they do not generate entire synthetic images. Instead, their model focuses on generating only specific parts of the image to replace artifacts. They used an inpainting technique inspired by Lugmayr, Andreas, et al. [21]. Another significant modification is the replacement of the U-Net architecture with a novel Swin Transformer-based network. This new network effectively utilizes attention mechanisms to better capture both local and global relationships in histology images, resulting in improved restoration quality. The inference stage of the artifact restoration process is illustrated in Figure 5.1. For the training, they sampled around 2500 images with and 2500 without artifacts from the Camelyon17 [1] dataset. The dataset images were resized to 256×256 pixels. FigureZ5.2 compares CycleGAN with Artifusion on five real-world samples with artifacts. It is visible that CycleGAN modified the style in the whole image, not just the affected part. The blue and green columns illustrate the gradual denoising process and the final restoration (green column) with Artifusion.

Marco Aversa et al. [4] introduced the DiffInfinite framework for generating large-

[1] https://camelyon17.grand-challenge.org/

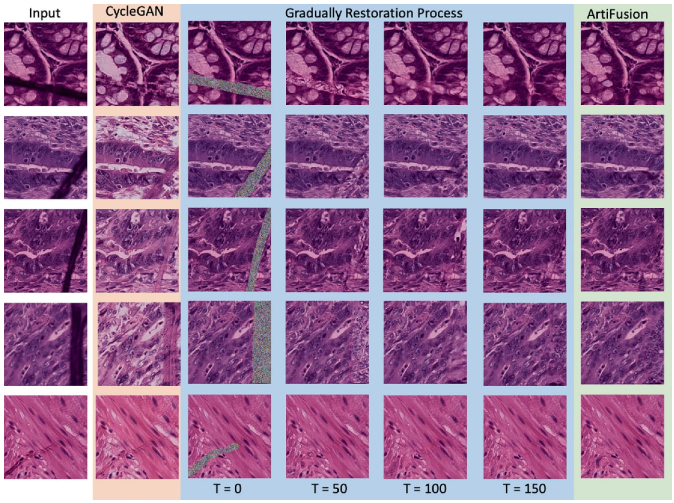

Figure 5.2: Qualitative comparison of CycleGAN and Artifusion [14]

scale synthetic histological images. The authors state that they can generate images of size 8000×8000, 16 times larger synthetic images than in the previously discussed works. Their proposal consists of 3 steps and is presented in Figure 5.3. In step *a.*), they randomly pick a smaller region from the large image (in size of 256×256, for example) and apply one step of reverse diffusion to them from x_t to x_{t-1}. In the reverse process, the generation is conditioned by the given instance mask; every mask instance is synthesized as a separate patch and is merged later. Step *c.*) is utilized as a helper step to keep track of every pixel's timestamp. The value is decreased by one after the reverse process on the pixel coordinates for the selected patch. The image generation continues in a loop until every value in the helper matrix from step *c.*) is not equal to zero. This framework employed a VAE for encoding latent space, ensuring faster sampling. Another important point of this work is that it does not use the conventional DDPM but instead uses denoising diffusion implicit models (DDIMs), providing efficient sampling for large image synthesis.

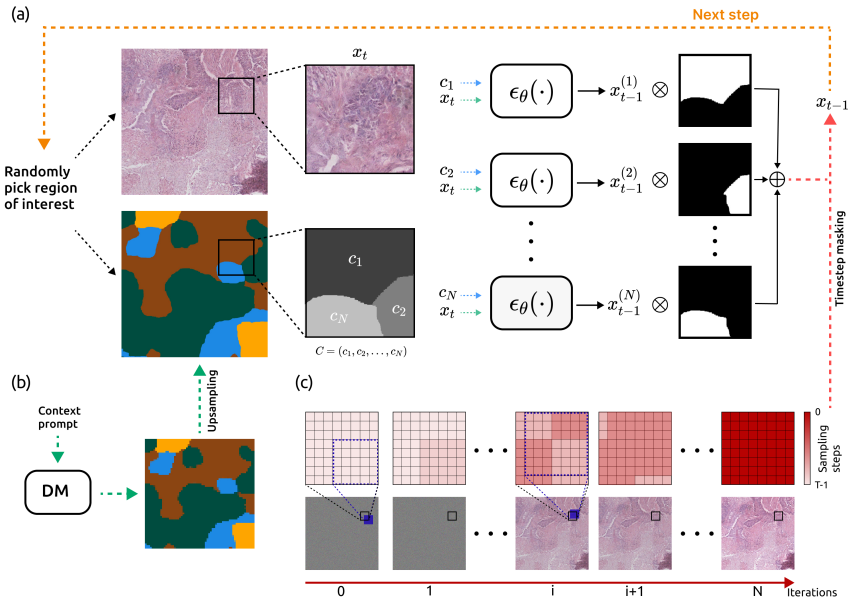

Figure 5.3: Large image synthesis multi-stage framework proposed by Aversa, Marco et al. [4]

5.2 Data Augmentation for Medical Image Segmentation

Mathias Öttl et al. [26] aimed to enhance the performance of the U-Net segmentation model by augmenting the dataset with synthetic images. Their research focused on segmenting breast cancer tumors, particularly concerning the Human Epidermal growth factor Receptor 2 (HER2) and its subtypes. HER2 comprises several subclasses, and effective treatment depends on this combination; these subtypes were imbalanced. This imbalance can lead to decreased segmentation performance. To address this issue, the authors compared three different approaches: image generation using Generative Adversarial Networks (GANs), Diffusion probabilistic models (DDPMs), and image inpainting. Their dataset augmentation pipeline is illustrated in Figure 5.4. In their setup, they split 40 whole slide images (WSIs) into training, validation, and test sets consisting of 24, 8, and 8 images,

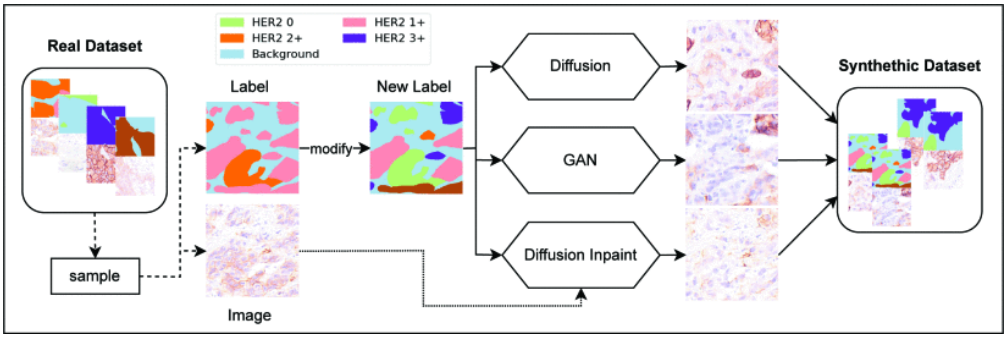

Figure 5.4: Pipeline for dataset augmentation by Mathias Öttl et al. [26]

respectively. The images generated by DDPMs beat those produced by GANs and inpainting models, achieving an improvement of 2.43

Xinyi Yu et al. [38] proposed a two-stage synthetic image generator for nuclei segmentation datasets based on Diffusion Probabilistic Models (DDPM). In the first stage, the model generates nuclei instance maps using an unconditional U-Net-based model. In the second stage, these instance maps are refined into complete synthetic nuclei images using the SPADE architecture, as inspired by Wang et al. [35]. This framework is illustrated in Figure 5.5. The dataset included images that were 1000×1000 pixels in size, and experiments with the Hover-Net and PFF-Net segmentation models demonstrated that augmenting the dataset with just 10% synthetic data significantly improved segmentation performance.

5.3 Comparing Generative Models for Medical Image Synthesis

Marco Aversa et al. [23] compared latent diffusion models (LDMs) and GANs for medical image synthesis across modalities like MRI, CT, and histopathology. Using metrics such as Fréchet Inception Distance (FID) and Structural Similarity Index Measure (SSIM), they concluded that LDMs outperformed GANs in image

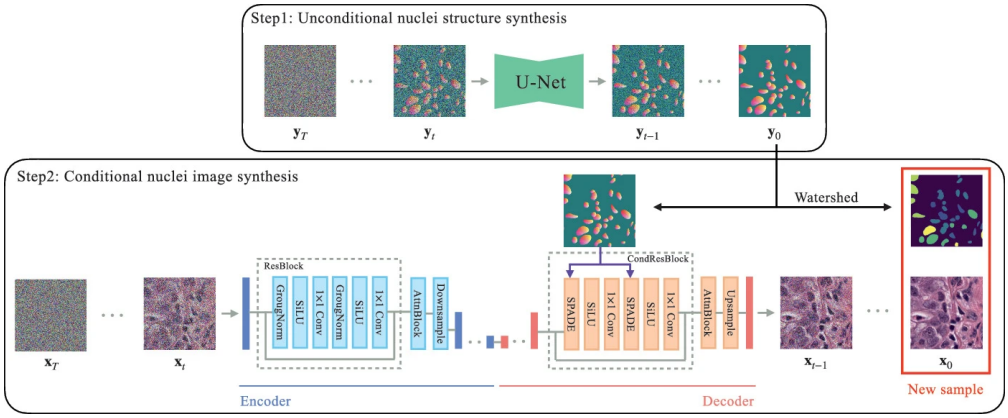

Figure 5.5: Two-stage synthetic nuclei image generator by Xinyi Yu et al. [38]

diversity and fidelity, especially in complex anatomical structures. While GANs often produced repetitive patterns and artifacts, LDMs generated artifact-free and realistic images, highlighting their suitability for data augmentation.

Similarly, Alimanov et al. [2] synthesized retinal images for segmentation using a multistep DDPM-based approach. Their pipeline included binary vessel tree masks as guidance, followed by a super-resolution network to upscale the generated images to 512×512. The U-Net backbone for vessel tree generation used advanced activation functions such as GELU and SiLU, and incorporated Vision Transformer (ViT) blocks for better feature learning. Figure 5.6 presents their architecture.

5.4 Conclusion of related works

To summarise our related works [14, 26, 38, 4, 2], the denoising diffusion probabilistic model and its variations show high potential in medical imaging. They can enhance the robustness of discriminative models by augmenting their training dataset with synthetic images. While some of the studies refer to the GANs as state-of-the-art solutions in image synthesis, it is observed that DDPMs and their

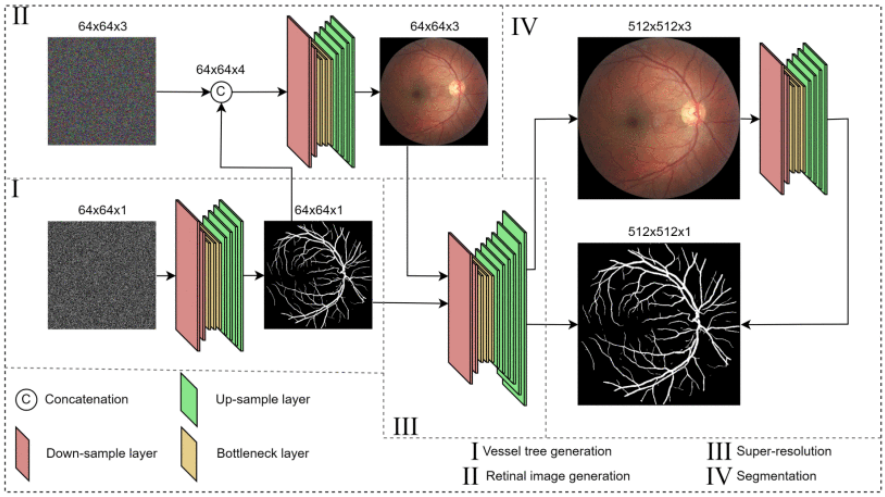

Figure 5.6: Composition of four deep learning models for retinal image dataset generation by Alimanov et al. [2]

variant can overperform in image diversity, a crucial point in training deep neural networks. An important observation is an ability to synthesize the images in higher spatial dimensions, reducing the time for acquiring synthetic datasets with relevant amounts of data but increasing our hardware requirements.

Chapter 6

Our Solution

This chapter presents the solutions developed to achieve the goals outlined in Section 1.1. Firstly, we will introduce a high-level overview of our solution. Next, we will describe the datasets used for experiments and continue with a detailed walkthrough of our model. Lastly, we will present the experimental setup and results. Based on the related works discussed in Section 5, we hypothesize that this solution enables us to test a multiple approach of dataset augmentation with synthetic images, which will enhance the segmentation models' metrics to learn better the representation of insufficient classes noted in Section 1.1.

6.1 High Level Overview of the Solution

Figure 6.1 illustrates the pipeline of our work. At the start, we have our initial dataset, which is used to train our DDPM network. This network will be used for semantic image synthesis and image inpainting; for both processes, one trained model can be used. Here, we experimented with training a DDPM model to sample in pixel and also in latent space, which will be discussed later in Section 6.3. With

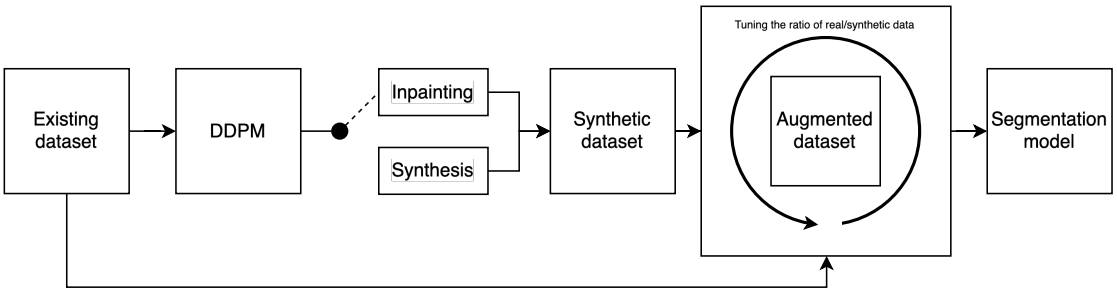

Figure 6.1: Overview of the dataset augmentation process using image synthesis and inpainting. Synthetic data is generated and merged with the original dataset to train a segmentation model.

the synthesis model, we sampled the synthetic dataset. In the augmented dataset we experimented with various ratio combination of real and synthetic data. Later, this augmented dataset was used to train the segmentation model.

6.2 Used Datasets

During our research we work with WSI data obtained from the Institute for Clinical and Experimental Medicine (IKEM) in Prague. However, we also intend to test our approach on another dataset collected from a public challenge, ICIAR 2018, which focuses on breast cancer histology images.[1]

6.2.1 IKEM - Heart Tissue Dataset

The images provided by IKEM consist of heart tissue biopsies taken after heart transplantation (see Figure 6.2). These images display various higher-level biological structures, including blood vessels, areas of inflammation, and the endocardium. We have access to 51 images, either fully or partially annotated, each with a resolution of approximately $10,000 \times 10,000$ pixels. In addition to these

[1]ICIAR 2018 Challenge: `https://iciar2018-challenge.grand-challenge.org/Home/`

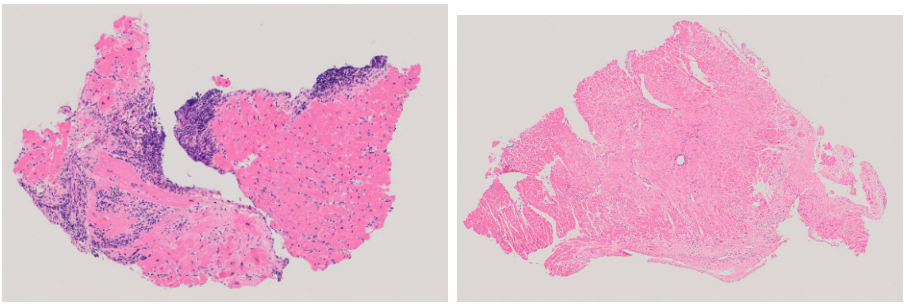

Figure 6.2: Sample whole slide images provided by IKEM

images, annotations of the biological structures are available in GeoJSON format. The following classes have been annotated: endocardium (31.22%), inflammation (38.40%), blood vessels (8.14%), and fatty tissues (22.03%).

6.2.2 ICIAR 2018 - Breast Cancer Histology Dataset

We collected 10 samples from breast histology images. Initially, the challenge provided a segmentation mask for four classes of tissue: normal, benign, in situ carcinoma, and invasive carcinoma. However, our primary interest lies mainly in blood vessels, which are present in the previous dataset. Because of this, in collaboration with the Faculty of Medicine, Comenius University Bratislava, we requested our annotation of blood vessels. Figure 6.3 provides us with a sample of this dataset, and at first sight, we can see that the two datasets are pretty different, allowing us to test our hypothesis on multiple datasets.

6.2.3 Dataset pre-processing

The pre-processing approach for both datasets was executed similarly, with only minor differences. The primary goal was to extract image patches of sufficient size to serve as input for our models — 256×256 pixels for the IKEM dataset and 512×512 pixels for the ICIAR dataset, allowing for more contextual information

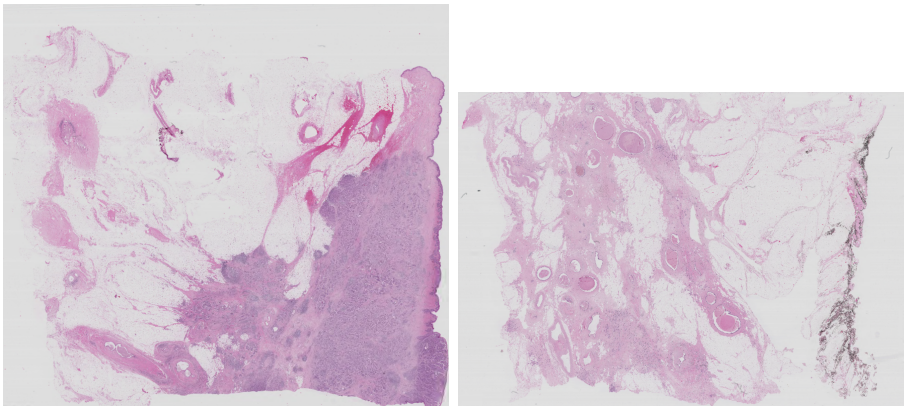

Figure 6.3: Sample whole slide images collected from ICIAR 2018

around the selected regions. We systematically checked the images using a sliding window of size 256×256 with a stride of 128×128. A patch was saved if specific criteria were met — namely, the patch contained at least 50% tissue and at least 10% of a target label. The overlap between the patches was selected due to the small amount of data. With this approach, we got around 30000 patches for the IKEM dataset and 28000 for ICIAR.

6.3 Architecture of the Synhtesis Model

This section will present our synthesis models' architectures in depth. We begin by introducing the noise estimator and detailing its components. Following this, we also introduce the Autoencoder, which encodes the input image into latent space, enabling the use of the LDM.

6.3.1 Semantic Synthesis Model

We decided to use the state-of-the-art U-Net architecture for deep-learning medical imaging solutions. In our work, the network is applied as a noise estimator for the synthesis process and is shown in Figure 6.4. The network usually consists

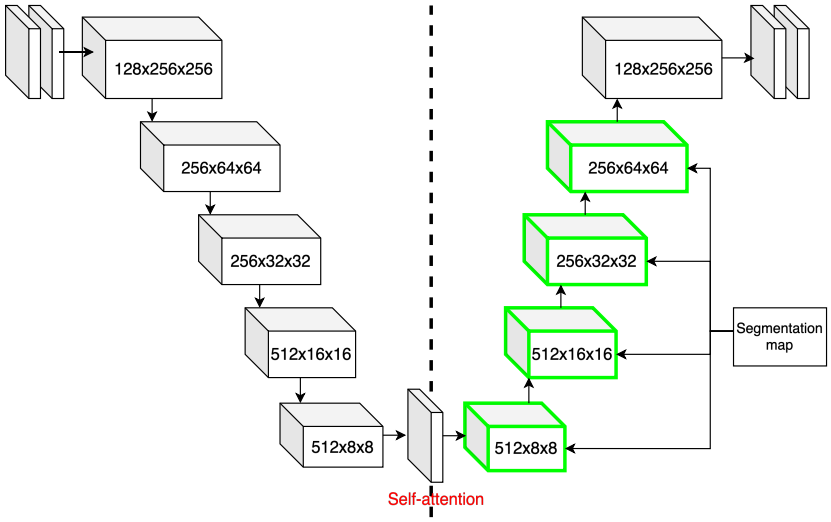

Figure 6.4: The architecture of our model, following conventional U-Net architecture and with self-attention in the bottleneck.

of two parts: an encoder and a decoder. The blocks are built from ResNet-based architecture, which extracts hierarchical features through the use of residual connections [13]. We applied a method proposed by [35] to achieve precise control over the generated synthetic images. The decoder (marked with green in Figure 6.5) reconstructs the synthetic image and incorporates additional context from the encoder and the segmentation map. The semantic information is embedded in the network through Spatially Adaptive Normalization (SPADE) layers [27]. This approach helps the model better align with the provided feature map and semantic information. Finally, the network employs SiLU [28] activation functions to introduce smooth, non-linear transformations, which enhance gradient flow in deeper models.

6.3.2 Autoencoder

For the experiment in latent space, we needed an autoencoder over our synthesis model; for this purpose, we selected an autoencoder architecture as a vector quan-

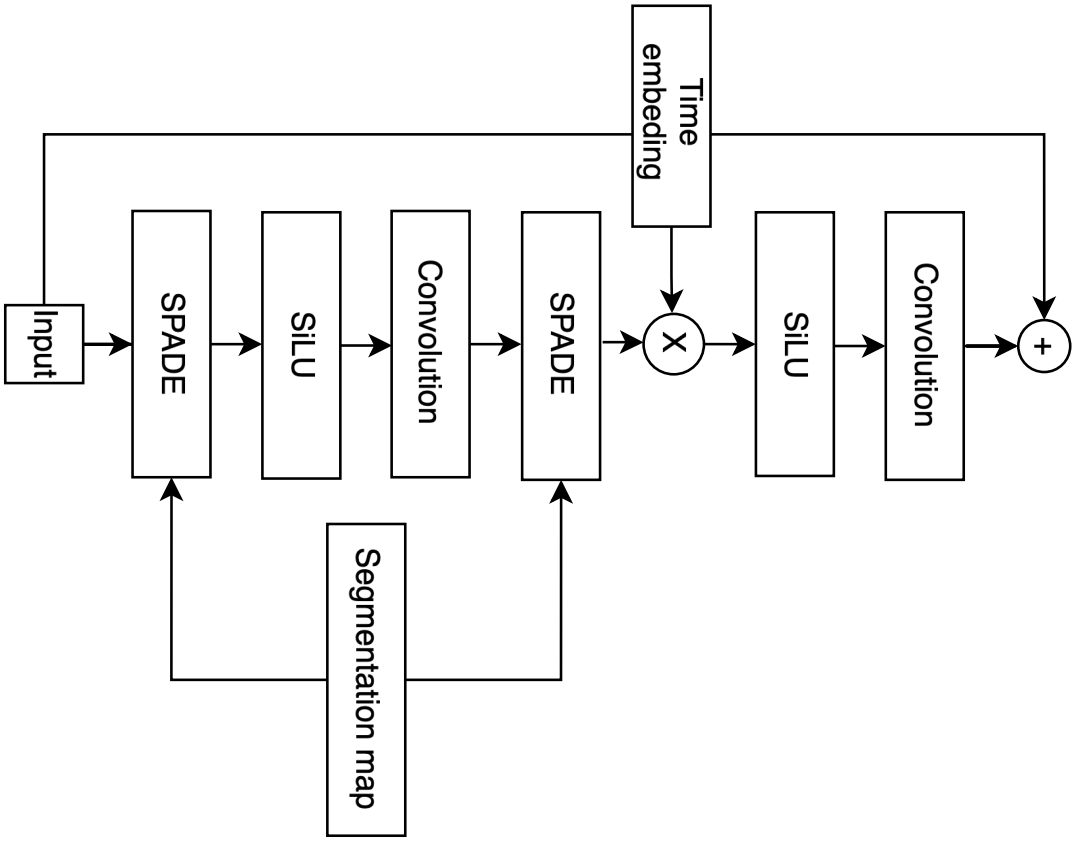

Figure 6.5: The architecture of the decoder block, with SPADE layer to embed semantic information.

tized variational model [32]. The encoder and decoder blocks are built in similar way, both of the consist from residual block and attention mechanisms for effective feature extraction. Between them is a vector quantization step, which replaces continuous latent vectors with discrete representations drawn from a codebook.

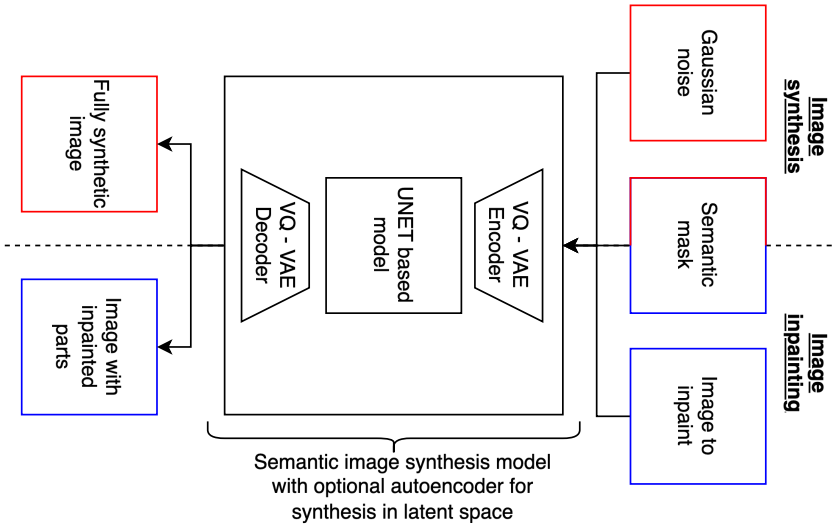

Figure 6.6: Sampling process of our synthesis model, the red side presents the fully synthetic image synthesis and the blue side stand for image inpainting. The VQ-VAE part is optional and used only for latent space sampling.

6.3.3 Inference Process of Image Synthesis

The inference process of our solution is presented in Figure 6.6. The process has two main branches, and the same synthesis model can be used for both of them. The red part of the diagram describes the image synthesis process. In that case, the model gets a Gaussian noise and semantic mask as its input. Then, it iteratively refines the noise based on the semantic mask to generate a new synthetic image. The aim of the blue side is image inpainting; in that case, the input is a real image with covered parts, and these covered parts are signed to the label to modify these parts of the input. In the end, we do not get a fully synthetic image, only a partially modified one. For the synthesis in latent space, we added an optional Vector Quantized Variational Autoencoder [32] part to the synthesis model. The synthesis models are trained as a traditional DDPM [16], but thanks to the nature of the inpainting process [21], we can use this model during the inference in both ways without any other modification.

6.4 Semantic Image Segmentation

We employed a ResNet-based U-Net architecture [13] for semantic image segmentation to delineate the biological structures within our heart tissue images. The model is structured as an encoder-decoder, wherein the ResNet blocks extract hierarchical features from the input images. To enhance feature selection in the decoder, we incorporated an Attention Gate (AG) mechanism as introduced in [25] into our U-Net-based segmentation model. This attention mechanism is designed to refine the spatial feature maps by selectively suppressing irrelevant activations while emphasizing the most informative regions. Given a gating signal $g \in \mathbb{R}^{C \times H \times W}$ from the decoder and skip connection features $x \in \mathbb{R}^{C \times H \times W}$ from the encoder, the AG computes an attention map $\alpha \in [0,1]^{1 \times H \times W}$ as follows:

$$g' = W_g g + b_g \tag{6.1}$$

$$x' = W_x x + b_x \tag{6.2}$$

$$\psi = \sigma\left(W_\psi \cdot \mathrm{ReLU}(g' + x') + b_\psi\right) \tag{6.3}$$

$$\alpha = \mathrm{Upsample}(\psi) \tag{6.4}$$

Here, W_g, W_x, W_ψ are learnable weights, σ denotes the sigmoid activation, and 'Upsample' is used to match the spatial resolution. The final output is the element-wise multiplication of the attention map with the original skip connection:

$$\tilde{x} = \alpha \odot x \tag{6.5}$$

The model is trained using a combination of Binary Cross-Entropy and Dice loss. During training, it learns to predict pixel-wise class labels for each image, and its

performance is evaluated using the Dice score, Precision, and Recall.

6.5 Experimental Setup

Hardware

The synthesis model was trained on an NVIDIA RTX 6000 Ada Generation graphical card with 48 GB memory, and the segmentation model was trained using an NVIDIA GeForce GTX 4090 graphics card with 24GB of memory.

Hyperparameters

We tested with a multiple-model configuration, changing the levels in the U-Net architecture and testing self-attention applications at different levels. The final architecture is visible in Figure 6.4. We down-sample the spatial dimension four times, and a self-attention layer is used only in the bottleneck of the model. The detailed description of the model blocks was provided in Section 6.3.1. With this configuration, the model has around 90 million learnable parameters. The hyperparameters for the diffusion process were set to the conventional one, and the denoising process was set to 1000. The linear beta schedule ran from 0.0001 to 0.02. The synthesis model was trained for 500 epochs, the batch size was 16, the learning rate was set to 0.0002, and we used Adam for the optimization. Input image values were scaled to the range [-1, 1] before input into the model. The loss function used to train the diffusion model is the Mean Squared Error (MSE), as defined in Equation 6.6. The MSE loss is then computed between the ground truth noise and the predicted noise: where ϵ_i is the true noise added to the i^{th} sample, and $\hat{\epsilon}_i$ is the predicted noise.

Table 6.1: Blood Vessel class representation percentage across dataset variations

Dataset Variation	**Blood Vessel (%)**
Real Dataset	8.14
Real + 2.5K Synthetic pcs.	11.22
Real + 5K Synthetic pcs.	14.06
Real + 7.5K Synthetic pcs.	18.97
Real + 10K Synthetic pcs.	19.16

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\epsilon_i - \hat{\epsilon}_i\right)^2, \tag{6.6}$$

We set the hyperparameters to batch size of 16 and 100 epochs for the segmentation experiments. The learning rate was set to 0.0002, and an exponential scheduler was to gradually decrease the learning rate during training, thereby enhancing convergence stability [9]. Finally, we trained a fifteen-segmentation model, three for every dataset variation described in Table 6.1. In the synthetic dataset, only labels of blood vessels were present; with the extension of this class, the other classes decreased slightly. The percentage shows how many percent of every mask was marked as blood vessels.

6.6 Evaluation of the Results

The evaluation is split into two parts, based on the used datasets. Both parts have two main groups of experiments: one is connected with image synthesis and the other with image segmentation of the corresponding dataset.

6.6.1 Results for the IKEM Heart Tissue Dataset

The quantitative metrics for the data synthesis experiment are in Table 6.2. Our goal was to compare the quality of synthetic data sampled in pixel space against

Dataset group	**KID (↓)**	**FID (↓)**	**LPIPS (↑)**
Fully synthetic - sampled in pixel space	0.070	77.395	0.58
Fully synthetic - sampled in latent space	0.054	59.905	0.59
Inpainted - sampled in latent space	0.032	47.760	0.58

Table 6.2: Evaluation metrics in pixel and latent space for inpainted, and synthetic data. ↓ indicates that lower values are better, and ↑ indicates that higher values are better.

data from the latent space. For evaluation, we used the three metrics discussed in 4.5: KID, FID, and LPIPS. In every case, the number of samples was set to 10,000. FID and KID are significantly lower for the datasets sampled in the latent space than those sampled from the pixel space. Among datasets from the latent space, the inpainted has the lower values, which aligns with expectations, as the mixed dataset contains original images, naturally providing better alignment with the ground truth. Based on these findings, we selected the inpainted dataset from the latent space as our synthetic dataset for semantic segmentation experiments

6.6.1.1 Generation of Fully Synthetic Images

In this section, we will evaluate how well the generated images align with real ones from the viewpoint of non-domain specialists. The Figure 6.7 is divided into quarters, each presenting an image pair. In every pair, the left is the original image, and the right is generated based on the real images' semantic mask. We notice that the main structure of the tissue (pink regions) is well reproduced, and the edge between the tissue and background (white areas) is well aligned with the original images. The purple "dots" in the synthetic images represent cells and are visually similar to the original images. However, the weakness of our model is specifically visible in pairs on the left side. In these cases, the original images contain holes in their inner regions, which are blood vessels, and we can see that

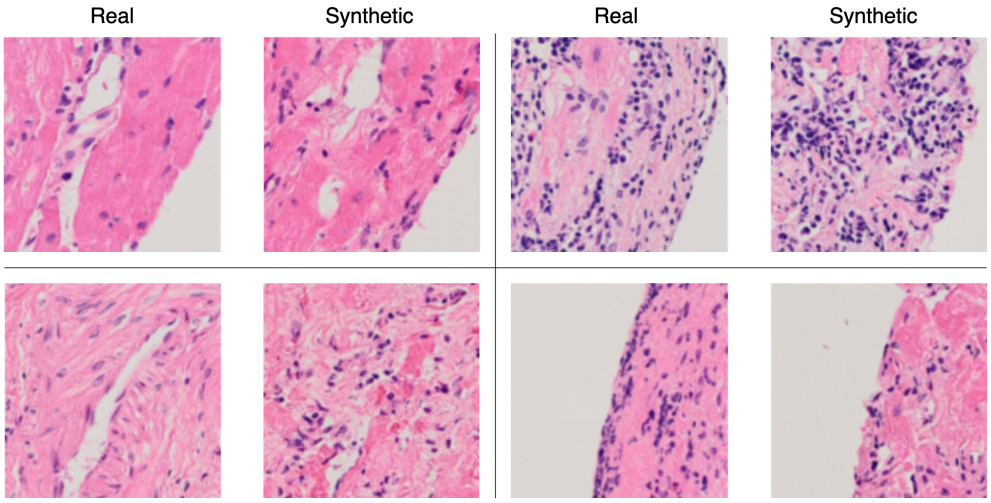

Figure 6.7: Synthetic sample pairs generated by our model, the synthetic image was generated based on the semantic mask of the real image.

the model could not recreate it in high quality.

6.6.1.2 Histology image modification with inpainting

During the evaluation of image modification through inpainting, we focus primarily on blood vessels, as highlighted in our goals (Section 1.1). Figure 6.8 is organized as follows:

- **First row:** Examples of real images containing regions of blood vessels marked by green lines.
- **Second row:** Examples of images without any semantic labels, referred to as "clean tissue." These images serve as the basis for modification.
- **Third row:** Modified images, with green lines indicating the inpainted regions.

Firstly, we want to highlight the positive aspect that the modified regions are well aligned with the original image, with no visible transition between the modified and

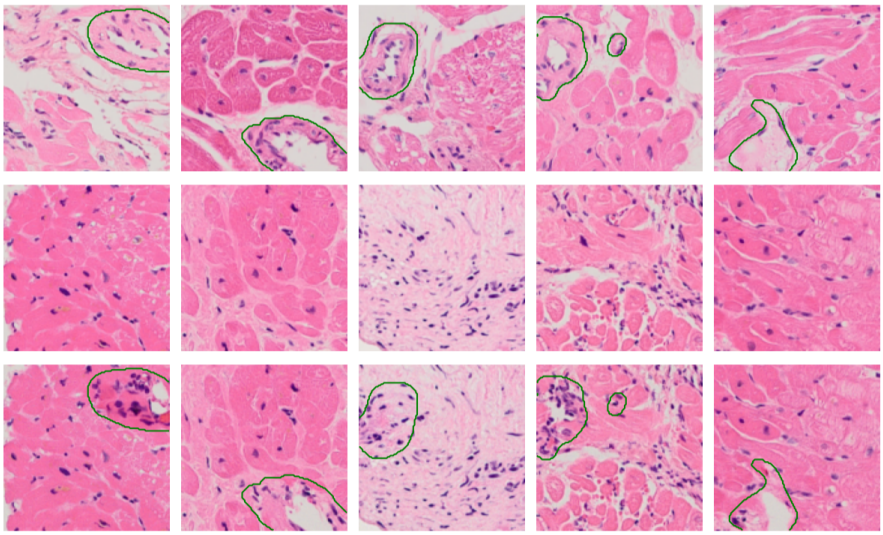

Figure 6.8: Inpainted blood vessels

unmodified areas. Every column shows visible traces of the network attempting to create blood vessels in the form of holes in a given area, except for the third column, where almost no visible change is observed.

6.6.1.3 Segmentation performance

From the 51 available samples, we randomly selected 3 WSI images, which were not used to train any segmentation model. These samples were used to calculate the testing metrics for the segmentations. Every data point visible in Figure 6.9 is an average from three runs on the same augmented dataset, to reduce variations due to initializing the weights in the network. Dice score and Precision decreased in every case. However, the recall improved, which can indicate that the model is missing fewer segmentations. The dropping Precision suggests a higher rate of false positives as the model becomes more sensitive. Not missing the segmentation label is more important than the low percentage of false positive predictions in medical imaging. We visualized (Figure 6.10) the true positive, false positive, and false negative predictions per image sample to better understand the models'

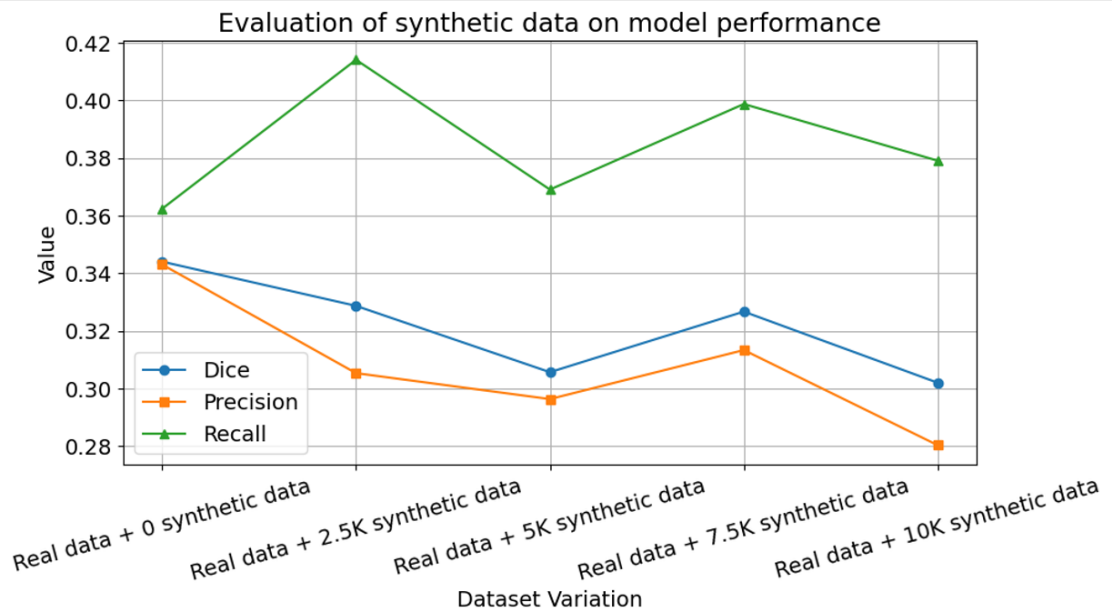

Figure 6.9: Evaluation metrics comparison across trained segmentation models. The chart displays Dice scores for blood vessels across dataset variations.

behavior. The top figure shows that in a few cases (blue regions), the model could not detect any part of the vessels. Green regions describe the true positives, and it is observable that the majority of the vessels were detected at least partially. In that case, there is a minority of false positive predictions (red regions). The tissue in the figure below has a leaky structure. These holes, which are marked as blood vessels, probably confuse the model. Again, it is visible that the model at least partially marked almost every blood vessel correctly. However, the model sensitivity in that case is very high, which can overwhelm the pathologist with too many false-positive cases and slow them down.

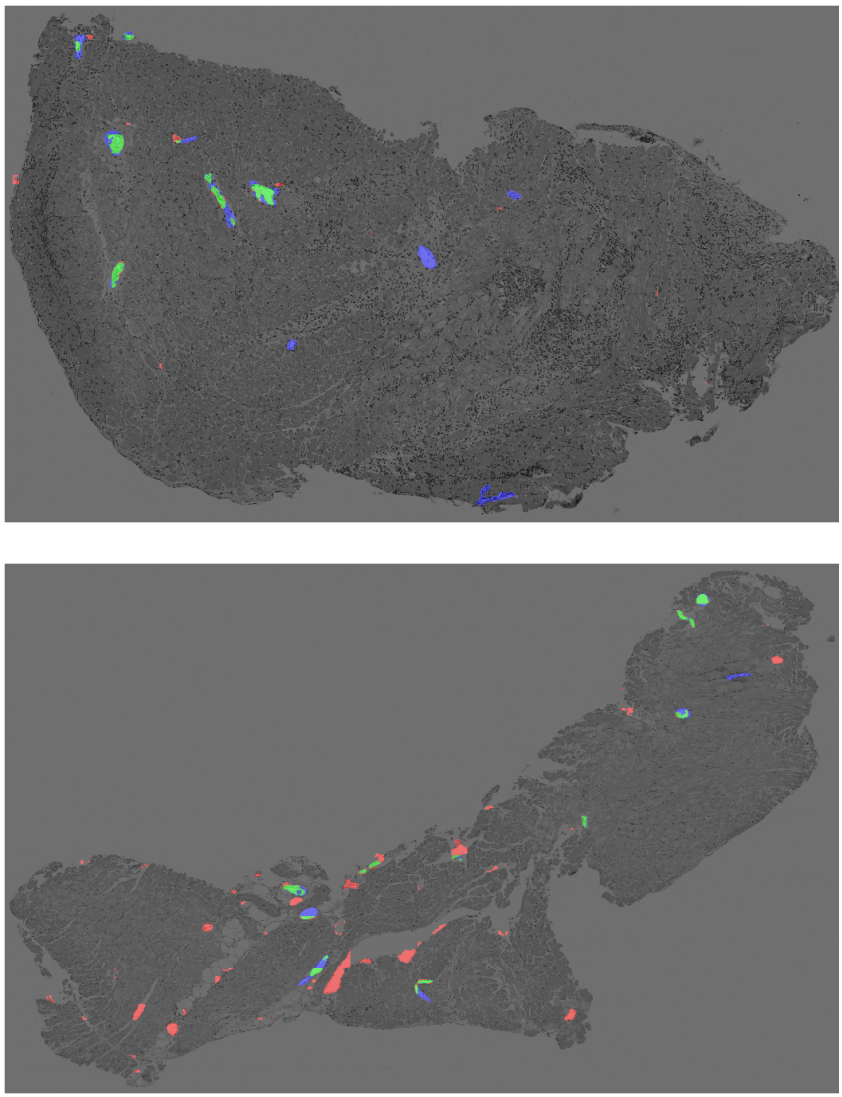

Figure 6.10: Examples of segmentation results. The green color indicates true positive predictions, while the red color indicates false positive predictions, blue is used for false negative predictions.

6.6.2 Results for the ICIAR 2018 Breast Cancer Histology Dataset

Since synthesis in latent space significantly overperformed the synthesis in pixel space (Table 6.2), we changed the first part of our experiments for the ICIAR data. In this case, we do not sample in pixel space; however, we want to examine the importance of a number of channels in latent space. The subsequent experiments with segmentation follow the same pattern as previously.

6.6.2.1 Comparison of latent space sizes' effect

We intended to train two VQ-VAE models to decode our image data into the latent space, where the synthesis will be performed. For both cases, the input shape into the autoencoder was $3\times512\times512$; for the first model, the shape of latent space was $4\times64\times64$, and for the second, $8\times64\times64$. The models were trained over 200 epochs with patches of ICIAR data. In Figure 6.11, we can visually compare the reconstructed image for both autoencoders; however, from a visual perspective, there is almost no difference between them. After preparing the autoencoders, we trained two synthesis models with the same setup to compare them. For comparison, we created two inpainted datasets with 10000 patches and used the same metrics KID, FID, and LPIPS as previously. The quantitative metrics in Table 6.3 show that the model with four channels in latent space slightly overperformed the second one, so we will continue with that dataset for the segmentation experiments.

Latent space shape	**KID**	**FID**	**LPIPS**
$4 \times 64 \times 64$	0.005	8.550	0.602
$8 \times 64 \times 64$	0.008	10.700	0.604

Table 6.3: Comparison of synthesis models trained on latent spaces of size $4 \times 64 \times 64$ and $8 \times 64 \times 64$.

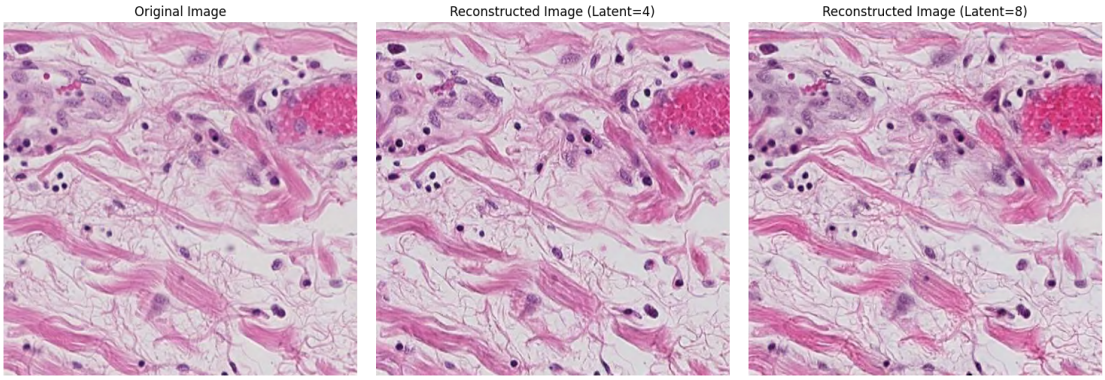

Figure 6.11: Comparison of reconstructed images from two VQ-VAE models with different latent space shapes.

6.6.2.2 Segmentation performance evaluation

Similar to the segmentation experiments for the IKEM dataset (Section 6.6.1.3), firstly, we train the segmentation model without synthetic data. However, in this case, because of time constraints, we were not able to experiment with various ratios of the synthetic data in the augmented dataset. We used the entire synthetic dataset from Section 6.6.2.1 for the augmented dataset. Figure 6.12 presents the results from the segmentation; again, we trained three separate models for both cases, and the values are the averages from them. We separated one WSI image from the available ten samples for the testing round. The results are almost identical to those of the IKEM dataset: recall increased, and precision with dice score dropped. However, the models' overall performance was better.

For qualitative analysis of the segmentation results, we used Figure 6.13, which shows the true positive (green), false positive (red), and false negative (blue) predictions of the model. The first column contains the results from one selected run without and the second with synthetic data. The model with the augmented dataset started to notice blood vessels better. We can see this from the increase in green areas in the upper image. Probably, the synthetic data introduced un-

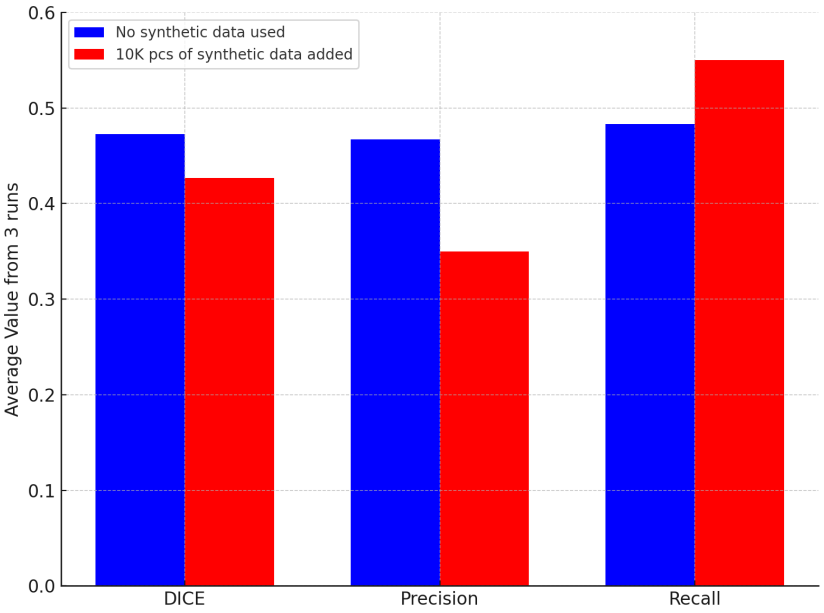

Figure 6.12: Blood vessels segmentation metrics comparison for iciar dataset. The blue bars represent the metrics from training on the original dataset, while the red bars are for training on the augmented dataset.

wanted noise into our dataset, and because of this, the segmentation model started to predict much larger false positives. This combination of results gives us an explanation for our quantitative metrics. Recall increased due to the larger area of true positive predictions, but the model did not start to recognize new previously undetected vessels, it just produced a large number of false positives, so the precision decreased.

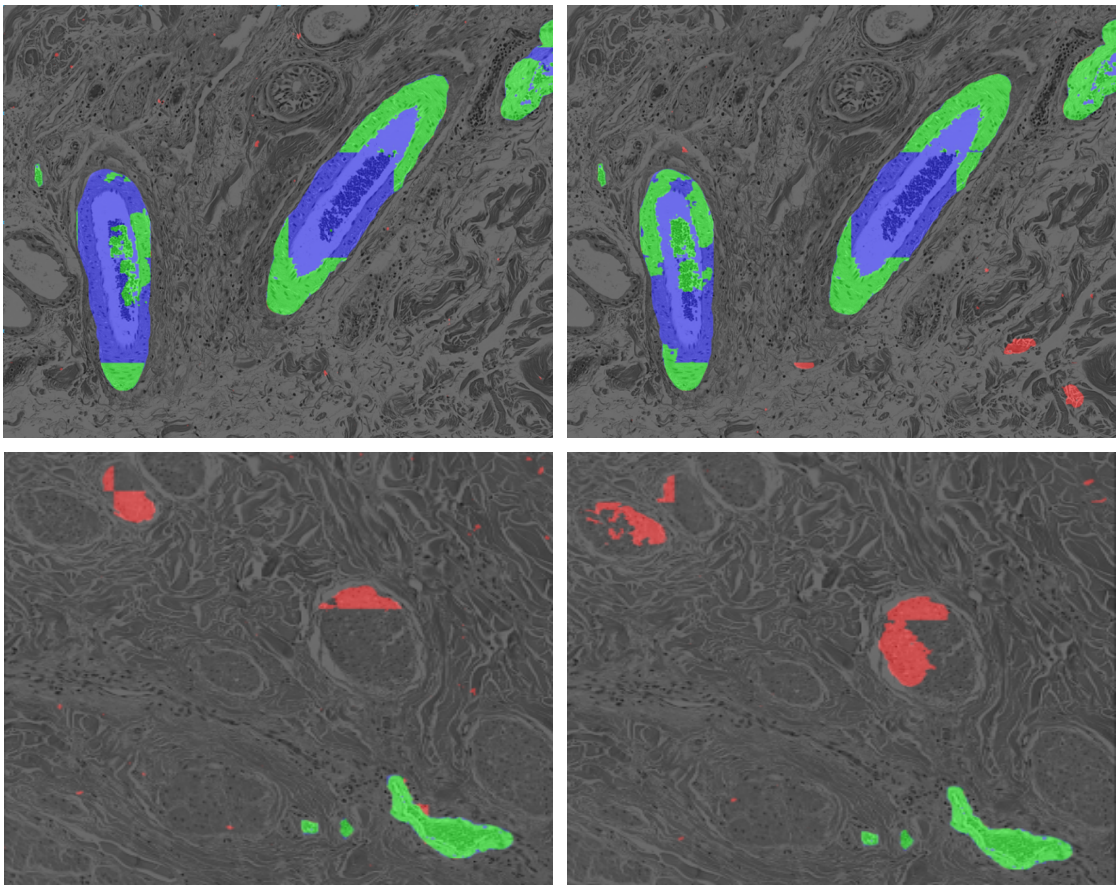

Figure 6.13: Qualitative comparison of semgnetation results. The first column shows the result for segmentation trained without synthetic data, the second with synthetic data.

Chapter 7

Conclusion

In this work, we presented our solution for addressing challenges in synthetic image generation and inpainting histological data, focusing on semantic blood vessel segmentation. Our pipeline combines image generation networks DDPM, LDM, and ResNet U-Nets with attention gates for segmentation to explore the influence of synthetic data on the segmentation models' metrics. The segmentation model was trained with augmented data, where various real and synthetic data ratios were present.

In our first experiment, we compared the quality of synthetic images sampled from pixel space against those sampled from latent space. For the quantitative evaluation, we used three metrics: KID, FID, and LPIPS. They all indicated that samples from the latent space better align with the ground truth. In the latent space, we sampled fully synthetic and inpainted data, which are partially real and partially modified. Among datasets from latent space, the in-painted has better values, which aligns with expectations, as the mixed dataset contains original images, naturally providing better alignment with ground truth.

For the synthetic dataset in IKEM segmentation experiments, we selected the in-

painted dataset sampled from the latent space with 10,000 patches. We decided to systematically augment the real dataset with synthetic data by 2,5000 patches in every round. Finally, the class representation of blood vessels increased from 8.14% to 19.15%. We trained three models for every combination of real and synthetic data. Overall, for this experiment, we used fifteen separate segmentation models. We evaluated the segmentation qualitatively, which suggests that the segmentation model performed well when the tissue did not have a leaky structure; in contrast, the model was confused with the holes and marked them as blood vessels. For quantitative evaluation, we had three WSI samples and calculated the dice score, precision, and recall over them. The recall increased, and the dice score with precision decreased along with the growth of synthetic data.

Since our experiments with image synthesis showed that data from latent space produce better results, we decided to examine the effect of latent space size on synthetic data quality. For this purpose, we trained two LDM models; in the first case, the number of channels in the latent was set to four, and for the other, it was set to eight. We did not find any significant difference during the visual inspection of the reconstructed images. The quantitative metrics selected the smaller latent space as better.

Because of the time constraint, for the segmentation experiments with ICIAR data, we trained the models with zero synthetic data and an augmented dataset using the whole synthetic set. The quantitative results were similar to the segmentation result with IKEM datasets; the dice score and precision dropped after the augmentation, and the recall increased. After visually inspecting some segmentation areas, we discovered that the model with the augmented dataset started to predict the blood vessels more precisely with a higher segmented area against the base model. However, we do not recognize any case where the augmented model found previously not segmented blood vessels. The other side is that the model started

to predict more false positive predictions, which dropped the precision and dice score, probably because of the noise from the synthetic data.
To sum up, our synthetic data in both cases improved the model sensitivity to better segment the blood vessels; however, their quality was probably not good enough, as is indicated by the decreasing dice score and precision.

7.1 Possible improvements

We see three main parts to the possible future work. One possible solution is extending our working pipeline with a classification model at the end. This model would classify the segmentation areas and filter out the false positive cases. Another solution would be to improve the sampling of synthetic data. Currently, we use a pure Gaussian noise as a starting point for DDPM; however, a slightly noisy image containing real blood vessels could be a better starting point to add some guidance to the model. The last possible improvement could be combining the two datasets for the synthesis process to improve the generalization of the blood vessel class.

Chapter 8

Resumé

Úvod

V tejto práci analyzujeme nedostatok dobre anotovaných digitálnych histologických snímok srdcového tkaniva, najmä cievnych štruktúr, ktorý obmedzuje výkonnosť segmentačných modelov založených na hlbokom učení. Navrhujeme a experimentálne porovnávame dva prístupy k augmentácii dát – úplnú syntézu obrázkov aj ich čiastočné editácie, a to v pixlovej aj latentnej reprezentácii. Naším cieľom je overiť, do akej miery doplnenie reálnej dátovej množiny o tieto syntetické snímky zlepší metriky segmentačných modelov pri zachytávaní podreprezentovaných biologických štruktúr.

Analýza

V práci sa zaoberáme medicínskym zobrazovaním s dôrazom na digitálnu histopatológiu. Medicínske obrázky, tvorené najmä neinvazívnymi technikami (CT, RTG a digitálne patológie), tvoria väčšinu dát v zdravotníctve. Podrobnejšie sa venu-

jeme digitalizácii histologických preparátov (WSI – Whole Slide Images), ktoré sú detailné, priestorovo rozsiahle a kľúčové pre presnú diagnostiku, no zároveň kladú vysoké nároky na dátovú infraštruktúru, anotácie odborníkov a spracovanie obrazu.

Predstavujeme počítačové videnie a jeho moderné metódy založené na hlbokých neurónových sieťach. Podrobnejšie popisujeme základy hlbokého učenia, vrátane štruktúry neurónových sietí, aktivačných funkcií, konvolučných vrstiev a procesu trénovania s dôrazom na sieťovú architektúru U-Net, ktorá je štandardom pri segmentácii medicínskych obrazov. Popisujeme tiež metriky pre kvantitatívne hodnotenie výsledkov segmentácie (Dice koeficient, presnosť, citlivosť).

V ďalšej časti sa zameriavame na syntézu obrazov pomocou generatívnych modelov – GAN, VAE a predovšetkým difúzne modely (DDPM a latentné difúzne modely – LDM), ktoré umožňujú generovať realistické syntetické snímky na doplnenie existujúcich dátových súborov. Detailne analyzujeme princípy a aplikácie týchto modelov, ako sú semantická syntéza obrazu a techniky „image inpainting“ (napr. RePaint). Pre hodnotenie kvality syntetických obrázkov uvádzame metriky FID a KID.

Cieľom tejto práce je na základe uvedených metód navrhnúť vlastné riešenie pre generovanie plne aj čiastočne syntetických histologických snímok, ktoré by zlepšili výkon segmentačných modelov, najmä v prípadoch s obmedzeným množstvom anotovaných údajov.

Súvisiace práce

V tejto kapitole sme analyzovali existujúce štúdie zamerané na syntézu histologických obrázkov pomocou metód založených na difúznych pravdepodobnostných

modeloch (DDPM).

Práca Zhenqiho He a kol. využíva DDPM na odstraňovanie artefaktov z histologických snímok (tzv. Artifusion). Na rozdiel od generovania celých obrázkov používajú modifikovaný U-Net s architektúrou Swin-Transformer, čo umožňuje presnú rekonštrukciu iba poškodenej časti obrazu bez nežiadúcej zmeny celého štýlu preparátu.

Marco Aversa a kol. predstavili metódu DiffInfinite pre generovanie veľkorozmerných histologických snímok (až 8000 × 8000 pixelov) pomocou viacstupňového procesu spätnej difúzie. Ich riešenie využíva latentné kódovanie (VAE) a efektívnejšie varianty DDPM (DDIM), čím zrýchľujú proces syntézy pri zachovaní vysokej kvality generovaných snímok.

Mathias Öttl a kol. sa zamerali na augmentáciu datasetov syntetickými snímkami pre lepšiu segmentáciu HER2 nádorov prsníka. V porovnaní modelov GAN, DDPM a inpaintingu dosiahli DDPM najlepšie výsledky (Dice skóre 0,854), pričom generovali rozmanitejšie a realistickejšie obrazy oproti GAN, ktorý produkoval repetitívne vzory.

Xinyi Yu a kol. navrhli dvojfázový generátor syntetických jadier buniek. Najprv DDPM generuje inštančné masky jadier, ktoré sú následne premenené na realistické syntetické snímky pomocou SPADE architektúry. Už malé množstvo syntetických obrázkov (10 %) významne zvýšilo presnosť segmentačných modelov.

Marco Aversa a kol. tiež porovnali latentné difúzne modely (LDM) a GANy naprieč rôznymi medicínskymi modalitami. Potvrdili, že LDM prekonávajú GANy v kvalite, diverzite a realistickosti syntetických medicínskych obrazov.

Podobne Alimanov a kol. využili DDPM na generovanie syntetických obrazov

sietnice na segmentáciu ciev. Ich riešenie pozostáva zo série modelov zahŕňajúcich ViT bloky, ktoré zabezpečujú lepšie učenie priestorových vlastností.

Tieto práce ukazujú, že DDPM a ich varianty majú vysoký potenciál pre medicínske aplikácie, najmä v oblasti syntézy obrázkov a augmentácie datasetov, pričom významne zvyšujú diverzitu generovaných dát a zlepšujú robustnosť segmentačných modelov.

Naše riešenie

V tejto kapitole predstavujeme naše riešenie zamerané na augmentáciu datasetov pomocou syntetických histologických obrázkov na zlepšenie segmentácie biologických štruktúr. Naším cieľom je overiť, či doplnenie datasetu syntetickými obrázkami skutočne zlepší výkonnosť segmentačných modelov.

Navrhnutý prístup pozostáva z dvoch hlavných krokov. V prvom kroku sme trénovali model založený na architektúre U-Net s ResNet encoderom a dekóderom využívajúcim priestorovo adaptívne SPADE vrstvy, umožňujúce syntézu a inpainting histologických obrázkov na základe sémantických masiek. Model bol implementovaný v pixelovom aj latentnom priestore, pričom latentný priestor využíval VQ-VAE autoenkóder.

V druhom kroku sme použili synteticky generované obrázky na augmentáciu datasetu a trénovali segmentačný model s ResNet U-Net architektúrou, obohatenou o Attention Gate mechanizmus. Ten umožnil lepšie zvýrazniť relevantné oblasti obrázka počas segmentácie. Model bol optimalizovaný kombináciou Binary Cross-Entropy a Dice loss funkcie.

Experimenty prebehli na dvoch datasetoch – súbore snímok srdcového tkaniva získanom z IKEM a snímok prsného tkaniva z verejnej výzvy ICIAR 2018. Dáta

boli predspracované na menšie obrazy s rozmermi 256×256 a 512×512 pixelov. Do datasetov boli pridané syntetické snímky triedy krvných ciev v rôznych pomeroch.

Naše výsledky naznačujú, že využitie syntetických obrázkov na augmentáciu datasetu môže významne zvýšiť zastúpenie nedostatočne reprezentovaných tried (napríklad krvných ciev), a tým zlepšiť presnosť segmentačných modelov.

Vyhodnotenie

V tejto kapitole sme zhodnotili vplyv augmentácie datasetov syntetickými obrázkami na výkonnosť segmentačných modelov. Experimenty sme realizovali na dvoch datasetoch: IKEM (srdcové tkanivo) a ICIAR 2018 (prsné tkanivo).

Pri IKEM datasete sme najprv kvantitatívne porovnali syntézu v pixlovom a latentnom priestore pomocou metrík KID, FID a LPIPS. Lepšie výsledky boli dosiahnuté v latentnom priestore, pričom najlepšie skóre vykázal dataset upravený metódou inpaintingu. Pri kvalitatívnom hodnotení boli hlavné štruktúry dobre reprodukované, ale model mal problémy presne generovať detaily ako krvné cievy. Pri inpaintingu boli upravené oblasti bez viditeľného prechodu, ale nie vždy boli reprodukované všetky detaily.

Segmentácia na augmentovanom datasete vykazovala zlepšenie metriky Recall, čo znamená, že model úspešnejšie detegoval skutočné cievy. Naopak, pokles Precision a Dice skóre naznačuje, že model zároveň produkoval viac falošne pozitívnych predikcií, čo je z medicínskeho hľadiska menej kritické než vynechanie patologických štruktúr.

Pri ICIAR datasete sme porovnali dva modely s rozdielnym počtom kanálov v latentnom priestore (4 vs. 8). Model so štyrmi kanálmi dosiahol lepšie výsledky,

preto sme ho využili na ďalšie experimenty. Podobne ako v prípade IKEM datasetu, augmentácia zvýšila Recall, ale znížila Precision a Dice. Z kvalitatívneho hodnotenia bolo jasné, že syntetické dáta pomohli modelu zvýrazniť existujúce štruktúry ciev, no zároveň výrazne zvýšili počet falošne pozitívnych detekcií, čo znížilo celkovú presnosť segmentácie.

Získané výsledky naznačujú, že využitie syntetických obrázkov v augmentácii datasetov môže pomôcť zvýšiť citlivosť segmentačných modelov, no zároveň vyžaduje ďalšie vylepšenia na zníženie množstva falošných detekcií.

Záver

V našej práci sme navrhli riešenie problémov syntetického generovania a inpaintingu histologických obrázkov so zameraním na sémantickú segmentáciu krvných ciev. Skúmali sme vplyv augmentácie datasetov syntetickými dátami na výkonnosť segmentačných modelov. Pri syntéze dát sme porovnali generovanie v pixlovom a latentnom priestore, pričom latentný priestor preukázal lepšiu kvalitu syntézy podľa metrík KID, FID a LPIPS. Spomedzi latentných variantov dosahovali lepšie výsledky obrazy upravené inpaintingom oproti plne syntetickým.

Pri segmentačných experimentoch na datasete IKEM sme augmentovali pôvodné dáta rôznymi pomermi syntetických obrázkov. Zistili sme, že pridávanie syntetických dát zvýšilo citlivosť (Recall) modelu pri detekcii krvných ciev, avšak zároveň znížilo Precision a Dice skóre kvôli nárastu falošne pozitívnych predikcií. Podobné výsledky sme dosiahli aj pri dátach ICIAR, kde augmentácia tiež viedla k zlepšeniu Recall a zároveň k poklesu presnosti.

Pri analýze vplyvu veľkosti latentného priestoru sme zistili, že menší latentný priestor (4 kanály) poskytoval lepšie kvantitatívne výsledky než väčší (8 kanálov).

Budúcu prácu vidíme v troch oblastiach. Prvou možnosťou je rozšírenie segmentačného procesu o klasifikačný model, ktorý by filtroval falošne pozitívne detekcie. Druhou možnosťou je zlepšenie procesu syntézy použitím šumu obsahujúceho reálne štruktúry ciev namiesto čisto gaussovského šumu. Treťou oblasťou je zlúčenie datasetov IKEM a ICIAR pri tréningu syntetických modelov, čo by mohlo zlepšiť generalizáciu modelu a výslednú kvalitu syntetických obrázkov.

References

[1] Charu C Aggarwal et al. *Neural networks and deep learning*. Vol. 10. 978. Springer, 2018.

[2] Alnur Alimanov and Md Baharul Islam. "Denoising diffusion probabilistic model for retinal image generation and segmentation". In: *2023 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2023, pp. 1–12.

[3] Yannick Assogba, Adam Pearce, and Madison Elliott. "Large scale qualitative evaluation of generative image model outputs". In: *arXiv preprint arXiv:2301.04518* (2023).

[4] Marco Aversa et al. "Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology". In: *Advances in Neural Information Processing Systems* 36 (2024).

[5] Sugata Banerji and Sushmita Mitra. "Deep learning in histopathology: A review". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.1 (2022), e1439.

[6] Samah Saeed Baraheem, Trung-Nghia Le, and Tam V Nguyen. "Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook". In: *Artificial Intelligence Review* 56.10 (2023), pp. 10813–10865.

[7] Mikołaj Bińkowski et al. “Demystifying mmd gans”. In: *arXiv preprint arXiv:1801.01401* (2018).

[8] Wouter Bulten et al. “Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study”. In: *The Lancet Oncology* 21.2 (2020), pp. 233–241.

[9] Florinel-Alin Croitoru et al. “Learning rate curriculum”. In: *International Journal of Computer Vision* 133.1 (2025), pp. 291–314.

[10] Anton Eklund. *Cascade Mask R-CNN and Keypoint Detection used in Floorplan Parsing*. 2020.

[11] Matej Halinkovic et al. “Intrinsically explainable deep learning architecture for semantic segmentation of histological structures in heart tissue”. In: *Computers in Biology and Medicine* 177 (2024), p. 108624.

[12] GM Harshvardhan et al. “A comprehensive survey and analysis of generative models in machine learning”. In: *Computer Science Review* 38 (2020), p. 100285.

[13] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[14] Zhenqi He et al. “Artifact Restoration in Histology Images with Diffusion Probabilistic Models”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 518–527.

[15] Martin Heusel et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30 (2017).

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[17] Shanshan Huang et al. “Controllable image synthesis methods, applications and challenges: a comprehensive survey”. In: *Artificial Intelligence Review* 57.12 (2024), p. 336.

[18] David H Hubel and Torsten N Wiesel. “Receptive fields of single neurones in the cat’s striate cortex”. In: *The Journal of physiology* 148.3 (1959), p. 574.

[19] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[20] Zewen Li et al. “A survey of convolutional neural networks: analysis, applications, and prospects”. In: *IEEE transactions on neural networks and learning systems* 33.12 (2021), pp. 6999–7019.

[21] Andreas Lugmayr et al. “Repaint: Inpainting using denoising diffusion probabilistic models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11461–11471.

[22] Shervin Minaee et al. “Image segmentation using deep learning: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3523–3542.

[23] Gustav Müller-Franzes et al. “A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis”. In: *Scientific Reports* 13.1 (2023), p. 12098.

[24] Niall O’Mahony et al. “Deep learning vs. traditional computer vision”. In: *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer. 2020, pp. 128–144.

[25] Ozan Oktay et al. “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018).

[26] Mathias Öttl et al. “Improved her2 tumor segmentation with subtype balancing using deep generative networks”. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–5.

[27] Taesung Park et al. “Semantic image synthesis with spatially-adaptive normalization”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2337–2346.

[28] Prajit Ramachandran, Barret Zoph, and Quoc Le. “Swish: a Self-Gated Activation Function”. In: (Oct. 2017).

[29] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[31] Hamid Reza Tizhoosh and Liron Pantanowitz. “Artificial intelligence and digital pathology: challenges and opportunities”. In: *Journal of pathology informatics* 9.1 (2018), p. 38.

[32] Aaron Van Den Oord, Oriol Vinyals, et al. “Neural discrete representation learning”. In: *Advances in neural information processing systems* 30 (2017).

[33] Kunfeng Wang et al. “Generative adversarial networks: introduction and outlook”. In: *IEEE/CAA Journal of Automatica Sinica* 4.4 (2017), pp. 588–598.

[34] Lei Wang et al. “A state-of-the-art review on image synthesis with generative adversarial networks”. In: *IEEE Access* 8 (2020), pp. 63514–63537.

[35] Weilun Wang et al. “Semantic image synthesis via diffusion models”. In: *arXiv preprint arXiv:2207.00050* (2022).

[36] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. "Tackling the generative learning trilemma with denoising diffusion gans". In: *arXiv preprint arXiv:2112.07804* (2021).

[37] Yanqing Yang et al. "Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network". In: *Sensors* 19.11 (2019), p. 2528.

[38] Xinyi Yu et al. "Diffusion-based data augmentation for nuclei image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 592–602.

[39] S Kevin Zhou et al. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises". In: *Proceedings of the IEEE* 109.5 (2021), pp. 820–838.