

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

FIIT-182905-110830

Bc. Patrik Kozlík

**Deep learning methods for histological
image processing**

Master thesis

Supervisor: prof. Ing. Vanda Benešová PhD.

May 2026

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

FIIT-182905-110830

Bc. Patrik Kozlík

**Deep learning methods for histological
image processing**

Master thesis

Study program: Intelligent Software Systems

Degree course: Computer Science

Place of elaboration: Institute of Computer Engineering and Applied Informatics

Supervisor: prof. Ing. Vanda Benešová PhD.

May 2026



MASTER THESIS TOPIC

Student: **Bc. Patrik Kozlík**
Student's ID: 110830
Study programme: Intelligent Software Systems
Study field: Computer Science
Thesis supervisor: prof. Ing. Vanda Benešová, PhD.
Head of department: Ing. Katarína Jelemenská, PhD.

Topic: **Deep learning methods for histological image processing**

Language of thesis: English

Specification of Assignment:

Spracovanie histologických obrazových medicínskych dát s využitím metód počítačového videnia a hlbokých neurónových sietí má stále významnejšie uplatnenie. Hlboké neurónové siete sa javia ako účinná metóda pri diagnostikovaní chorôb či segmentácii medicínskych snímok a začínajú nahrádzať tradičné prístupy počítačového videnia. Pri správnom klasifikovaní medicínskych snímok je preto potrebná extrakcia jednotlivých črt snímky, ktoré sú niekedy aj pre patológa veľmi ťažko rozpoznateľné. Napriek veľkému pokroku v oblasti digitálnej patológie a použitia metód hlbokého učenia, stále zostáva mnoho výskumných úloh otvorených pre sofistikovanejšie riešenia. Analyzujte súčasný stav v oblasti analýzy histologických snímok pomocou počítačového videnia a hlbokého učenia. Preskúmajte existujúce metódy spracovania histologických snímok, najmä techniky segmentácie, klasifikácie a detekcie patologických štruktúr. Identifikujte otvorené výskumné problémy spracovania histologických snímok v digitálnej patológii. Navrhните riešenie využívajúce najmodernejšie metódy počítačového videnia a hlbokého učenia pre riešenie vybraného otvoreného výskumného problému. V rámci vlastného riešenia navrhните vhodný model neurónovej siete, jeho architektúru ako aj stratégiu učenia modelu. Navrhnuté riešenie implementujte a aplikujte ho na reálnu dátovú sadu histologických snímok v programovacom jazyku Python s využitím rámcov a knižníc ako PyTorch a OpenCV. Vizualizujte výsledky a interpretujte výstupy modelu. Overte a vyhodnoťte úspešnosť navrhnutého riešenia a porovnajte výsledky s existujúcimi metódami na základe vhodných metrík ako napr. presnosť, F1-score, AUC-ROC, IoU.

Deadline for submission of Master thesis: 04. 05. 2026
Approval of assignment of Master thesis: 13. 04. 2026
Assignment of Master thesis approved by: prof. Ing. Vanda Benešová, PhD. – Study programme supervisor

Declaration of honor

I declare that the entirety of this thesis is my original work, based on consultations with my supervisor and the referenced literature.

In Bratislava, 4.5.2026

Bc. Patrik Kozlák

Acknowledgement

I sincerely thank my supervisor, prof. Ing. Vanda Benešová, PhD., for her help and support during the work on my Master's thesis. Her advice, willingness to always help, motivation to progressively improve, and the opportunities she provided through various moments and event participations that contributed to my growth in the research field have greatly motivated me.

I would also like to express my sincere gratitude to Ing. Maroš Kollár for his support, guidance, and valuable recommendations throughout this work.

Finally, I would like to thank my family and friends for their continuous support.

Annotation

Slovak University of Technology Bratislava

Faculty of Informatics and Information Technologies

Degree Course: Intelligent Software Systems

Author: Bc. Patrik Kozlák

Master Thesis: Deep learning methods for histological image processing

Supervisor: prof. Ing. Vanda Benešová PhD.

May 2026

Histopathological whole-slide images (WSIs) are hard to analyse and computationally demanding. The diagnosis relies on relations between local and global context, and clinical use requires interpretability. Research hypothesis asks whether hierarchical transformers with explicit relational attention can improve multi-scale WSI modeling and explainability. We propose the Relation-Aware Hierarchical Image Pyramid Transformer (RA-HIPT), fusing whole-slide and patch embeddings via cross-attention with graph regularization and a custom loss function aimed at relation discovery. We performed evaluation in self- and weak supervision on controlled synthetic WSIs and BRACS. For evaluation we use attention maps, region-region graphs, class-relation visualizations, and confusion analysis. A synthetic pipeline combines VQ-VAE, semantically conditioned diffusion, and inpainting-based stitching. Empirically, VQ-VAE reached FID 13.14; conditioned sampling on BCSS with DDIM-50 achieved FID 122.84 and mean SSIM 0.13. Alongside we performed seam ablations and a qualitative assessment of structures with 15 participants. With RA-HIPT, heatmaps and relation graphs highlighted tissue structure. These findings demonstrate synthetic dataset quality, model interpretability, and the model's ability to capture biologically relevant structures.

Anotácia

Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

Študijný program: Intelligent Software Systems

Autor: Bc. Patrik Kozlík

Diplomová práca: Deep learning methods for histological image processing

Vedúci diplomovej práce: prof. Ing. Vanda Benešová PhD.

May 2026

Histologické snímky (WSI) sú náročné na analýzu a spracovanie. Diagnóza spočíva vo vzťahoch medzi lokálnym a globálnym kontextom štruktúr pričom klinické nasadenie vyžaduje vysvetliteľnosť. Výskumná hypotéza skúma, či hierarchické transformery s explicitnou pozornosťou vzťahov môžu zlepšiť zachytenie vzťahov a vysvetliteľnosť. Navrhujeme Relation-Aware Hierarchical Image Pyramid Transformer (RA-HIPT), ktorý spája embeddingy celej snímky a výrezov cez cross-attention s grafovou regularizáciou a vlastnou stratovou funkciou zameranou na objavovanie vzťahov. Vyhodnotenie prebehlo v samo- a slabo supervizovanom režime na syntetických WSI a na BRACS dátach. Na vyhodnotenie slúžia mapy pozornosti, grafy vzťahov medzi regiónmi, vizualizácie triedových vzťahov a analýza výsledkov klasifikácie. Syntetický proces kombinuje VQ-VAE, sémanticky podmienenú difúziu a inpaintingové spájanie. Empiricky VQ-VAE dosiahol FID 13,14; podmienené generovanie na BCSS s DDIM-50 dosiahol FID 122,84 a priemerný SSIM 0,13. Zároveň sme vykonali ablačné štúdie a kvalitatívne hodnotenie štruktúr s 15 účastníkmi. S RA-HIPT heatmapami a grafmi vzťahov sme zachytili štruktúry tkaniva. Tieto zistenia dokazujú kvalitu syntetického datasetu, vysvetliteľnosť modelu a schopnosť modelu zachytiť biologicky relevantné štruktúry.

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Research Questions	4
2	Medical Imaging	9
2.1	Biomarkers	10
2.2	Digital Histopathology	12
2.3	Types of Tasks	13
2.4	Challenges	14
3	Deep Learning	19
3.1	Convolutional Neural Network	20
3.2	Vision Transformers	21
4	Feature Space	27
4.1	Encoders	28
4.1.1	Autoencoders	28
4.1.2	Variational Autoencoders	29
4.2	Latent Space	30
4.3	Feature Importance	31

4.3.1	Mahalanobis Distance	32
4.4	Dimensionality Reduction	33
4.5	Interpretable Latent Space	36
4.6	Disentanglement	37
4.6.1	Architectural Approache	38
4.6.2	Data Augmentation and Supervision	38
4.6.3	Loss Functions and Regularization	39
4.6.4	Latent Space Interpretability Metrics	42
4.6.5	Challanges of Disentanglement	44
5	Generative Neural Networks	47
5.1	Vector Quantized Variational Autoencoder	48
5.2	Denoising Diffusion Models	50
5.2.1	Denoising Diffusion Probabilistic Models (DDPM) [43]	52
5.2.2	Denoising Diffusion Implicit Models (DDIM) [96]	53
5.2.3	Pseudo Linear Multistep Methods (PLMS) [59]	54
5.2.4	Inpainting with Diffusion Models	54
5.2.5	Semantic Mask-Conditioned Generation	56
6	Explainability	59
6.1	Post-hoc Vizualization Methods	60
6.1.1	Grad-CAM	61
6.1.2	Layer-wise Relevance Propagation	62
6.2	Ante-hoc Methods	63
7	State of the art	67
7.1	BoQ: A Place is Worth a Bag of Learnable Queries [7]	68
7.2	Weakly Supervised Vector Quantization for Whole Slide Images Classification [94]	70

7.3	Enhancing Interpretability Through Loss-Defined Classification Objective in Structured Latent Spaces [36]	72
7.4	Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning (HIPT) [22]	75
7.5	ZoomLDM: Latent Diffusion Model for multi-scale image generation [118]	77
7.6	Summary	79
8	Our Approach	81
8.1	Approach Overview	82
9	AI-driven Biomarker Prediction	87
9.1	Whole Slide Embeddings Extraction	88
9.2	Vision Transformer with Dual-Branch Attention Mechanism	89
9.2.1	Tissue Branch	89
9.2.2	Relation Branch	90
9.3	Attention Module based Custom Optimized Loss Function	91
9.4	Loss Scheduler	96
9.5	Self Supervised Learning	97
9.6	Graph Structure based RA-HIPT Modification (Graph-RA-HIPT) .	98
9.6.1	Graph Structure Based Regularization Module	98
9.6.2	Graph Structure Focused Loss functions	99
10	Controlled Synthetic Validation Dataset	105
10.1	Overview and Pipeline	106
10.2	Latent Space (encoder-decoder)	108
10.3	Semantic-mask-conditioned Diffusion	108
10.3.1	Classifier-free Guidance	108
10.4	Inpainting (Repaint)	109

10.4.1	Grid Mask Inpainting	110
10.4.2	Higher Resolution Region Generation with In- painting	110
10.5	Color Normalization	111
10.5.1	Reinhard Normalization	111
10.5.2	Edge Normalization	112
10.5.3	Latent Noise Initialization Normalization	113
11	Experiments	115
11.1	Datasets	115
11.1.1	BReAst Carcinoma Subtyping Dataset (BRACS) [15]	116
11.1.2	Breast Cancer Semantic Segmentation Dataset (BCSS) [9]	117
11.1.3	Controlled Synthetic Validation Dataset	118
11.2	Evaluation Metrics	119
11.2.1	Synthetic Data Generation Metrics	120
11.2.2	AI-driven Biomarker Detection Methods for Solution Analysis	124
11.3	Synthetic Dataset Experiments	126
11.3.1	VQVAE Autoencoder Experiment	127
11.3.2	Synthetic Image Generation Experiments	129
11.3.3	Semantic Mask Conditioned Synthetic Image Generation Ex- periments	132
11.3.4	Inpainting Grid Mask Experiments	133
11.3.5	Guided Region Generation with Inpainting of 1024x1024 re- gions	135
11.3.6	Guided Region Generation of 4096x4096 regions via downsample- inpaint-upsample	136
11.3.7	Guided Whole Slide Image Generation via downsample-inpaint- upsample	137
11.3.8	Ablation Studies	137

11.3.9 Qualitative Analysis with Lay Users	141
11.4 AI-driven Biomarker Detection Preliminary Experiments	144
11.4.1 RA-HIPT Experiments (self-supervised, 4096x4096 regions, BCSS Dataset)	145
11.4.2 Graph RA-HIPT Experiments (self-supervised, 4096x4096 regions, BCSS Dataset)	147
11.4.3 Graph RA-HIPT Experiments (weak-supervised, 256x256 patches, Controlled Synthetic Dataset)	151
11.4.4 Graph RA-HIPT Experiments (weak-supervised, 256x256 patches, BRACS Dataset)	154
12 Viewer tool for qualitative analysis (HistoView)	163
12.1 Technology Stack	164
12.2 Functionalities of the Application	164
12.3 Use Cases	165
12.3.1 Qualitative Analysis of Synthetic Structures with Laik Users	165
12.3.2 Qualitative Analysis of the Predicted Heatmap Biomarkers .	166
13 Conclusion	169
13.1 Summary	169
13.2 Future Work	171
Declaration of Use of Artificial Intelligence	209
A Plan of the Work	A-1
B Conference Paper	B-1
C UX Testing: Structure Search in Synthetic Histology Images	C-1

D	Technical Documentation	D-1
D.1	File system structure	D-1
D.2	RA-HIPT: Installation and Setup	D-4
D.3	RA-HIPT: Demo Inference	D-5
D.4	Controlled Synthetic Dataset: Installation and Setup	D-7
D.5	Controlled Synthetic Dataset: Demo Inference	D-8
D.6	System design	D-10
D.6.1	RA-HIPT architecture	D-10
D.6.2	Controlled synthetic generation (WSI.Generator)	D-11
D.6.3	Shared engineering pattern	D-11
D.6.4	Verification and maintenance	D-12

List of Figures

1.1	Proposed image processing workflow of the histopathologist, with highlighted sections indicating the areas of focus in this study . . .	3
2.1	Example of histology image	10
2.2	Biomarker detection with a use of AI [80]	12
2.3	Diagram of the histological tissue processing [68]	13
3.1	Example of convolution neural network [55]	21
3.2	Example Vision Transformer architecture [116]	22
4.1	Example of autoencoder architecture [95]	29
4.2	Example variational autoencoder architecture [95]	30
4.3	Process of the latent space creation [74]	31
4.4	Visualization of disentangled latent features	32
4.5	Difference between Euclidean and Mahalanobis distance [18]	33
4.6	Example of PCA dimensionality reduction[4]	35
4.7	Data visualizations after dimensionality reduction with UMAP method	36
4.8	Changing of feature vectors of latent space trained on faces	37
5.1	Generative neural networks Trilemma [37]	48
5.2	VQ-VAE architecture diagram[107]	49

5.3	Forward and reverse process of DDM[117]	50
5.4	Image examples of using repaint with multiple different sampling steps[64]	55
5.5	Example of conditioned histology image generation[14]	56
6.1	Diagram of explainable AI methods	60
6.2	Visualization of Grad-CAM map [92]	61
6.3	Visualization of LRP map [71]	63
7.1	Architecture of Bag of Queries	68
7.2	Results of Bag of Queries architecture	69
7.3	VQ-MIL architecture	71
7.4	Results of experiments with VQ-MIL compared to state-of-the-art results	72
7.5	Flow of the latent boost approach	73
7.6	Results of the latent boost architecture	74
7.7	HIPT architecture	75
7.8	Heatmap results of HIPT architecture	76
7.9	HIPT architecture results and comparisons	77
7.10	ZoomLDM architecture diagram	77
7.11	Synthetic samples generated with ZoomLDM	78
7.12	Results of the ZoomLDM architecture on TCGA and BACH dataset	79
8.1	Current approach of biomarker prediction	83
8.2	Proposed approach for biomarker prediction, where blue part represent proposed solution for AI model focused on global-local relations and green part represent proposed controlled synthetic validation dataset used for hypothesis-driven validation	83

9.1	AI-driven biomarker detection and visualization pipeline, using RA-HIPT architecture	88
9.2	RA-HIPT transformer block with relation (cross-attention) and tissue (self-attention) branches	90
9.3	Graph RA-HIPT architecture	99
10.1	Pipeline of controlled synthetic validation dataset generation	106
10.2	Pipeline of controlled synthetic validation dataset generation. (1) Latent autoencoder (2) Semantic-mask-conditioned diffusion (3) Patch stitching (Inpainting - RePaint)	107
10.3	Straight mask example	110
10.4	Wavy mask example	110
10.5	Combination of Reinhard and edge normalization without inpainting	112
11.1	Class represented in BRACS dataset	116
11.2	Examples of BRACS dataset in different resolutions	116
11.3	Examples of BRACS dataset and its annotations	117
11.4	Example of BCSS dataset with segmentation masks	118
11.5	Example of controlled synthetic whole slide image with segmentation masks (made with our proposed synthetic image generation pipeline)	119
11.6	Graphs of VQVAE training	129
11.7	Reconstructed image patches examples	130
11.8	Synthetic data patch examples	131
11.9	Conditioned synthetic data patch examples	134
11.10	Example 1024 x 1024 synthetic region (left: mask legend, center: semantic mask, right: synthetic image)	135

11.11	Example 4096 x 4096 synthetic region (left: mask legend, center: semantic mask, right: synthetic image)	136
11.12	Examples of synthetic whole slide images with their semantic masks	138
11.13	SAM seam visibility (left: instance map; right: SAM boundaries + tile grid).	139
11.14	SAM seam recall and boundary-on-seam precision (B1–B8).	140
11.15	Felzenszwalb superpixels for B1 (tiling) vs. B5 (ours).	141
11.16	Felzenszwalb seam metrics (B1–B8).	142
11.17	LPIPS comparison of wavy vs. straight seam masks	142
11.18	Selected Results of the Structure Search Task	144
11.19	Graphs of RA-HIPT training	145
11.20	Example of patches used for inference of the model	146
11.21	Visualization of per head attention module of trained RA-HIPT . .	147
11.22	Visualization of heatmap made of attention weights from Graph RA-HIPT	148
11.23	Visualization region to region relations derived from the regularization graph structure	149
11.24	Analysis of embeddings collapse through the training	150
11.25	Visualization of PCA of region embeddings and Region graph degree distribution	151
11.26	Visualization of heatmap made of attention weights from Graph RA-HIPT on Controlled synth dataset	153
11.27	Semantic mask used for generation of wsi	154
11.28	Visualization region to region relations derived from the regularization graph structure	155
11.29	Confusion matrix of classification over synthetic dataset	156

11.30 Visualization of heatmap made of attention weights from Graph RA-HIPT on BRACS dataset WSI	157
11.31 Visualization of PCA of region embeddigns and Region graph degree distribution	158
11.32 Visualization region to region relations derived from the regulariza- tion graph structure	159
11.33 Visualization of class relations trained by the Graph RA-HIPT ar- chitecture	160
12.1 Preview of HistoView tool	165
12.2 Preview for synthetic structure analysis use case	166
12.3 Preview for biomarker detection with heatmap analysis use case . .	167

List of Tables

11.1 Comparison of sampling techniques for diffusion model generation .	129
11.2 Comparison of sampling techniques for diffusion model generation .	131
11.3 Comparison of sampling techniques for conditioned diffusion model generation over 10 000 patches	133
13.1 Nepodmienená latentná difúzia: porovnanie vzorkovačov	183
13.2 Podmienená generácia na BCSS: porovnanie vzorkovačov (metriky na 10 000 výrezoch).	184
A.1 Plan of the work for first semester	A-1
A.2 Plan of the work for second semester	A-2
D.1 Electronic appendix structure	D-2
D.2 Source Code directory layout	D-3
D.3 In-repository documentation (beyond this appendix)	D-4
D.4 RA-HIPT runtime prerequisites	D-4
D.5 RA-HIPT: what to configure before training or inference	D-5
D.6 RA-HIPT packaged demo (Google Drive assets)	D-6
D.7 RA-HIPT: other useful CLI tools (flags in each file)	D-7
D.8 WSI.Generator runtime prerequisites	D-7
D.9 WSI.Generator: where checkpoints and masks are expected	D-8

D.10 WSI.Generator: common inference and dataset pipelines	D-10
D.11 Shared layout between RA-HIPT and WSI.Generator	D-11
D.12 Where validation artefacts live and routine upkeep	D-12

List of used abbreviations

AI	Artificial Intelligence
WSI	Whole Slide Image
ROI	Region Of Interest
CNN	Convolutional Neural Network
VAE	Variational Autoencoder
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection
MIG	Mutual Information Gap
DCI	Disentanglement, Completeness, Informativeness
Grad-CAM	Gradient-weighted Class Activation Mapping
LRP	Layer-wise Relevance Propagation
BoQ	Bag of Queries
VQ-MIL	Vector Quantized Multiple Instance Learning
VQ-VAE	Vector Quantized Variational Autoencoder
DTFD-MIL	Dual-Task Feature Distillation for Multiple Instance Learning
TransMIL	Transformer-based Multiple Instance Learning
CLAM	Clustering-constrained Attention Multiple Instance Learning
TCGA	The Cancer Genomic Atlas
AUC	Area Under the Curve

DDPM	Denoising Diffusion Probabilistic Models
DDIM	Denoising Diffusion Implicit Models
PLMS	Pseudo Linear Multistep Methods
FID	Fréchet Inception Distance
IS	Inception Score
H&E	Hematoxylin and Eosin
HIPT	Hierarchical Image Pyramid Transformer
RA-HIPT	Relation-Aware Hierarchical Image Pyramid Transformer
BRACS	BReAst Carcinoma Subtyping Dataset
BCSS	Breast Cancer Semantic Segmentation Dataset
ViT	Vision Transformer
GAN	Generative Adversarial Network
EMA	Exponential Moving Average
U-Net	U-Net Architecture
DDM	Denoising Diffusion Models
MIL	Multiple Instance Learning
XAI	Explainable Artificial Intelligence
KL	Kullback–Leibler Divergence
ELBO	Evidence Lower Bound
AdamW	Adam with Weight Decay

ResNet	Residual Network
DINOv2	Self-Supervised Vision Transformer
MSLS	MapillarySLS Dataset
LDM	Latent Diffusion Model
CIFAR-10	Canadian Institute For Advanced Research 10-class image dataset
CIFAR-100	Canadian Institute For Advanced Research 100-class image dataset
CT	Computed Tomography
DINO	Self-Distillation with NO labels
F1	F1 Score
GPU	Graphics Processing Unit
GSV	Google Street View
LPIPS	Learned Perceptual Image Patch Similarity
MLP	Multilayer Perceptron
MNIST	Modified National Institute of Standards and Technology dataset
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
SAM	Segment Anything Model
SSIM	Structural Similarity Index Measure
SSL	Self-Supervised Learning

TCGA-BRCA The Cancer Genome Atlas Breast Invasive Carcinoma cohort

UI User Interface

VGG-16 Visual Geometry Group 16-layer network

VQ Vector Quantization

Chapter 1

Introduction

Although histological images are unfamiliar to most people, their analysis plays a critical role in identifying biomarkers that ultimately save countless lives. These histological images are feature-rich and contain an extreme number of details. Because of this, their processing is complex and time-consuming [58]. Neural networks are being used more and more often in this domain [34]. With the help of AI, pathologists are able to save from 20% to 90% of the time needed for image analysis [17]. In diagnosis, every second matters, and that is why this time-saving approach is life-saving. However, this technology does not have only advantages; it also has some disadvantages, one of which is AI interpretability.

Early **biomarker detection**, which is one type of histological image processing, can reveal important information and show the exact path for treating specific types of diseases.[115] That is why AI has started to be used more frequently. AI is capable of finding relations that are not visible at first sight and can even process the entire image.[120] These relations are often found in image areas that are far apart, and the model is not able to detect them, which causes the AI model to lose an important concept of the disease. That is why we need to focus more

on improving AI-driven biomarker detection with an focus on **global and local feature relations**.

Interpretability became one of the most mentioned problems in the field of medicine.[31] From the histological image point of view, it is crucial for the doctor to be able to explain how the AI system, which performed the prediction, decided, especially when it comes to the detection of new biomarkers. The core problem is that the patient has the right to request this decision. Without knowledge of how the system decided, this system cannot be used in a production environment, and the doctor cannot use its results for diagnostics.[90]

The contribution of this work is divided into two main parts. The first part focuses on integrating **global and local feature relations** into the architecture to enable it to understand both the overall whole-slide image context and the fine details of the image. To verify the hypothesis that these relations can be identified and meaningfully connected, attention-based heatmap visualizations are generated over the learned embeddings, allowing for the interpretation of the model’s behavior. The second part involves the creation of a **conditioned synthetic dataset** designed for validating the previously mentioned hypothesis. The dataset is assembled by arranging predefined structures at precise spatial positions. This synthetic dataset is then used to support and confirm the proposed relational modeling approach. The overall pipeline aims to validate the solution for biomarker detection.

1.1 Motivation

To understand the challenges of the pathologists’ diagnostic process, we first consider the proposed examination flow illustrated in Fig. 1.1. In comparison, the traditional image processing workflow contains AI-driven analysis, which is not

interpretable and processes the image patch by patch. The key components that differ from the traditional workflow are highlighted, and we now discuss the motivation behind them.

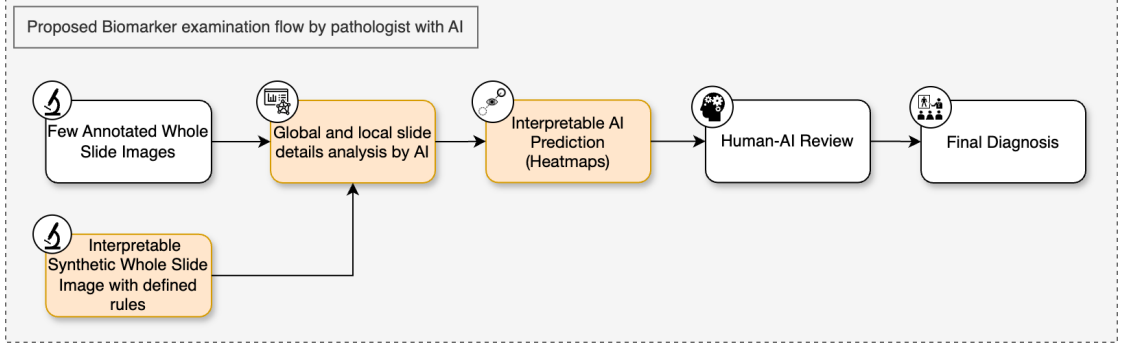


Figure 1.1: Proposed image processing workflow of the histopathologist, with highlighted sections indicating the areas of focus in this study

The first step involves **global and local slide analysis by AI**, focusing on the integration of global and local feature awareness into the model. These relationships are essential in the diagnostic process, as pathologists evaluate the slide holistically rather than as isolated regions. This step also represents the pathologist’s process of analyzing the slide at multiple magnifications, which closely reflects the real diagnostic workflow.[39, 86]

In many real-world settings, the limited availability of annotated whole-slide images makes it difficult to validate architectural hypotheses. Large whole-slide images and the high demand for expert annotation slow down the process of validation and AI model training.[72] In particular, our work aims to test whether a neural network can effectively integrate local and global features within a single model.[21] To enable controlled and interpretable validation of this hypothesis, we introduce a second step based on **interpretable synthetic whole-slide images** generated according to predefined rules and concepts.[79] This approach allows us to directly assess whether the model captures the important relations, providing a

technique for hypothesis-driven validation that is not possible with real-world data alone. In the other words we create **controlled synthetic validation dataset for down-stream models**.

The third component is the **interpretable AI** prediction stage, where heatmaps provide essential insights for biomarker validation.[42] Pathologists must understand which slide regions contribute to the model’s decision; without this information, they cannot assess whether the model relies on meaningful features.[53] By incorporating these steps, the models used by pathologists become more interpretable and better aligned with their natural decision-making process, which is crucial for reliable diagnostic support.[53]

These hypotheses and workflow modifications are considered mainly from a user experience point of view. The entire proposed pipeline and the change in the pathologist’s workflow represent a different research question and need to be further validated in the future, after the proposed neural network architecture modifications are implemented.

1.2 Research Questions

Based on the motivation and problem background we stated these three research questions:

RQ1: Can a neural network architecture that explicitly integrates local and global features capture predefined local-global relationships in whole slide images?

Whole slide images contain information at multiple spatial scales, where local patterns are often meaningful only in the context of global spatial organization. Conventional architectures typically focus on either local or global features, which

limits their ability to model interactions between these scales. This research question directly addresses the central hypothesis of this thesis: whether explicitly integrating local and global representations within a single neural network enables the model to capture such structured relationships.[39, 86]

Hypothesis:

- H_0 : Integrating local and global features within the neural network architecture does not improve the model’s ability to capture predefined local-global relationships in whole slide images.
- H_1 : Integrating local and global features within the neural network architecture improves the model’s ability to capture predefined local-global relationships in whole slide images.

RQ2: To what extent do interpretable synthetic whole slide images enable controlled validation of local-global feature integration compared to real-world data?

Validation of local-global feature integration is challenging when using real-world whole slide images, as the true underlying relationships between local and global structures are unknown and difficult to isolate. Interpretable synthetic whole slide images, generated according to predefined rules and concepts, provide full control over the data generation process. This allows explicit encoding of known relationships, making it possible to assess whether a model captures the intended interactions. This research question motivates the use of synthetic data as a validation tool rather than a replacement for real data.[21, 79]

Hypothesis:

- H_0 : Interpretable synthetic whole slide images do not provide a more controlled or informative framework for validating local-global feature integra-

tion than real-world whole slide images.

- H_1 : Interpretable synthetic whole slide images provide a more controlled and informative framework for validating local-global feature integration than real-world whole slide images, due to their predefined rules and explicitly encoded local-global relationships.

RQ3: Can a neural network that integrates local and global features produce representations and outputs that are both interpretable and explainable, enabling expert validation of the learned local–global relationships?

Neural networks that integrate local and global features aim to model multi-scale relationships in whole slide images, yet it remains unclear whether these relationships are learned in an interpretable and explainable manner. In digital pathology, such explainability is essential to allow expert validation of model behavior. This research question therefore investigates whether a model integrating local and global features can produce explainable representations that enable meaningful validation by a pathologist, using synthetic whole slide images as a controlled evaluation setting.[53]

Hypothesis:

- H_0 : A neural network that integrates local and global features does not produce interpretable or explainable representations that allow meaningful expert validation of the learned local–global relationships.
- H_1 : A neural network that integrates local and global features produces interpretable and explainable representations that allow meaningful expert validation of the learned local–global relationships.

Chapter 2

Medical Imaging

In the last years, deep neural networks started to be one of the most useful tools in hospitals as an intelligent assistant for image processing. We know multiple different types of data used in the field of medicine. Some of them are image data and numerical data. Both of these data can be processed by deep neural networks. Numerical data can represent exact numerical values of medicals.[58] [34]

Image data can exist in a wide variety of formats, each used for a different purpose. Some of them are magnetic resonance imaging (MRI), computed tomography (CT), X-ray, ultrasound, and photographic images.[1] These data contain biomarkers that are crucial for the diagnostic process. In our work, we focus primarily on image data, specifically digital histopathology images, an example of which can be seen in Fig. 12.1. These data come with multiple processing methods and a wide range of challenges to address, such as biomarker detection.

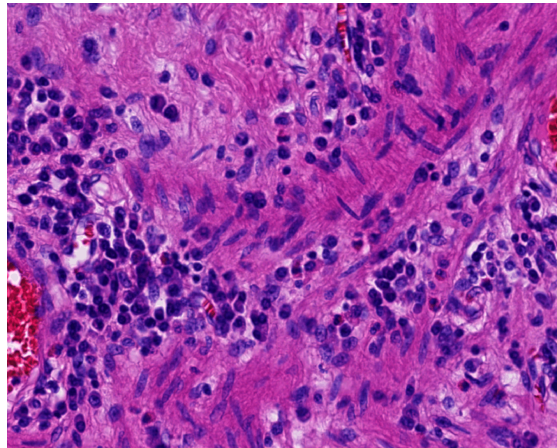


Figure 2.1: Example of histology image

2.1 Biomarkers

Biomarkers are measurable biological indicators that reflect biological processes, disease progression, or response to therapy. They can exist in different structures such as cells, molecules, or proteins. A combination or even a single biomarker can indicate the presence of a disease or reveal underlying biological processes. Nowadays, digital biomarkers are often combined with biological ones, but their detection remains a challenging task.[81, 122]

One type of biomarker is the outcome-driven biomarker. These biomarkers focus on predicting the clinical outcomes of patients. Some of these outcomes include disease progression or treatment response, which can be guided by biomarker analysis performed by doctors. Types of these biomarkers include prognostic, diagnostic, and predictive. Prognostic biomarkers measure biological characteristics. Diagnostic biomarkers measure biological molecules reflecting the presence or absence of disease. Finally, predictive biomarkers indicate whether a patient is likely to respond to treatment.[81, 100]

The second type of biomarker is the systems biology-driven biomarker. These

aim to understand biological systems as a whole, processing multiple sources of data. Such sources include genomics, transcriptomics, proteomics, and other omics data, which together create structures or relationships such as network biomarkers. Types of these biomarkers include phenotype, molecular, and cross-tiered. Phenotype biomarkers measure characteristics and traits based on which classification can be made. Molecular biomarkers can measure and indicate genetic mutations or circulating tumor cells. Finally, cross-tiered biomarkers integrate multiple levels of biological information used for a more complex understanding of diseases.[81, 100]

Nowadays, biomarkers are starting to be uncovered by AI. It is easier to use advanced algorithms to analyze huge datasets, rather than for doctors themselves. AI can uncover more innovative and surprising connections. For these types of tasks, it is extremely important to use explainable AI, so doctors can understand how AI performed the detection. The process of AI integration into the biomarker detection pipeline can be seen in Fig. 2.2. These concepts show that it is important and crucial to use AI for biomarker detection, as it can reveal relations that cannot be seen at first sight.[8, 38]

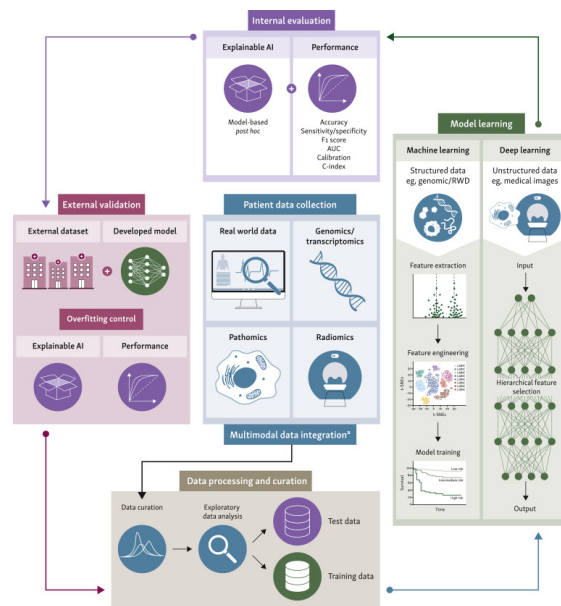


Figure 2.2: Biomarker detection with a use of AI [80]

2.2 Digital Histopathology

The core thing we need to focus on and understand is the data that are generated in the field of digital histopathology. These data are high-resolution digital images of tissue sections. They are created with a multi-step process. A section of the tissue is extracted, placed in paraffin, stained, and afterward sliced and placed between glass slides 2.3. In this phase, the slide can be scanned by scanners, so-called whole slide image (WSI) scanners. After scanning, we have a digitalized image of the tissue ready to process. This preparation step allows us to select a task that arises from a problem that we are trying to solve.[68, 106, 47]

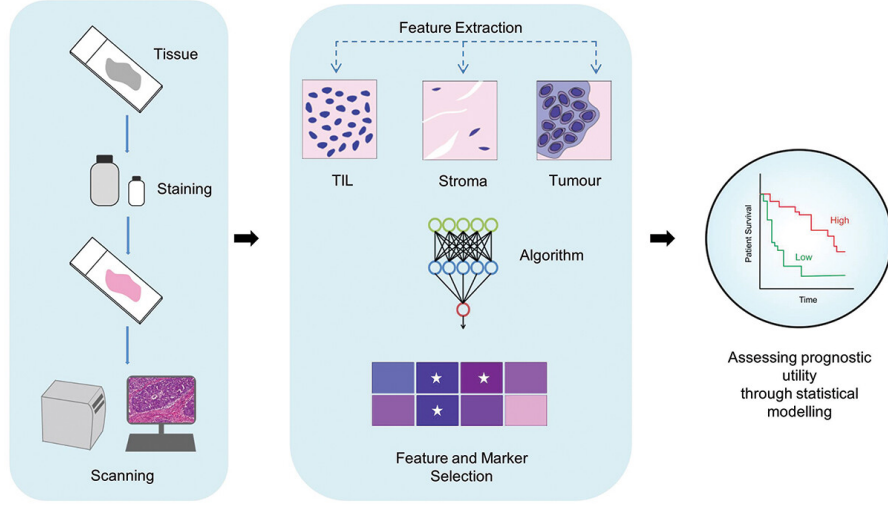


Figure 2.3: Diagram of the histological tissue processing [68]

2.3 Types of Tasks

The nature of the task reflects the problem that we are trying to solve. This creates, especially in the field of histopathology, opportunities where different deep learning tasks can be applied.

Classification is the first and most common task that is performed on data. The main goal of classification is to assign a label to a data sample. This is especially important for computer-aided diagnosis and distinguishing between healthy and non-healthy patients. Classification can be binary or multi-class, depending on the nature of the problem and the number of classes to classify. From the histopathology point of view, there is a vast number of research problems where classification can be applied, such as the classification of cells, tissues, tumors, or different structures located in the tissue.[63, 77]

Segmentation is another task that is also common in the field of medicine. The algorithm in this task tries to assign a class not to the whole image but to each

pixel of the image. This process outputs a segmentation mask, which divides the image into parts that can be later analyzed and processed. This task is important in situations when we do not just need to classify whether there is something, but we also need the exact position and pixel-precise borders of the object. From the histological point of view, segmentation can be used for the segmentation of cells, tissue types, tumors, and more. These segmented objects can be used for precise quantification, understanding of tissue structure, or further analysis.[98, 27]

Detection is a similar task to segmentation, but in this task, we do not need pixel-wise classification. We only need to know where the object is located. This is especially important for the localization of regions of interest. Detection is represented as bounding boxes surrounding the object. In histological images, detection can be used for counting and locating cells, locating tumor zones, or even identifying events in the tissue.[27, 5]

Image Generation is also an important task when it comes to the domain of histology. At first sight, it may not be obvious where it can be applied, but it has a crucial role. Generation of histological images is important when it comes to one of the main problems of histology, data. In situations where we lack annotated data, synthetic data with labels can be generated. Because of this, image generation is used in the field of histopathology for generating labeled data or even generating different types of image stainings.[67, 105]

2.4 Challenges

The complexity of histological images comes with its own challenges, and it is important to be aware of them when working with these data.[47]

Not Enough Annotation is a core problem in digital histopathology. Creating annotations is time-consuming and highly domain-specific. The problem stems from the pathologist’s lack of time to create annotations. Making annotations is not part of their main job, which makes this even more complicated. Because of that, the vast number of created datasets is small, which makes deep neural network training harder. In these situations, techniques like self-supervised or active learning are used.[42]

Image Size is one of the first challenges when it comes to histopathological data. The common size of one WSI is 100 000x100 000 pixels, which makes it hard to process. Because of this, preprocessing techniques like tiling or patch extraction need to be applied.[50, 106]

Image Staining is another challenge that comes out of the histopathology data nature. Each laboratory uses different scanners and different staining intensities, which creates different colors of digitized data. Even if we had the exact same tissue slide and it would be stained and scanned in different laboratories, it would be different. Because of this, stain normalizations need to be used when processing these data.[50, 102]

Tissue and Cell Variability is a challenge that is highly connected to the features of histopathological images. They contain a vast amount of fine details, which are invisible to the untrained eye. Differences in shape, size, and appearance across patients, organs, and types of diseases make generalization highly difficult.[102, 58]

Explainability and Trust is one of the most significant challenges. In many situations, pathologists make life-changing decisions. When deep learning enters

the decision-making process, it needs to be explainable; pathologists need to know how it decided, and the system also needs to be trustworthy. The main problem is that in most deep learning applications, these requirements are not met. Even the pathologist is not able to provide their insight directly into the application, which makes these applications limiting and unreliable.[103, 35, 31]

Chapter 3

Deep Learning

Nowadays, deep learning has a leading role in the progress of technologies in nearly all sectors of everyday life. Some of the fields where deep learning is applied are more robust and some of them less. One of the biggest of all of them is the medical sector. In medicine, deep learning is facing a vast amount of problems. Few of them are prejudiced doctors on weakly interpretable results of used models, a small amount of data for training, and a huge amount of feature-rich data[75]. Especially histological data, which are one of the most complex and feature-rich.[82]

The primary method of processing histological data is the use of deep neural networks. These neural networks are used for tasks like segmentation, object detection, or classification. For extracting features and classification, state-of-the-art networks such as convolutional neural networks and transformers are used. Both of them might be supervised or self-supervised.[82] This highly depends on the amount of data that we have for training.

All the mentioned architectures work with a so-called latent space, which they also create[73]. This space is a low-dimensional representation of the input data.

The architectures that are most widely used for creating these low-dimensional representations are encoders.

3.1 Convolutional Neural Network

Convolutional neural networks (CNNs) are one of the most used models in image processing. They are used for classification, object detection, or recognition tasks. These models automatically learn the hierarchical features of the image.[101]

$$y[n] = (x * h)[n] = \sum_{m=-\infty}^{\infty} x[m]h[n - m] \quad (3.1)$$

For this process, convolution, whose equation can be seen in Eq. 3.1, is used. The core parts of this equation are $x[m]$, which represents the input signal, and the $h[n - m]$ representing kernel. The parameter n is also important and represents the output time index.[25]

$$G[i, j] = \sum_{u=-k}^k \sum_{v=-k}^k I[u, v] K[i - u, j - v] \quad (3.2)$$

The second implementation can be seen in Eq. 3.2, which represents a convolution layer. The equation is derived from cross-correlation. The $I[u, v]$ part represents the image, and the $K[i - u, j - v]$ represents the kernel, where u, v, j and i represents input image and kernel pixel locations.[104]

Most CNN architectures contain four main layers: convolution, pooling, activation, and a fully connected layer. An example of this architecture can be seen in Fig. 3.1. The convolution layer is the core component of the network; it applies a filter (kernel) to the data, which creates feature maps. The pooling layer is

used for down-sampling the image while preserving the most important features. The activation layer adds non-linearity to the network and controls the signal flow between layers by applying an activation function to the output of each neuron. The last fully connected layer is used for performing the final classification of the extracted features.[55]

Another advantage of convolution is that it can be used in a fully-convolutional network, which allows the model to process input of various sizes and classify image regions or even individual pixels.

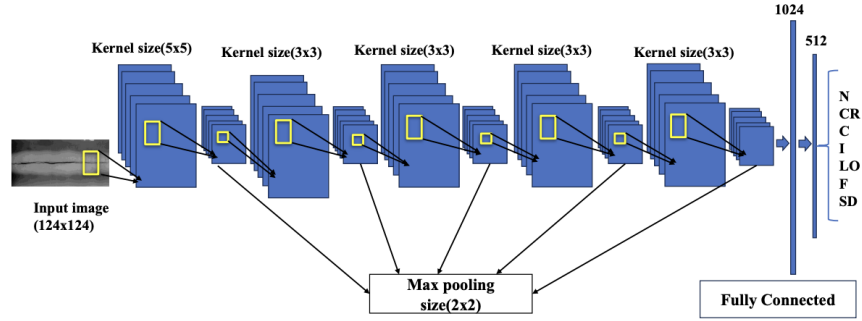


Figure 3.1: Example of convolution neural network [55]

3.2 Vision Transformers

With the occurrence of transformers used for natural language processing tasks, the question arises whether this type of architecture can be used for computer vision tasks. It turned out that not only can transformers be used for image processing, but in some cases, they even achieve much better results than state-of-the-art CNNs.[116, 32]

The architecture of vision transformers is not much different from ordinary transformers. The main difference is that the vision transformer does not contain a decoder part, which the ordinary transformer contains. The encoder, the part

that remains, is made of seven different components: image splitting into patches, patch embedding, normalization, multi-head attention, skip connection, and multi-layer perceptron. The whole architecture of the vision transformer can be seen in Fig. 3.2. [32, 108]

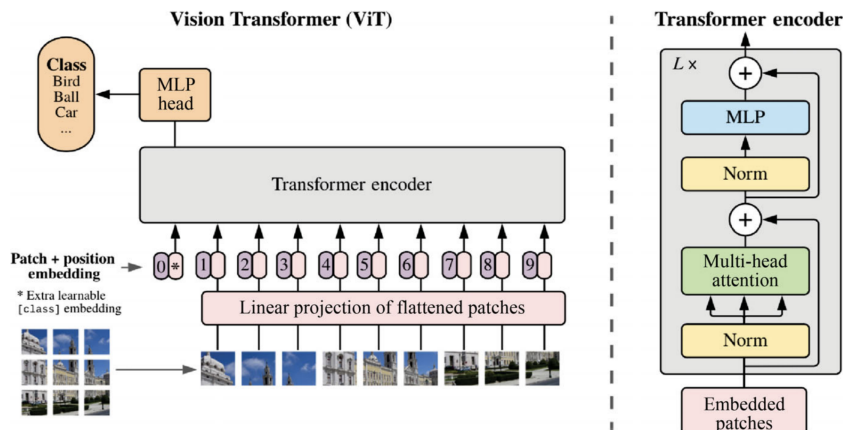


Figure 3.2: Example Vision Transformer architecture [116]

Image splitting to patches converts a whole image into a sequence of flattened patches, similar to tokens in original natural language processing transformers. The concept is not about convolving over local receptive fields like CNNs do, but about operating over fixed-size image tokens in sequences.

Patch embedding can be divided into two main components. It combines normal patch embeddings and positional embeddings, so the spatial dimension is not lost. Overall, patch embedding is responsible for projecting patch vectors into a higher-dimensional embedding.

Normalization in the context of vision transformers is an important component. It is responsible for training stability, convergence speed, and overall model

performance. This is done by normalizing input data or activations so they have a standardized distribution.

Multi-head attention is the core concept of transformer architecture. This block lets the model capture different parts of patches and capture global dependencies. Each head of this attention block is responsible for capturing different relations and contextual patterns. Behind each attention head is a calculation of attention, which can be seen in Eq. 3.3. The value Q means query, question to the token. Another value K represents the key, corresponding to the answer to the token question. The third component V represents the core meaning of input embedding. At the end, value d represents the dimensionality of query and key vectors.[108]

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d}} \right) V \quad (3.3)$$

Skip connections are the same technique that was introduced in the ResNet architecture. Their goal is to improve the flow of gradients and avoid the vanishing gradient problem. In other words, they help the model learn identity maps and support the gradient flow in deeper architectures.

Multi-layer perceptron is the last layer of the vision transformer architecture. It processes the output of the encoder containing multi-head attention and enriches feature representations. Another property provided by this layer is non-linearity and higher capacity for feature transformation.

The main **advantages**[32] of vision transformer architecture beside CNN's are:

- Does not contain inherent biases like CNNs
- Architecture is more flexible

- Easily adaptable for downstream tasks
- Capture global dependencies
- Does not sacrifice spacial resolution on deeper layers

Chapter 4

Feature Space

In recent years, a vast number of deep neural network architectures have been developed, but only a few of them are focused on explainable and controlled working with the latent space. Latent space is an important part of deep learning because it contains all important features. Working with these features is a core concept when it comes to a better explainable model, especially in the field of histopathology. Understanding these features encoded in the latent space is the first step in understanding how the model has decided. Because of this, latent space undergoes dimensionality reduction, which makes it easier to visualize and also interpret.[91, 36]

Interpretable latent space is a core element in the domain of histopathology, but it is not easy to generate. Most representative latent spaces are produced by encoder architectures through the process of creating a lower-dimensional representation of the input. However, due to the unknown and complex nature of this process, which is a crucial drawback in the field of medicine, an extensive area for research has emerged.[36, 36, 57]

4.1 Encoders

The encoder architecture is a special type of neural network architecture used to transform input, such as text, images, or audio, into low-dimensional representations. This representation is also called latent vectors or embeddings. These latent vectors together create the latent space, which contains the most important features of the input data. Usually, encoders are paired with decoders, which perform the opposite task and convert the latent vector back to the original image.[13]

These types of architectures have a number of crucial problems. The first problem is an information bottleneck. In this stage, the encoder compresses data into a fixed-size latent vector, which often leads to the loss of important information. Another problem is that these models commonly suffer from overfitting and poor generalization, especially those with deep architectures. Additionally, these models often suffer from poor interpretability, which is a common issue when working with latent vectors.[13, 23]

4.1.1 Autoencoders

Similar to encoders, autoencoders share a common concept with them. They contain two separate parts: an encoder and earlier-mentioned decoder, which is responsible for reconstructing the input data from latent vectors. For this task, they use a special reconstruction loss, typically implemented as mean squared error. An example of this architecture can be seen in Fig. 4.1. Autoencoders are commonly used for tasks such as dimensionality reduction, noise reduction, anomaly detection, and as part of generative models.[13]

Most of the problems that ordinary encoders have can be addressed by autoencoders and their variants. For the information bottleneck, β -Vae can be used, which creates richer representations. To address overfitting and poor generaliza-

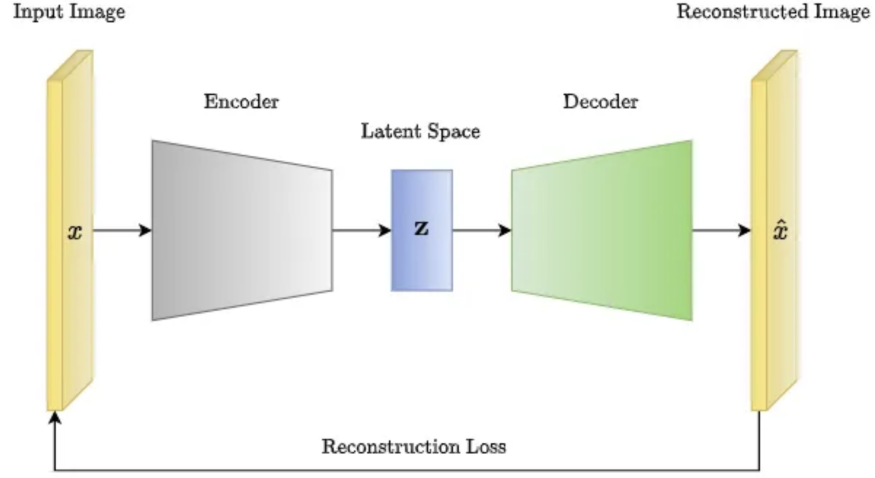


Figure 4.1: Example of autoencoder architecture [95]

tion, denoising autoencoders are employed, as they can learn a large number of features. Interpretability can be improved with disentangled autoencoders, which create a more separate and interpretable latent space.[13, 23]

4.1.2 Variational Autoencoders

Variational autoencoders are derived from ordinary autoencoders. The core difference between variational autoencoders and autoencoders is that variational autoencoders introduce probabilistic modeling. This allows the model to generate new data similar to what it was trained on. The architecture is primarily used for tasks such as the generation of images, sounds, or text samples.[73]

The key ideas behind variational autoencoders can be divided into three parts. The first part is the representation of encoded inputs in the latent space as a probability distribution. This gives the model more generative power, creates a smoother and more continuous latent space, and improves generalization. The second part is the reparametrization trick, which helps the model with backpropagation because

sampling from a distribution is not differentiable. The last part is its loss function, which consists of two components. One is the reconstruction loss, which measures the difference between input and output, and the second is the Kullback-Leibler divergence, which pushes the learned distribution to be closer to a standard normal distribution. All of the mentioned components can be seen in Fig. 4.2, which together represent the architecture of the variational autoencoder.[30, 41]

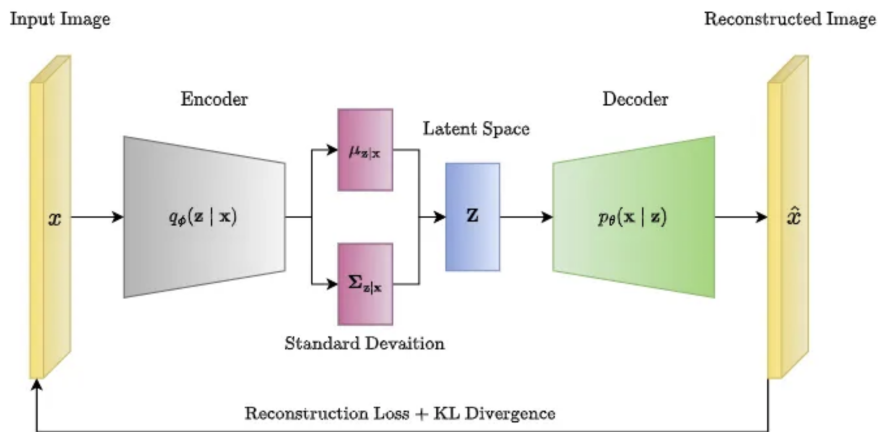


Figure 4.2: Example variational autoencoder architecture [95]

Overall, the model samples random points from a Gaussian distribution, feeds them into the architecture, and generates new data from them. Another advantage of this architecture is the well-structured and continuous latent space, which makes it more disentangled and representative.

4.2 Latent Space

Latent space is one of the most important concepts in deep neural networks, and it is highly connected to data features. It is a low-dimensional space that represents high-dimensional complex data. Latent space is created by neural network layers, mostly autoencoders. The process of latent space creation can be seen in Fig. 4.3.

The advantage of latent space is that it represents the most important features of the data. However, individual features are not obvious in the latent space and need to be identified through experiments to determine which part of the latent space represents which feature.[73, 91]

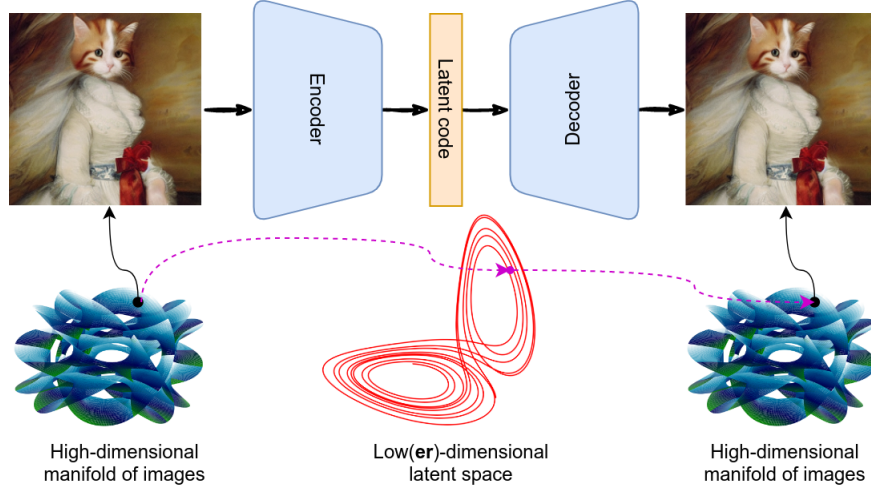


Figure 4.3: Process of the latent space creation [74]

The whole concept of interacting with latent space is promising and can be beneficial not only for the interpretation of features but also as an option to modify individual features in the domain of histology, where data are rich in features. [10]

4.3 Feature Importance

The most interesting and complex part of the latent space is the features embedded within it. In the past, feature extractors were hand-crafted, which made them easier to interpret but, on the other hand, time-consuming to create. Nowadays, deep neural networks perform this much faster but at the expense of explainability. Features carry information such as shape, color, texture, and more. It is not easy

to clearly determine which part of the latent space is responsible for which feature, as can also be seen in Fig. 4.4.[91]

The core problem is that features inside the latent space are entangled. Because of this, it is not trivial to extract and modify a single feature in the latent space, as a change in one feature can affect others. This entanglement makes working with features challenging. A crucial task in the medical domain is to determine how to unravel the connections between features and understand how they are related. The similarity of these features can be measured using the distances between them in the latent space. Some techniques for calculating these distances include the Euclidean and Mahalanobis distances.[91, 93]

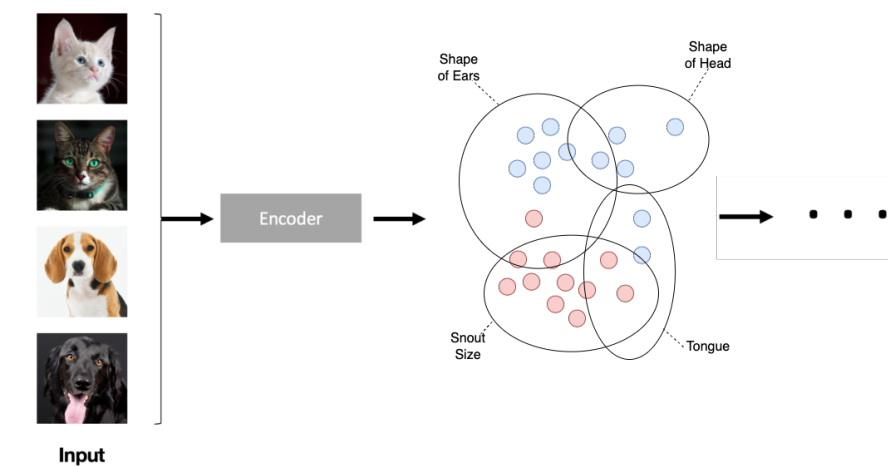


Figure 4.4: Visualization of disentangled latent features

4.3.1 Mahalanobis Distance

As mentioned, it is common and crucial to calculate the similarity of individual points in latent space distributions. Unlike the Euclidean distance, which treats each feature equally, the Mahalanobis distance accounts for the variance and correlation of the data. In Equation 9.15, $\mathbf{S}^{-1}$ represents the inverse of the covariance

matrix, ensuring that features with large variance contribute less to the distance. The second term, $\boldsymbol{\mu}$, represents the center of the distribution. This equation ensures that distance is calculated relative to the center of the distribution rather than the origin. A visual representation of how the Mahalanobis distance works can be seen in Fig. 4.5, where it is clearly shown that the distance aligns with the data distribution and does not rely on a fixed radius.[70, 26]

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \quad (4.1)$$

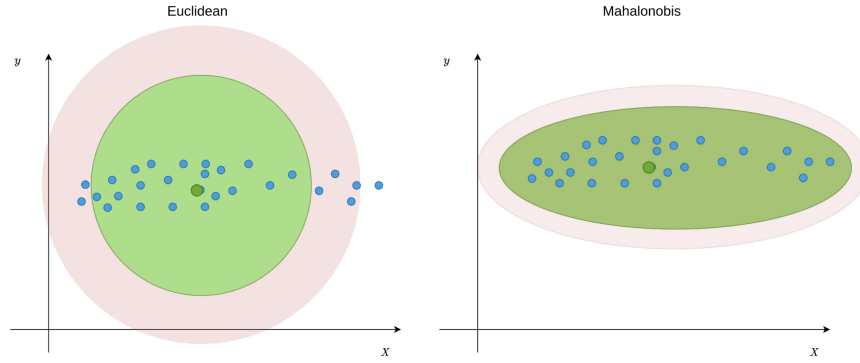


Figure 4.5: Difference between Euclidean and Mahalanobis distance [18]

Mahalanobis distance is mainly used in classification problems where there are several highly correlated groups, and features or classes have different spreads. It also handles scale differences and is sometimes used for outlier detection. Based on these properties, Mahalanobis distance tends to be highly advantageous in the medical domain, where features are often strongly correlated.

4.4 Dimensionality Reduction

Histological images contain a vast number of features, some of which are easier to detect and some harder. However, it is not a trivial task to determine which

features are the most important. In high-dimensional space, data become sparse, and distance measures like Mahalanobis lose their meaning. These are a few of the reasons why dimensionality reduction is used. Other purposes for using dimensionality reduction include addressing the curse of dimensionality, reducing overfitting, improving computational efficiency, enabling visualization, and reducing noise.[97, 3]

Methods of dimensionality reduction can be divided based on the type of data they process. The first are linear methods, which rely on the linear structure of the data and assume that important structure lies on or near a linear subspace. The most common examples of linear methods are principal component analysis and linear discriminant analysis. The second are non-linear methods, which aim to unfold or flatten the curved structure of latent space into a low-dimensional one. Examples of these methods include t-distributed stochastic neighbor embedding and uniform manifold approximation and projection. In addition to these methods, neural network architectures such as encoders are also used for dimensionality reduction and feature extraction.[97]

Principal Component Analysis is one of the best-known linear methods used for dimensionality reduction. It transforms a dataset with correlated features into a set of linearly uncorrelated features, also known as principal components. It uses an orthogonal linear transformation to project data onto a new coordinate system. This transformation ensures that the first component has the highest possible variance, and each succeeding component has the highest possible variance while being orthogonal to its preceding component. These calculations are performed using the eigenvalues and eigenvectors of the data covariance matrix or through singular value decomposition. A visualization of this method can be seen in Fig. 4.6.[16, 4]

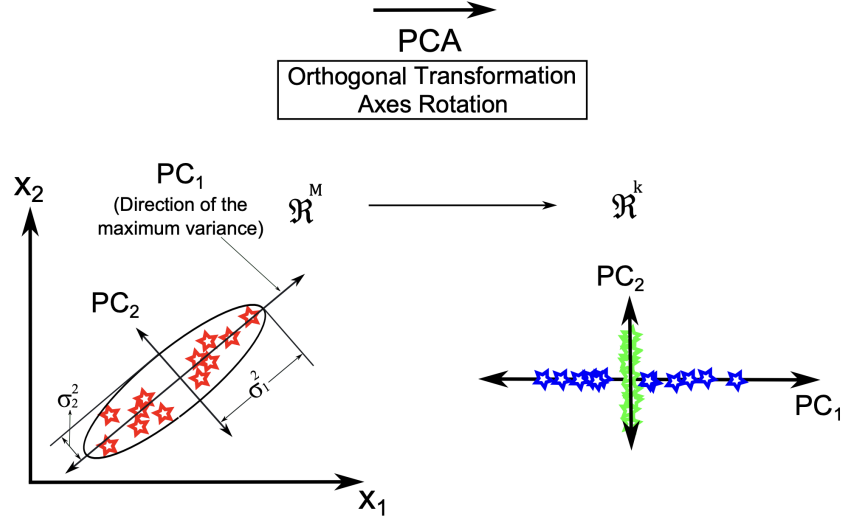


Figure 4.6: Example of PCA dimensionality reduction[4]

Uniform Manifold Approximation and Projection is, on the other hand, one of the best-known non-linear methods. By its nature, the algorithm preserves both the global and local structure of the data, making it effective for the visualization of high-dimensional data. Some of these visualizations can be seen in Fig. 4.7. The algorithm first constructs a weighted k-nearest neighbor graph to capture the local structures. After this, it optimizes low-dimensional embeddings to preserve the topological structure of the graph. Another advantage of UMAP is that it effectively scales to large datasets.[40, 69]

Dimensionality reduction is not only advantageous, but some of its aspects need to be carefully considered. Loss of interpretability when features are transformed, unclear meaning of new features, or potential loss of important features are some of these aspects. Because of this, it is important to carefully decide when to use it. Appropriate moments is before clustering or classification, for noise reduction and visualization, or when features are redundant and highly correlated.

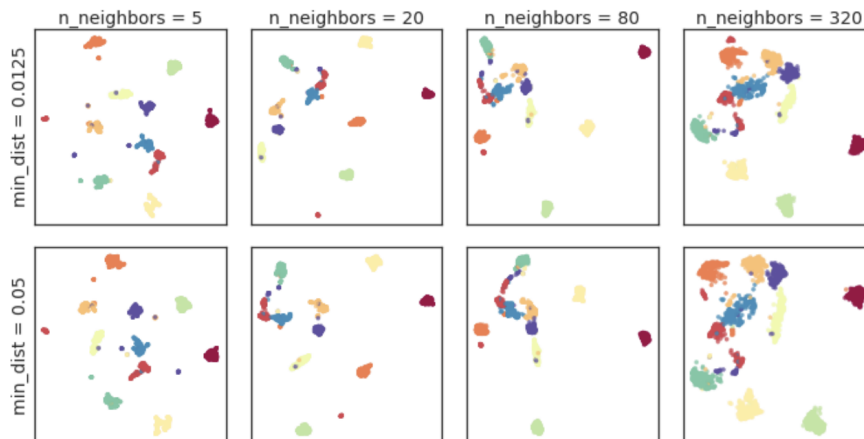


Figure 4.7: Data visualizations after dimensionality reduction with UMAP method

4.5 Interpretable Latent Space

Interpretable latent space gives us the opportunity to understand the decision-making process of neural networks. The most important question is: when can we say that a latent space is interpretable? A latent space is considered interpretable when its dimensions and directions align with meaningful, human-understandable concepts. This state of the latent space can also be described as a disentangled latent space.[56, 91]

If each latent dimension corresponds to a known attribute, we are able to understand and trust the entire model. By increasing the interpretability of the latent space, we can say that the model’s transparency and usefulness also increase. From the perspective of medical imaging, a fully interpretable latent space unlocks the potential for discovering patterns and connections between features, both known and unknown to pathologists.[91]

4.6 Disentanglement

The main goal of disentanglement is to encode variations of features into separate latent variables. Each of these variables can represent different features, such as shape, color, or texture. The aim of disentanglement is to enable changes in individual latent channels, which results in the modification of a single feature. By altering these separate feature vectors, different aspects of the image can be changed. An illustration of this process can be seen in Fig. 4.8.[56, 61, 99]

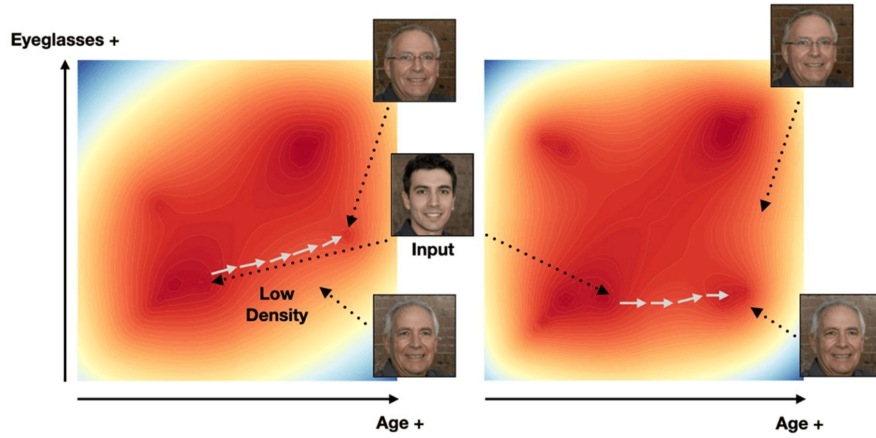


Figure 4.8: Changing of feature vectors of latent space trained on faces

There are a few benefits that disentanglement provides. **Invariance** is one of them. It is easier to extract representations that are invariant because they can be easily marginalized along dimensions. The second is **transferability**, which makes a disentangled latent space more suitable for transfer learning since all features are separated along the dimensions. Another benefit is that human experts can assign their domain-specific meaning to individual dimensions, which is called **interpretability**. The last is **conditioning and intervention**, which allows interpretable conditioning over a subset of the latent space.[56]

The creation of latent space as a whole and the creation of a disentangled

latent space are two different processes. When creating a disentangled latent space, several constraints need to be considered. Techniques for creating a disentangled latent space can be grouped into three categories: architectural approaches, data augmentation, and supervision with loss functions and regularizations.

4.6.1 Architectural Approache

Regarding the first technique, the architectural approach, we need to examine the individual parts of neural network layers. The structure and design of the neural network are modified, which enforces the creation of a structured latent space. This approach determines, through the architecture, how the latent space will be compressed, separated, and organized.[56, 60, 73]

Some types of these architectures are, for example, **autoencoders**, which create latent space directly in the bottleneck part of the architecture. Another type is **capsule networks**, which are less common. These networks group neurons into capsules, where each capsule serves as a cluster for a single feature [89]. The last type is the so-called **disentangled representation network**, which contains multiple branches or heads, each learning different features.

4.6.2 Data Augmentation and Supervision

Another technique, data augmentation and supervision, is also used. It focuses on two different topics.

The first one is controlled augmentations. Controlled augmentation focuses on applying specific sets of transformations, such as rotation, color normalization, or scaling. The goal of this approach is to force the model to learn these specific features from the augmentations and to connect the original image with its augmented sample. This creates a latent space that understands that a single object

can have different features while still corresponding to the same class.[60, 11]

The second aspect is contrastive learning [24]. The only requirement for this technique is to have two types of samples: positive and negative. These samples force the model to separate the latent space so that it can distinguish between positive and negative samples and determine which cluster a newly embedded input is closer to. This is useful in situations where we can explicitly define what positive and negative samples look like.

4.6.3 Loss Functions and Regularization

At first sight, the loss function and regularization may seem like separate concepts, but in the context of disentanglement, this is not entirely the case. In the process of creating a disentangled latent space, these two components complement each other. The core idea is that the loss function is divided into a main loss function and regularizations. The role of the main loss function is to define the primary objective of model training, allowing it to solve the overall learning task, such as reconstruction or classification. In addition to the main loss function, regularizations act as auxiliary components. They add constraints or preferences to the model, such as desired shapes and properties of the expected latent space.[11, 36]

β -VAE Loss is a modification of the standard variational autoencoder loss, which is designed to create a disentangled latent space. The only one addition is the hyperparameter β which controls the trade-off between reconstruction accuracy and disentanglement, by increasing β also increases disentanglement. Looking at Eq. 4.2, the first part $\mathcal{L}_{\text{BETA}}(\phi, \beta) = -E_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z})$ is the reconstruction loss, controlling how well the output is reconstructed. The second part $\beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \parallel p_{\theta}(\mathbf{z}))$ represents how close the approximate posterior is to the

prior, as measured by the KL divergence.[114]

$$\mathcal{L}_{\text{BETA}}(\phi, \beta) = -E_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) + \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \parallel p_{\theta}(\mathbf{z})) \quad (4.2)$$

FactorVAE Loss is an extension of the base variational autoencoder, with the main difference being that it directly penalizes the total correlation of the latent space. This means it accounts for the dependency between latent features in the latent space. The primary goal of this loss function is to achieve a disentangled latent space with independent latent features. In Eq. 4.3, $\text{TC}(z)$ represents the penalization of total correlation, which can be weighted by the hyperparameter γ . [51]

$$\mathcal{L}_{\text{FactorVAE}} = E_{q(z|x)}[\log p(x|z)] - D_{\text{KL}}(q(z|x) \parallel p(z)) - \gamma \cdot \text{TC}(z) \quad (4.3)$$

Contrastive loss is a simple loss function that pulls similar input latent representations closer to each other, and different representations are pushed away. Because of this, the core part of the loss function is the calculation of distance. In Eq. 4.4, distance is represented by D . The parameter y is also important, as it acts as a switch between pulling embeddings together and pushing them apart.[30]

$$\mathcal{L} = y \cdot D^2 + (1 - y) \cdot \max(0, m - D)^2 \quad (4.4)$$

Triplet Loss is a function similar to contrastive loss, where the function is calculated with samples. But in this loss function, we don't have just two samples but three: anchor, positive, and negative. The goal of this function is to make the distance between the anchor and the positive sample smaller and anchor and the

negative sample bigger. In Eq. 4.5 can be seen two parts, first $\|f(x_a) - f(x_p)\|_2^2$ representing the distance between anchor and positive. The second part $\|f(x_a) - f(x_n)\|_2^2$ representing the distance between negative and anchor. Moreover, in the equation is parameter α representing the least wanted margin.[36]

$$\mathcal{L} = \max(\|f(x_a) - f(x_p)\|_2^2 - \|f(x_a) - f(x_n)\|_2^2 + \alpha, 0) \quad (4.5)$$

Magnet Loss has a similar concept as a triplet or contrastive loss. However, it uses a bit smarter approach. It is designed to learn better class-wise separation and intra-class compactness. The advantages of this method are that it uses clustering to model class distribution, push samples closer to the cluster centroid, and separate them from other centroids. Looking at Eq. 4.6, it contains $-\frac{1}{2\sigma^2}\|f(x) - \mu_{c(x)}\|^2$, which produces normalized distance between embedded input and cluster centroid. The numerator represents the distance between the input to its corresponding cluster, and the denominator represents the sum of distances between all clusters.[85]

$$\mathcal{L}(x) = -\log\left(\frac{\exp\left(-\frac{1}{2\sigma^2}\|f(x) - \mu_{c(x)}\|^2\right)}{\sum_{j \in \mathcal{N}(x)} \exp\left(-\frac{1}{2\sigma^2}\|f(x) - \mu_j\|^2\right)}\right) \quad (4.6)$$

Orthogonal Jacobian Regularization is a regularization method that shape the Jacobian matrix to be orthogonal, which helps to avoid entangled and redundant features. Another advantage of this regularization is that it helps avoid mode collapse and preserves the manifold structure. Looking at Eq. 4.7, J means Jacobian matrix, I represents identity matrix.[113]

$$\mathcal{L}_{\text{OJR}} = \|J^\top J - I\|_F^2 \quad (4.7)$$

4.6.4 Latent Space Interpretability Metrics

Creating a disentangled latent space is important, but without proper metrics to measure whether disentanglement was successful, we cannot draw any conclusions. For this reason, a large number of different metrics have been developed. These metrics can be divided into two main categories based on the data they require. The first category requires labels, including metrics such as mutual information gap and disentanglement, completeness, and informativeness. The second category does not require labeled data, with metrics such as the Silhouette score and the Davies-Bouldin index.[19, 62]

Mutual Information Gap also called MIG is a strong indicator representing how much each latent dimension captures factors of variation, like shape, texture, and colour, while penalizing redundancy. It measures mutual information between true factors and latent features. Another thing, that is also measured, is how well the model aligns latent dimensions with generative factors. In the Eq. 4.8 three core parts can be seen. The first part $I(\mathbf{v}_{j_k}, \mathbf{z}_k) - \max_{j \neq j_k} I(\mathbf{v}_j, \mathbf{z}_k)$ represents mutual information calculation between first and the second most informative latent feature. In this part, v_{j_k} represents positive match and the v_j represents candidate negative samples. The second part $\frac{1}{H_{Z_k}}$ is responsible for normalization and the last $\frac{1}{K} \sum_{k=1}^K$ perform averaging over all factors.[62]

$$\frac{1}{K} \sum_{k=1}^K \frac{1}{H_{Z_k}} \left(I(\mathbf{v}_{j_k}, \mathbf{z}_k) - \max_{j \neq j_k} I(\mathbf{v}_j, \mathbf{z}_k) \right) \quad (4.8)$$

Disentanglement, Completeness, Informativeness in a short DCI is a set of three complementary scores, which provide an understanding of how well the representation captures the ground true generative factors. Disentanglement, which can be seen in Eq. 4.9, measure if each latent dimension influences only one latent

factor. In this equation are three important parameters R_{ik} representing importance score, H_j representing entropy and $\log K$ representing maximum possible entropy. Completeness in Eq. 4.10 measures if each factor is explained by one latent dimension. In this equation is important parameter $Q_j^{(k)}$, which represents the importance score of the latent variable j for predicting factor k . And the informativeness, shown in Eq. 4.11, represents how well latents predict all factors. The $\text{score}(v_k, \hat{v}_k)$ part in an informativeness metric represents the computation of model performance.[33]

$$\text{Disentanglement} = \sum_{j=1}^d \left(\frac{1}{K} \sum_{k=1}^K R_{ik} \right) \cdot \left(1 - \frac{H_j}{\log K} \right) \quad (4.9)$$

$$\text{Completeness} = \frac{1}{K} \sum_{k=1}^K \left(1 + \frac{\sum_j Q_j^{(k)} \log Q_j^{(k)}}{\log d} \right) \quad (4.10)$$

$$\text{Informativeness} = \frac{1}{K} \sum_{k=1}^K \text{score}(v_k, \hat{v}_k) \quad (4.11)$$

Silhouette score is the metric that is most commonly connected to cluster quality. In the topic of latent space, similar factors form clusters. This metric tells how well each point fits into its own cluster in comparison to other clusters. In Eq. 4.11 can be seen two main parts $b(i)$, which represents minimum average distance between point i and all points in any other cluster. The second part is $a(i)$, which represents the distance between point i and all other points in the same cluster.[87]

$$S = \frac{1}{N} \sum_{i=1}^N \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4.12)$$

Davies–Bouldin Index is the last mentioned metric, which measures the quality of the cluster. It is possible to detect intra-cluster similarity, which represents how compact each cluster is. And inter-cluster difference, which represents how well separated the clusters are. Looking at the Eq. 4.13 can be seen that three distances are calculated. The first distance $\Delta(C_i)$ is the average distance of all points in cluster i from its centroid. The second distance $\Delta(C_j)$ is the average distance to the other cluster. And the $\delta(C_i, C_j)$ represents the distance between the centroids of cluster i and cluster j . [78]

$$DB(C) = \frac{1}{K} \sum_{i=1}^K \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\} \quad (4.13)$$

4.6.5 Challenges of Disentanglement

The most important aspect of working with latent space is understanding its problems and challenges. By being aware of them, we can aim to address these issues or at least account for them. Some of these challenges include identifiability and impossibility, the role of inductive biases, interpretability and utility, and, lastly, assumptions about factors. [60, 61]

Identifiability and Impossibility is the first challenge, highlighting a fundamental limitation of disentanglement. According to this problem, it is not possible to create a high-quality disentangled latent space using only unsupervised training. For the same data, an infinite number of disentangled spaces exist, making it difficult to determine which one is the best and most valuable. This challenge indicates that relying on random chance is insufficient, and some form of inductive bias is essential. This can be introduced through model architecture, changes in the training process, or modifications to the data. A few studies have achieved reasonable results with unsupervised learning, but very specific model classes are

required.

Interpretability and Utility encompass the concept of disentanglement. This concept suggests that humans can distinguish between different features, which in turn implies improved generalization. However, results show that no clear reduction in complexity is achieved. Different disentanglement metrics often disagree on which model is best for creating a disentangled latent space. Because of this, it is difficult to determine what disentanglement truly means, whether it generalizes better, and whether the model is more interpretable in certain domains.

Assumptions on Factors relates to the core concept of disentanglement, which is the separation of individual features. The problem is that in real-life scenarios, features are often correlated. In situations where features are correlated and the model attempts to disentangle them, disentanglement may not be applied correctly.

Chapter 5

Generative Neural Networks

In recent years, generative neural networks have become more and more popular in the field of image generation. They are used in multiple areas such as multimedia, art, autonomous driving, and medical imaging. In the medical imaging field, these networks are used for tasks such as synthetic data generation, image reconstruction, inpainting of missing or corrupted regions, and modality translation.[88, 88]

The main diagram representing different types of generative neural network architectures can be seen in Fig. 5.1. From the diagram, it can be observed that each architecture offers two advantages and one disadvantage.[121]

Generative adversarial networks have fast and high-quality sampling, but in the field of histological images, where these images contain complex features and fine differences, mode diversity becomes crucial. The main architectural feature of GAN is that it contains two parts: a discriminator and a generator. From the perspective of architecture training, the generator focuses on image generation, while the discriminator decides whether the generated image looks real or fake.

This means that the main goal of the architecture is to train the generator so effectively that it is able to fool the discriminator into classifying the generated image as a real one.[48, 6]

In this work, we will focus primarily on variational autoencoders, because of their ability to sample quickly while preserving mode diversity, and on denoising diffusion models, which also preserve mode diversity but provide high-quality sampling. These architectures will be further explained in the following sections.

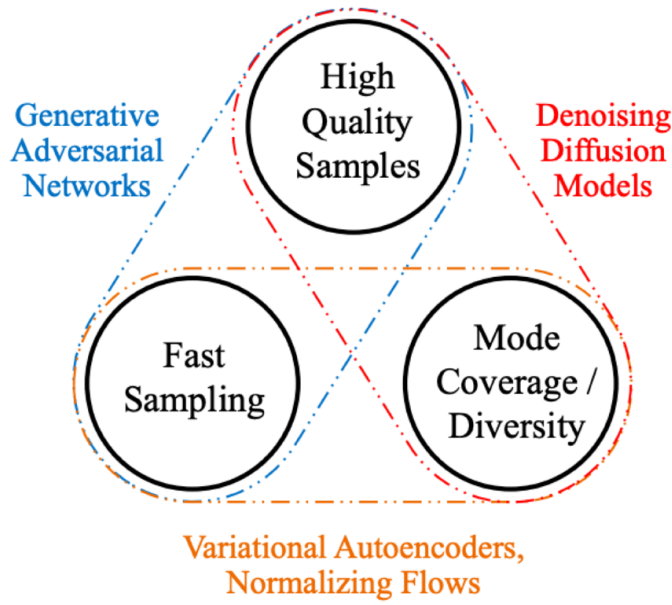


Figure 5.1: Generative neural networks Trilemma [37]

5.1 Vector Quantized Variational Autoencoder

First, we will focus on the autoencoder architecture, as it serves a complementary role to our main objective, which is synthetic data generation with denoising diffusion models. The core principles of autoencoders were explained in Sec. 4.1.1. In this section, we will focus on the VQ-VAE architecture and its ability to perform the **image reconstruction** task, whose goal is to reconstruct the original image as

accurately as possible. Autoencoders trained in this manner are commonly used in denoising diffusion models to encode images into latent space and to decode noisy latent representations back into images.

VQ-VAE is a generative model that works in a similar way to an ordinary autoencoder. The only modification is the addition of a learned codebook, which can be seen in Fig. 5.2. This codebook is used for discretization of the latent space, which gives the model three advantages.[107]

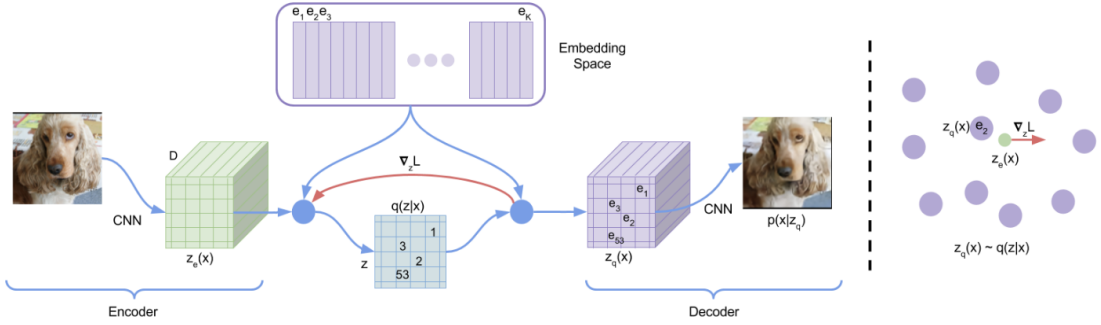


Figure 5.2: VQ-VAE architecture diagram[107]

The first advantage is that the model creates a finite set of learnable embeddings, enabling quantization of continuous encoder outputs to their nearest neighbors. This makes reconstruction less blurred. The second advantage is training stability, achieved by avoiding posterior collapse, since variance optimization in the KL divergence is eliminated due to deterministic quantization. The last advantage is that tokens represent semantic prototypes. Codebook learning with exponential moving average (EMA) handles updates dynamically and supports semi-supervised techniques.[107]

5.2 Denoising Diffusion Models

Denoising diffusion models are one part of the mentioned generative model trilemma. These models focus on high-quality sampling and mode diversity. The core concept of sampling with DDMs is the noising and denoising process, inspired by non-equilibrium thermodynamics. Another important concept is the Markovian chain, which is parameterized and used for the forward diffusion process. This process progressively adds Gaussian noise to the data until pure noise is obtained. The forward and reverse processes can be seen in Fig. 5.3.[43, 109, 28]

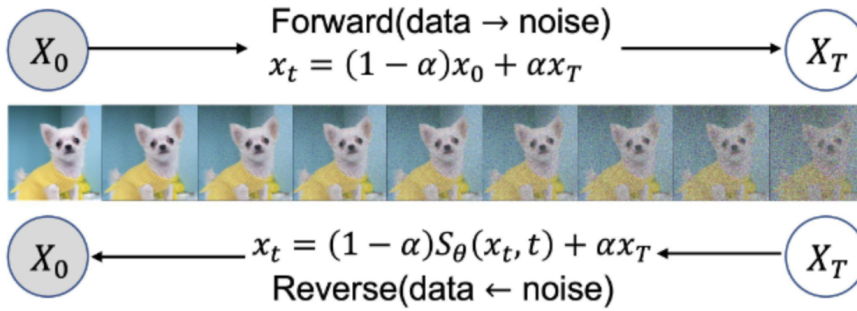


Figure 5.3: Forward and reverse process of DDM[117]

Forward process is a fixed Markov chain, which adds small Gaussian noise to data. This process is represented by Eq. 5.1.[43]

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (5.1)$$

where:

- $\mathbf{x}_t$ is the image at timestep t in the diffusion process
- $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian (normal) distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

- β_t is the noise schedule parameter at timestep t , controlling the amount of noise added at each step

Reverse process is the complementary part to the forward process. It begins with noise $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and learns transitions to iteratively denoise the sample back into a data. Training involves optimizing a variational bound on the negative log-likelihood, often simplified to predicting the noise added during the forward diffusion. The formulation of the reverse process is shown in Eq. 5.2.[43]

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, t)) \quad (5.2)$$

where:

- $\mathbf{x}_t$ is the noisy image at timestep t
- θ denotes the learnable parameters of the neural network
- $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian (normal) distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Markovian chain plays a role in both the forward and reverse processes. Markovian chain represent that the transition depends only on the current state. This process help to simplify the parametrization of Gaussian conditionals when noise is small, ensure trackable posteriors, allow closed form KL divergences in ELBO. The reverse chain learnable parameters capture data distribution through sequential one-step predictions.[43]

The key applications of DDMs include conditional generation of images and audio, synthetic generation of medical images, drug discovery via molecular simulation, and the creation of digital art.[109, 28]

5.2.1 Denoising Diffusion Probabilistic Models (DDPM) [43]

Denoising diffusion probabilistic models (DDPM), in comparison to ordinary DDMs focus on noise scheduling, model parametrization, training and sampling.

In DDPMs, the noise level is fixed using a specific **linear or cosine scheduler**. This allows the model to sample any noisy timestep from clean data using a single formula. As a result, there is no need to simulate step-by-step noising during training, which makes the process much faster.

Another modification in DDPMs is that they do not use mean prediction to estimate noise; instead, they employ a **UNET** architecture. This makes training simpler and improves the accuracy of noise prediction.

DDPMs use single efficient **loss focused only on denoising quality** across all noise levels. This is achieved by ignoring of constant terms in the variational bound. This equation can be seen in Eq. 5.3.

$$L_{\text{simple}} = E [\|\epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2] \quad (5.3)$$

where:

- $E[\cdot]$ denotes the expectation over the data distribution and noise
- ϵ is the true noise that was added during the forward process
- $\epsilon_{\theta}(\cdot, t)$ is the neural network that predicts the noise, parameterized by θ
- $\mathbf{x}_0$ is the original clean image
- $\bar{\alpha}_t$ is the cumulative product of $(1 - \beta_t)$ values up to timestep t , defined as $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$
- t is the timestep in the diffusion process

This model also provides a **concrete sampling procedure**, which first generates pure noise, then computes the predicted mean for each of the 1000 backward steps. A small amount of noise is added at each step to simulate the true posterior. This process helps generate diverse and high-fidelity samples.

5.2.2 Denoising Diffusion Implicit Models (DDIM) [96]

Further modifications of the DDPM led to the creation of DDIM. This approach enables deterministic and faster sampling while preserving quality.

The first modification in this process is the **non-Markovian posterior**, which allows skipping steps. The number of steps used can be flexible, which can speed up sampling but may compromise quality. Another feature of DDIM is that it has a **deterministic option** and can be executed either as DDIM or as DDPM.

The equation used in the DDIM can be seen in Eq. 5.4. This equation allow process to be fully deterministic or stochastic.

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \sigma_t \mathbf{z} \quad (5.4)$$

where:

- $\mathbf{x}_t$ is the noisy image at timestep t
- $\bar{\alpha}_t$ and $\bar{\alpha}_{t-1}$ are the cumulative products of $(1 - \beta_s)$ up to timesteps t and $t - 1$ respectively
- $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is the predicted noise by the neural network at timestep t
- $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is random noise sampled from a standard normal distribution

- The third term $\sigma_t \mathbf{z}$ adds stochastic noise, allowing for different sampling trajectories
- When $\sigma_t = 0$, the process becomes fully deterministic (DDIM), and when $\sigma_t > 0$, it includes stochasticity

5.2.3 Pseudo Linear Multistep Methods (PLMS) [59]

Pseudo linear multistep method was build on top of DDIM. The main modification of this architecture is its ability to use multiple previous noise predictions to better estimate current noise prediction. These prevoius noise predictions are combined in linear sum and produce higher order estimation of the noise. The equation of the k-th order calculation can be seen in Eq. 5.5. PLMS provide much smoother and accurate trajectory at lower step counts. It also provide sharper edges, better global structure and also fewer "burn" artifacts.

$$\epsilon_t^{(\text{PLMS})} = \sum_{i=0}^{k-1} w_i \epsilon_{t-i} \quad (5.5)$$

where:

- ϵ_{t-i} are the previous noise predictions
- w_i are fixed coefficients derived from the Adams–Bashforth method, which determine the weight of each previous prediction

5.2.4 Inpainting with Diffusion Models

Image inpainting, also called image completion, is a task that focuses on filling in missing regions of an image. From a computational perspective, this task can also be seen as pixel-conditioned generation. The goal is to create an image that looks

natural and is free of any synthetic artifacts. Inpainting can be applied to various tasks, such as restoring damaged photos, removing unwanted objects, or generating content behind masked parts of an image. Historically, GANs were mostly used for this task, but this architecture had multiple problems.[52, 64]

Problems introduced by the use of GANs include the fact that models were trained on specific mask types, which limited their ability to generalize. Additionally, these models often generated simple, incorrect, and blurred regions rather than semantically meaningful content.[64]

RePaint method was developed to address these problems. It proposes using a pretrained unconditioned DDPM-based architecture as a prior, which applies reverse diffusion on unmasked pixels, allowing the model to fill in unknown regions. The key idea is to alter the sampling routine so that known image pixels are re-injected at each reverse diffusion iteration. This enables the model to achieve high-quality and diverse inpainted images. Another modification is the addition of a special resampling mechanism, which introduces extra forward and backward diffusion steps to refine consistency and improve alignment. An example of inpainting with multiple sampling steps can be seen in Fig. 5.4.[64]

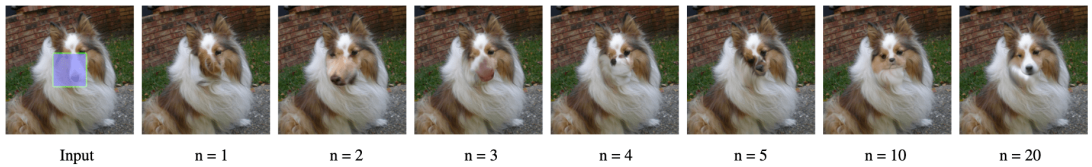


Figure 5.4: Image examples of using repaint with multiple different sampling steps[64]

5.2.5 Semantic Mask-Conditioned Generation

Semantic mask-conditioned generation is a method for synthetic image generation. In this method, the model generates an image conditioned on a semantic layout, allowing us to specify which structures should appear in which regions. Unlike inpainting, it does not require matching any original image pixel by pixel. Conditioned generation is ideal in situations where control over spatial structures is needed and no image reference exists. This method is mostly used in tasks such as scene synthesis and medical imaging, where images need to be generated based on annotations or for data augmentation. An example of semantic mask-conditioned generation can be seen in Fig. 5.4.[46, 110]

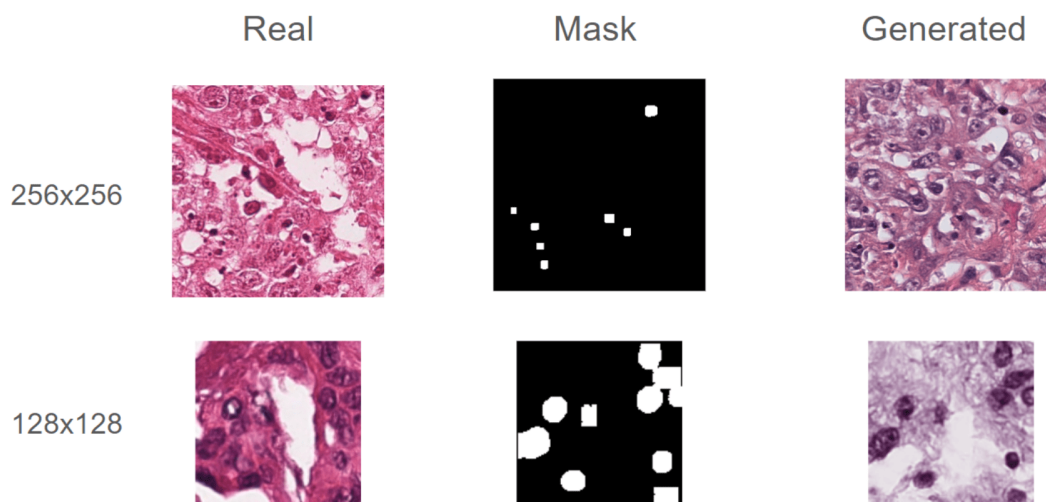


Figure 5.5: Example of conditioned histology image generation[14]

Models used for conditioned generation include GANs and DDMs. Nowadays, DDMs are more commonly used because of their ability to generate globally coherent images and more diverse outputs. DDMs can employ multiple conditioning mechanisms, with the three most common being: spatial conditioning, where the image is pixel-wise aligned with the semantic mask; semantic embeddings, where

each class is mapped to learned embeddings; and soft/hard conditioning, where the model is guided (soft) or forced (hard) to generate the conditioned image.[46, 110]

Chapter 6

Explainability

The main question is why deep learning requires explainability. The answer is straightforward: in some cases, doctors remain concerned to use AI due to its limited explainability and concerns regarding generalization. Deep neural networks are often perceived as black-box systems, which decreases trust in their decisions. Nevertheless, clinicians are interested in using these tools to speed up diagnostic workflows or highlight abnormalities that might be difficult to detect visually. Another fundamental issue is that patients have the right to understand how decisions affecting their health are made, particularly when they rely on AI-driven predictions. Combined, these factors make explainability in medical AI indispensable, reinforcing the need for continual development of methods that improve transparency and interpretability.[20, 90, 103]

When a model fails in a critical setting, providing a clear explanation is often necessary. However, terminology used to classify explainability methods is not fully standardized, which leads to inconsistent naming across the literature. Despite these differences, explainability techniques can generally be divided into two groups based on whether they are applied before or after model training. Post-hoc

methods are applied retroactively, meaning they attempt to explain decisions after a model has been trained. In contrast, ante-hoc methods focus on interpretability by design. Models are built inherently transparent, making their decision process understandable from the start. A more detailed overview of explainable AI techniques is illustrated in Fig. 6.1. [20, 65, 44]

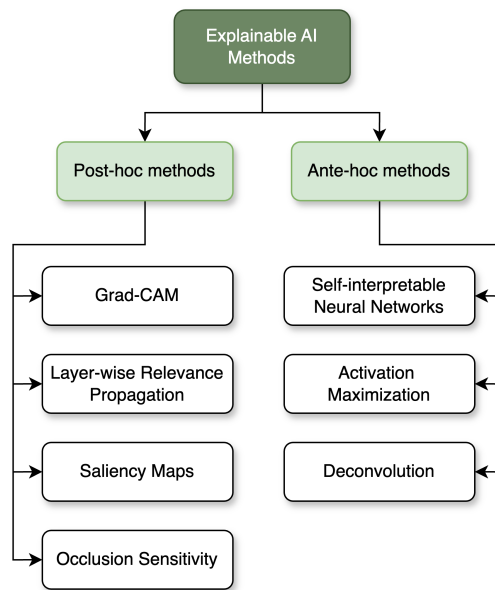


Figure 6.1: Diagram of explainable AI methods

6.1 Post-hoc Vizualization Methods

The post-hoc method can be integrated more easily than ante-hoc. It provides an explanation after the model is trained, which creates a big advantage since the model does not need to be adjusted directly. These methods are model-agnostic, which makes them easy to implement on any model. Post-hoc methods are often more adaptable, and their impact can therefore be much greater if they are able to provide accurate explanations.[65]

On the other hand, at first sight, the problem with post-hoc can be overlooked. A vast number of post-hoc methods exist, but each of these methods provides different accuracy of relevancy maps. In a situation like this, a problem with the reliability of these methods arises. This problem needs to be accounted for when working with these methods. Some of the well-known post-hoc methods are, for example, Grad-CAM or LRP.[65, 112]

6.1.1 Grad-CAM

The first post-hoc method is Gradient-weighted Class Activation Mapping, known as Grad-CAM. It is an explanation technique designed for CNN-based neural networks. Grad-CAM visualizes and highlights the image regions that contribute most to the final prediction. In practice, the output is a heatmap that is most often overlaid on the original image, as shown in Fig. 6.2.[92]



A man is sitting at a table with a pizza

Figure 6.2: Visualization of Grad-CAM map [92]

a mathematical point of view, the process of heatmap generation can be divided into two main parts. The first part is the extraction of importance weights, which

can be seen in Eq. 6.1. We first calculate the gradient y^c with respect to the feature map A of a convolution layer, on which global average pooling is applied. This calculation obtains weights α_k^c . The second step is to calculate the final Grad-CAM heatmap, whose equation can be seen in Eq. 6.2. This calculation represents a simple weighted combination of feature maps, followed by the ReLU activation function.[92]

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (6.1)$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (6.2)$$

The advantages of using Grad-CAM are that it increase explainability of the model's decisions, it helps with potential debugging of model performance by visualizing the regions that contributed most to the prediction, and improve model trustworthiness.

6.1.2 Layer-wise Relevance Propagation

The second post-hoc technique is Layer-wise Relevance Propagation, or LRP. This technique is similar to Grad-CAM, but the core difference is that it focuses on visualizing fine-grained input attributes, as shown in Fig. 6.3. It operates by propagating the prediction backward through the neural network. This process is controlled by a set of carefully designed propagation rules.[71]

Looking at LRP, from the mathematical point of view. For providing of explanation of the equation, which can be seen in Eq. 6.3, we need to focus on j and k . The j represents layer l and the k represents the layer $l+1$. Moving to next

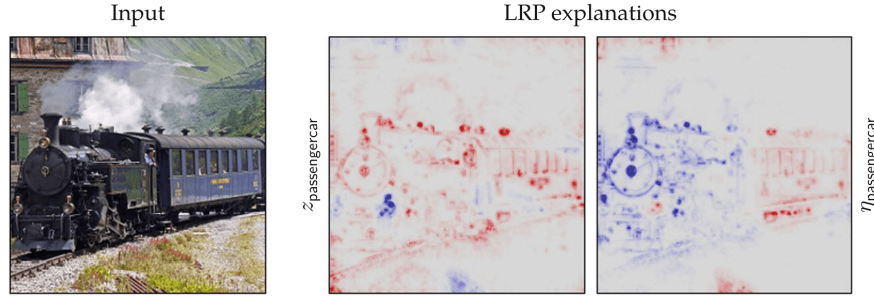


Figure 6.3: Visualization of LRP map [71]

part, z_{jk} represent contribution of neuron j to neuron k . The last important part R_k represents the propagation relevance score of the higher layer. By calculation of the provided equation can be obtained R_j , which represents the relevance score for layer j , which can be visualized as a relevance heatmap for the corresponding layer.[71]

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k. \quad (6.3)$$

The advantages of this method are similar to Grad-CAM, but there are a few differences between the these methods. LRP makes it possible to visualize more fine-grained attributes, which can be advantageous in the field of histology, where images contain many fine details. It also works with higher resolutions and is not limited to CNN architectures, but can be applied to any deep neural network architecture. Overall, LRP may be better in certain cases than Grad-CAM, but both methods are used for different suitable use cases.

6.2 Ante-hoc Methods

Ante-hoc methods, or intrinsic methods, are approaches that focus on changes inside the model architecture, rather than only processing the output of the model.

This makes these models transparent and interpretable from the beginning, meaning they are internally understandable to humans without any additional explanation tool.[65, 84]

It is also worth mentioning the problems with this method. From the deep neural network perspective, the integration of ante-hoc methods can cause a decrease in model accuracy, creating a trade-off between accuracy and explainability. Additionally, applying these methods makes models more robust and complex, which can lead to performance issues. The last problem is that, even though this method aims to improve explainability, it can still be complex, and cognitive overload for users may remain high. Examples of ante-hoc methods are self-explaining neural networks and prototype networks.[84]

Chapter 7

State of the art

The current state of research in the field of latent space, feature vectors, encoders and generative networks is rapidly advancing. Many papers focus on these topics. Encoders such as variational autoencoders, vector quantization autoencoders, and beta-variational autoencoders are showing exceptional results in creating representative and disentangled latent spaces. And on the other sides denoising diffusion models show exceptional results in synthetic image generation.

A few architectures, such as Bag of Queries, aim to integrate the most important features directly into the training process. This idea is powerful and has great potential. However, the problem with these approaches is that the features are not sufficiently explainable and cannot be easily understood. The integration of fully understandable features into the architecture and forcing the model to focus on them is not yet well-researched. That is why we do not know which architecture and approach will be best for our proposed solution, but we can draw insights from papers on similar topics.

7.1 BoQ: A Place is Worth a Bag of Learnable Queries [7]

This work proposes an architecture called Bag of Queries. Although the architecture focuses on place recognition tasks and accurately matching locations, it can also be beneficial in other fields. Using this architecture, universal place-specific attributes can be captured in the form of learnable queries. These input queries are directly integrated into the architecture via cross-attention. In addition, a special attention block is implemented, which provides explainability within the architecture and can generate saliency maps. One of the advantages of this architecture is the possibility of using different backbones, such as CNNs or Vision Transformers.

The whole architecture, shown in Fig. 7.1, is divided into three main parts: backbone, encoder, and learnable queries with cross-attention. The backbone is responsible for extracting local features. Afterwards, a cascade of encoder blocks performs sequential refinement of the output from the backbone. In the final step, learnable queries perform self-attention to integrate shared information into the training process.

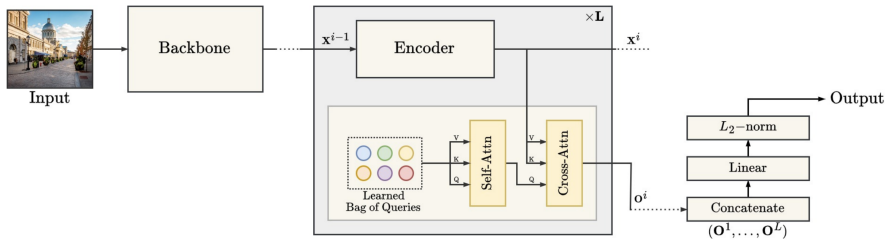


Figure 7.1: Architecture of Bag of Queries

In this work, multiple datasets were used, each containing various formats of street views from different angles. They include different environments, lighting

conditions, weather, perspectives, and even some historical grayscale images. Some of the most well-known datasets used are MapillarySLS (MSLS), Pitts250k, and AmsterTime. Overall, the combination of all datasets contained over 300 000 images in different resolutions.

The training process mentioned in this work was performed only with the GSV-Cities dataset. The Multi-Similarity loss was used as the loss function, and AdamW was used as the optimizer. The hyperparameter setup included a batch size of 4, a learning rate of 0.0002–0.0004, and 40 epochs. All images were rescaled to 320x320 pixels, which was also innovative, as all state-of-the-art works were trained with higher resolutions.

Multiple experiments, in which the main metric was recall, were performed. The experiments can be divided into three main parts. The first tested the use of the self-attention block, which showed that the model’s recall clearly increased. The second experiment, varying the number of BoQ blocks, showed that it is ideal to use four BoQ blocks. The last important experiment focused on the backbone architecture. Results show that the best backbone is the vision transformer-based architecture DINOv2. Overall, the results indicate that using BoQ achieves the best recall compared to state-of-the-art architectures, as can also be seen in Fig. 7.2.

Method	Dim.	Pitts250k-test			MSLS-val			SPED			Nordland*		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
AVG [6]	2048	78.3	89.8	92.6	73.5	83.9	85.8	58.8	77.3	82.7	15.3	27.4	33.9
GeM [41]	2048	82.9	92.1	94.3	76.5	85.7	88.2	64.6	79.4	83.5	20.8	33.3	40.0
NetVLAD [6]	32768	90.5	96.2	97.4	82.6	89.6	92.0	78.7	88.3	91.4	32.6	47.1	53.3
SPE-NetVLAD [55]	163840	89.2	95.3	97.0	78.2	86.8	88.8	73.1	85.5	88.7	25.5	40.1	46.1
Gated NetVLAD [58]	32768	89.7	95.9	97.1	82.0	88.9	91.4	75.6	87.1	90.8	34.4	50.4	57.7
Conv-AP [3]	4096	92.4	97.4	98.4	83.4	90.5	92.3	80.1	90.3	93.6	38.2	54.8	61.2
CosPlace [11]	2048	92.3	97.4	98.4	87.4	93.8	94.9	75.3	85.9	88.6	54.4	69.8	75.9
MixVPR [4]	4096	94.2	98.2	98.9	88.0	92.7	94.6	85.2	92.1	94.6	58.4	74.6	80.0
EigenPlaces [12]	2048	94.1	97.9	98.7	89.2	93.6	95.0	82.4	91.4	94.7	54.2	68.0	73.9
BoQ (Ours)	4096	95.0	98.4	99.1	91.1	94.8	95.7	85.4	93.1	95.4	69.5	83.4	87.0
BoQ (Ours)	16384	95.0	98.5	99.1	91.2	95.3	96.1	86.5	93.4	95.7	70.7	84.0	87.5

Figure 7.2: Results of Bag of Queries architecture

7.2 Weakly Supervised Vector Quantization for Whole Slide Images Classification [94]

The second work focuses on multiple instance learning. This is a weakly supervised learning approach, where data are structured into labeled bags, each containing multiple instances. In the domain of histology, this can be explained as follows: if a single tumor-positive instance is present on a slide, the entire slide is labeled as positive. This is especially important when processing histology images, which have very high resolutions and require vast computational resources.

When processing histological images, it is important to compress smaller tiles into embeddings and process them. The problem with current pre-trained vision encoders is that the generated embeddings are not task-specific and contain significant noise, which results in poor discrimination between tumor and healthy tiles. One of the technologies that targets these problems is VQ-MIL.

The core concept of the VQ-MIL architecture is derived from the VQ-VAE architecture. The main difference is the replacement of the reconstruction loss with a weakly supervised slide-level classification loss. The second difference is the replacement of the decoder with the MIL model. From an architectural point of view, VQ-MIL can be divided into three parts, as shown in Fig. 7.3. The first part is the vision encoding model, which divides the WSI into tiles and converts them into embedded vectors. The second part is the vector quantization module, which encodes embeddings into the nearest codewords of the vector quantization codebook. These encoded codewords are then passed to the final part, the MIL model, which aggregates all codewords and makes the prediction.

Two different datasets were used in this work. One was Camelyon16, which contained 400 whole-slide images, divided into 270 training slides and 130 test

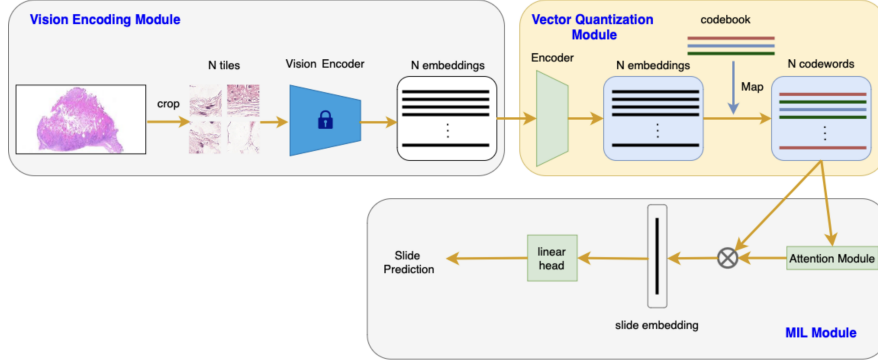


Figure 7.3: VQ-MIL architecture

slides with pixel-level annotations. The dataset includes slides with breast cancer metastases in lymph node tissue. The second dataset, TCGA, includes WSIs for two major lung cancer subtypes. It consists of 1 054 slides, all at 20x magnification.

The training process was performed on cropped tiles of size 224x224 pixels. A ResNet50 was used as the encoder, and AdamW served as the optimizer with a learning rate of 0.0001. The batch size was 1, due to the vast amount of details in the histological images. Overall, the model was trained for 400 epochs on Camelyon16 and 100 epochs on the TCGA lung cancer dataset. The metrics used in the experiments were accuracy and AUC.

Ablation studies performed with VQ-MIL show that the optimal codebook size is 32 and the best pseudo-bag size is 256. Overall, results of this method compared to state-of-the-art architectures like DTFD-MIL, TransMIL, or CLAM show that VQ-MIL performs best and is superior in both accuracy and AUC. A comparison of the results can be seen in Fig. 7.4.

Model	Accuracy	AUC
ABMIL	0.8474 (± 0.021)	0.9111 (± 0.021)
CLAM	0.8676 (± 0.036)	0.9489 (± 0.013)
TransMIL	0.8877 (± 0.057)	0.9618 (± 0.027)
DTFD-MIL	0.8809 (± 0.022)	0.9349 (± 0.012)
Ours	0.8884 (± 0.016)	0.9647 (± 0.010)

Figure 7.4: Results of experiments with VQ-MIL compared to state-of-the-art results

7.3 Enhancing Interpretability Through Loss-Defined Classification Objective in Structured Latent Spaces [36]

In this work, the proposed architecture is called Latent Boost. The architecture works with latent representations of the data. With this method, the authors aim to focus on explainability and transparency in a classification task. Latent representations are a promising solution to achieve this, as they contain rich information about the processed data. However, this approach faces several challenges that the work aims to address. The creation of latent space requires specific guidance to ensure that similar representations are clustered together. Another challenge is overall domain generalization, because latent representations can differ significantly between domains. The authors in this paper attempt to solve these problems.

The data used for experiments in this work consisted of three different datasets. One was Fashion MNIST, containing grayscale images of 10 categories at a resolution of 28x28 pixels. The other two datasets were CIFAR-10 and CIFAR-100, containing colored images of 10 and 100 classes.

From an architectural point of view, the core aspect is not a change in the

deep neural network architecture, but a change in the loss function. With this modification, the authors can integrate conditions on the latent space structure into the training process. The full architectural concept of the proposed solution is shown in Fig. 7.5.

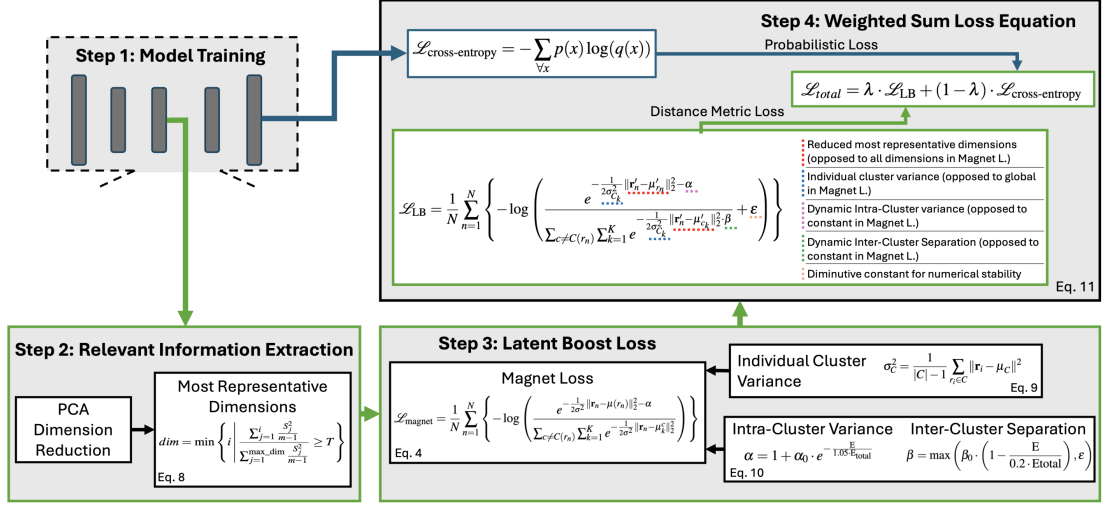


Figure 7.5: Flow of the latent boost approach

The first experiment shows a combined weighted loss function divided into two separate parts. These parts can be controlled with the hyperparameter λ , which balances the weight of each part.

Cross-entropy loss is the first and core part of the loss function. This loss is responsible for computing the weighted sum equation for multi-class classification.

Distance metrics loss is a combination of four different loss functions used as regularization for structured latent space construction. The first, **contrastive loss**, minimizes the distance between similar pairs of samples and maximizes the distance between different ones. The second, **triplet loss**, ensures that the

positive-to-anchor distance is smaller than the negative-to-anchor distance. The third, **n-pair loss**, is similar to triplet loss but uses multiple negative samples instead of one. The last, **magnet loss**, is responsible for optimizing data grouping into clusters. It penalizes samples based on the distance to correct and incorrect cluster centroids.

In the training process, the model architecture used for CIFAR-10 was VGG-16, and for CIFAR-100 it was ResNet-50. Both experiments used the same setup of 10 epochs, the Adam optimizer, and applied early stopping. The metrics used in these experiments were accuracy and F1-score.

After experiments with multiple loss functions, the final workflow combines the application of PCA for dimensionality reduction, calculation of individual cluster variation, and then calculation of magnet loss with cross-entropy loss. Results of this final method are shown in Fig. 7.6, where an increase in accuracy and a decrease in the number of epochs can be clearly seen. In addition to these metric evaluations, visualization of latent representations was also performed, which clearly shows a better-structured latent space.

Dataset	Metric	Baseline	Magnet	Latent Boost w/o PCA	Latent Boost	Improvement (%)
Fashion MNIST	Accuracy	88.59 $\pm$ 0.15	89.52 $\pm$ 0.34	88.12 $\pm$ 0.29	90.86 $\pm$ 0.21	2.56%
	Micro-F1	0.8867 $\pm$ 0.002	0.8946 $\pm$ 0.004	0.8971 $\pm$ 0.004	0.9038 $\pm$ 0.005	1.93%
	Nr. Epochs	40.67 $\pm$ 4.19	37.67 $\pm$ 6.18	34.86 $\pm$ 4.26	34.67 $\pm$ 3.09	-14.76%
CIFAR-10	Accuracy	85.88 $\pm$ 0.40	87.36 $\pm$ 0.82	88.12 $\pm$ 0.47	88.44 $\pm$ 0.28	2.98%
	Micro-F1	0.8586 $\pm$ 0.0042	0.8738 $\pm$ 0.0081	0.8811 $\pm$ 0.0040	0.8843 $\pm$ 0.0024	2.99%
	Nr. Epochs	76.33 $\pm$ 8.65	72.67 $\pm$ 1.70	71.94 $\pm$ 3.72	66.33 $\pm$ 4.50	-13.09%
CIFAR-100	Accuracy	61.77 $\pm$ 0.49	62.75 $\pm$ 0.51	63.05 $\pm$ 0.52	63.04 $\pm$ 0.48	2.06%
	Micro-F1	0.6163 $\pm$ 0.0064	0.6256 $\pm$ 0.0057	0.6218 $\pm$ 0.0059	0.6288 $\pm$ 0.0051	2.03%
	Nr. Epochs	86.67 $\pm$ 9.46	69.67 $\pm$ 8.73	69.23 $\pm$ 8.92	68.67 $\pm$ 6.02	-20.74%

Figure 7.6: Results of the latent boost architecture

7.4 Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning (HIPT) [22]

The core concept of this work is to process multiple resolutions and features of the whole slide image. It decompose processing into hierarchy of transformers operating at different spatial scales, what can be seen in Fig. 7.7. This approach results in preserving of global features and reduce computational demand. Resolution levels can be divided into three main parts. First are 16x16 patches capturing cellural features and micro-structural patterns. The second 256x256 patches capture cellural organizations and the last 4096x4096 capture whole tissue phenotypes. This sequential processing allow model to capture in top-level transformer fine details and know whole concept.

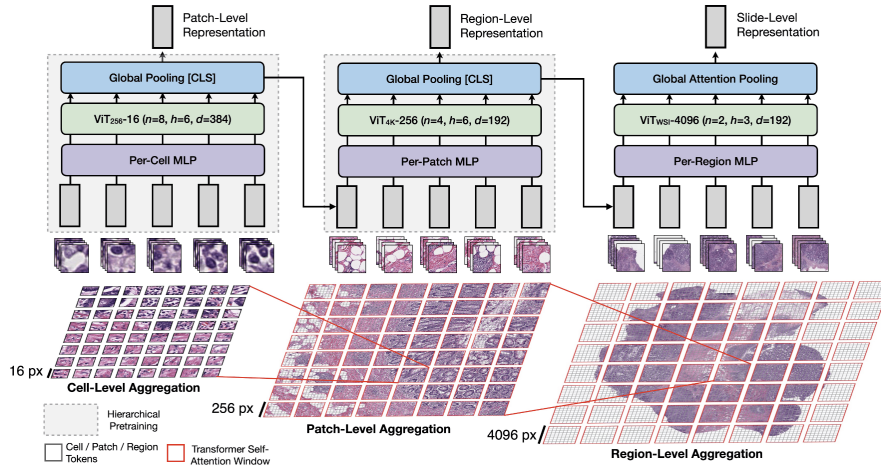


Figure 7.7: HIPT architecture

The dataset used in this work contains 10 678 weakly annotated WSIs of multiple cancer types. From these, around 400 000 patches of size 4096×4096 were extracted for transformer pre-training. Images were collected from multiple

medical centers, with variability in scanners and staining protocols. This dataset structure and its features are suitable for training a generalizable model.

Training of this architecture can be divided into multiple parts. The first part is pre-training of the patch-level transformer. This transformer is pre-trained in a self-supervised manner using contrastive learning on augmented views. This self-supervised training framework is called DINO. Training is performed at multiple scales to learn both fine details and global structures. After this training, the models are frozen and only the embedding aggregator is trained. The model was trained with the AdamW optimizer and survival cross-entropy loss. This hierarchical training strategy significantly reduces GPU memory usage.

The final output of the architecture is embeddings, which contain both global structures and fine details. These embeddings can then be used for multiple downstream tasks, such as survival prediction or identifying unique morphological phenotypes with heatmaps. These heatmaps can be seen in Fig. 7.8. Overall, the results and comparison of this architecture to other state-of-the-art solutions can be seen in Fig. 7.9. From these results, it is clear that the HIPT architecture achieves the best c-Index compared to other state-of-the-art models across multiple datasets.

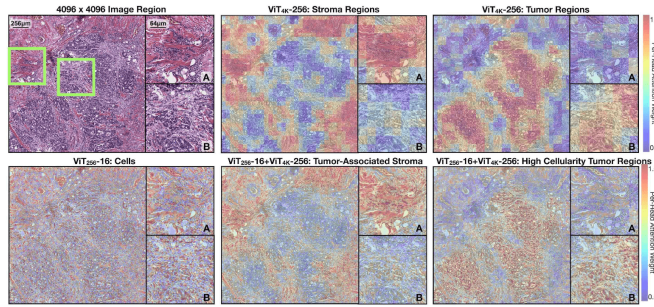


Figure 7.8: Heatmap results of HIPT architecture

Architecture	IDC	CRC	CCRCC	PRCC	LUAD	STAD
ABMIL [39]	0.487 ± 0.079	0.566 ± 0.075	0.561 ± 0.074	0.671 ± 0.076	0.584 ± 0.054	0.562 ± 0.049
DeepAttnMISL [80]	0.472 ± 0.023	0.561 ± 0.088	0.521 ± 0.084	0.472 ± 0.162	0.563 ± 0.037	0.563 ± 0.067
GCN-MIL [50, 86]	0.534 ± 0.060	0.538 ± 0.049	0.591 ± 0.093	0.636 ± 0.066	0.592 ± 0.070	0.513 ± 0.069
DS-MIL [46]	0.472 ± 0.020	0.470 ± 0.053	0.548 ± 0.057	0.654 ± 0.134	0.537 ± 0.061	0.546 ± 0.047
HIPT	0.634 ± 0.050	0.608 ± 0.088	0.642 ± 0.028	0.670 ± 0.065	0.538 ± 0.044	0.570 ± 0.081

Figure 7.9: HIPT architecture results and comparisons

7.5 ZoomLDM: Latent Diffusion Model for multi-scale image generation [118]

The last work we want to highlight is ZoomLDM. This architecture is a diffusion model used for large-image domain generation at multiple magnifications. It aims to solve problems such as failure to capture global structures and the wider context of images in the field of image generation. The architecture can be divided into two main parts, as shown in Fig. 7.10. The first part is SSL embedding extraction using a pretrained and frozen HIPT architecture. The second part uses these embeddings and summarizes them to project into the cross-magnification latent space. These embeddings are further used as a condition when generating patches from 256x256 to 4096x4096.

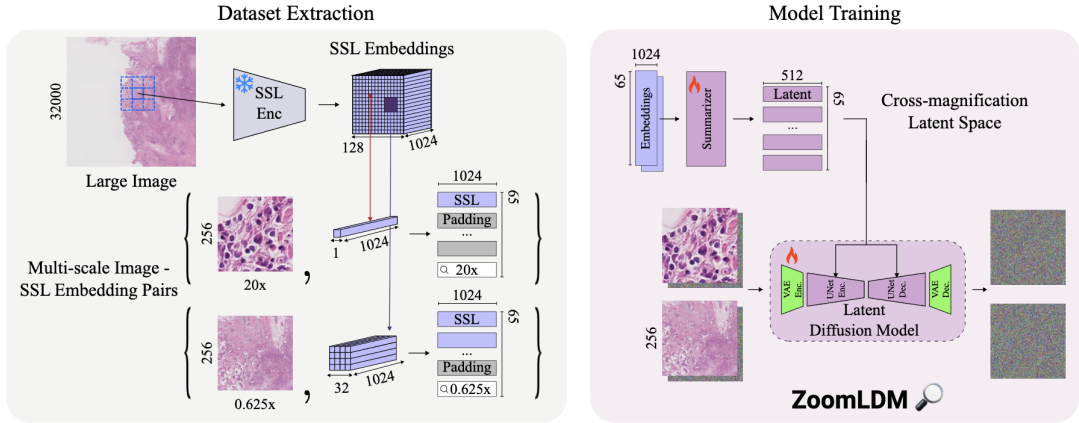


Figure 7.10: ZoomLDM architecture diagram

From the dataset point of view, the authors used a multi-scale histopathological dataset from TCGA-BRCA at eight different magnifications, ranging from 20x to 0.15625x. A total of 1136 whole-slide images were used to prepare the dataset by creating patches of multiple resolutions, such as 256x256 or 1024x1024. Core challenges with this dataset include data imbalance and the small number (2500) of low-magnification patches.

In the training process, a latent diffusion model with a UNET architecture was used, pretrained on ImageNet and conditioned by magnification and summarizer embeddings. The encoder used is VQ-f4, and the summarizer is implemented as a 12-layer vision transformer. The training process used a standard LDM-style noise schedule and the Adam optimizer with a 10^{-4} learning rate. The conditioned diffusion model was trained separately as a diffusion model. Weight sharing across all scales was also applied during training.

From the experimental point of view, ZoomLDM achieves state-of-the-art FID and visual quality across multiple magnifications, as shown in the synthetic samples in Fig. 7.11. It is clearly evident that the architecture can preserve both global and local tissue structures. Looking at the table shown in Fig. 7.12, it can be seen that the architecture was validated on two datasets, on which the model demonstrates state-of-the-art results.

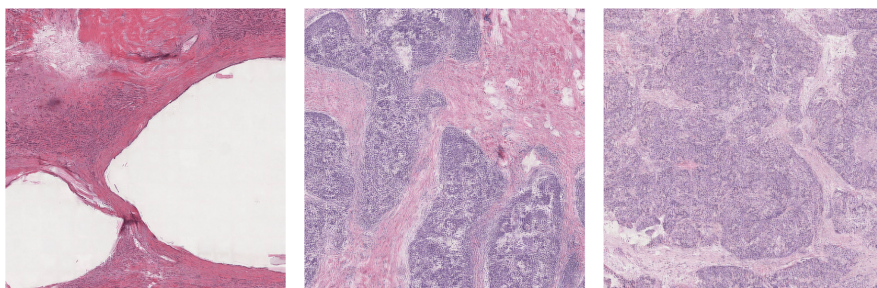


Figure 7.11: Synthetic samples generated with ZoomLDM

Method	Conditioning	TCGA BRCA					BACH				
		SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	CONCH $\uparrow$	UNI $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	CONCH $\uparrow$	UNI $\uparrow$
Bicubic	-	0.653	24.370	0.486	0.871	0.524	0.895	34.690	<u>0.180</u>	0.969	0.810
CompVis [38]	LR image	0.563	21.926	0.247	0.946	0.565	0.723	27.278	0.206	0.954	0.576
ControlNet [48]	LR image	0.543	21.980	0.252	0.874	0.563	0.780	27.339	0.276	0.926	0.721
ZoomLDM	Uncond	0.591	23.217	0.260	0.936	<u>0.680</u>	0.739	29.822	0.235	0.965	0.741
	GT emb	0.599	23.273	0.250	<u>0.946</u>	0.672	0.732	29.236	0.245	<u>0.974</u>	0.753
	Infer emb	<u>0.609</u>	<u>23.407</u>	0.229	0.957	0.719	<u>0.779</u>	<u>30.443</u>	0.173	0.974	<u>0.808</u>

Figure 7.12: Results of the ZoomLDM architecture on TCGA and BACH dataset

7.6 Summary

The mentioned papers [7, 94, 36, 118, 22] cover five different topics, but all are closely related to the concept of latent space. It is difficult to draw a clear conclusion about latent space, as it remains an open research task, as these works have shown. Latent space can be seen as a labyrinth that hide all important features, with the explainability contained within it. State-of-the-art solutions indicate that performing research in this direction can be highly beneficial and may lead to new and innovative discoveries regarding the relationship between global and local features, particularly in biomarker detection and synthetic data generation. This is especially relevant in the field of medicine, where this research direction is not yet fully explored, despite the enormous importance of the features.

Chapter 8

Our Approach

Biomarker detection, in the context of cancer detection, is one of the most important topics. For our approach, we aim to create a model that focuses on biomarker detection with the assistance of AI. Our approach can be divided into three main parts: integration of global-local relations into the architecture, creation of a controlled synthetic validation dataset, and preservation of the system’s interpretability.

The first part we propose is the integration of **global-local relations** into the neural network architecture. Enabling this feature makes the neural network model aware of image context and allows it to reason more deeply and complexly, especially in the context of histological images, which contain complex relations between different parts of the image.

The second part focuses on **controlled synthetic validation dataset**. This pipeline creates conditioned, related synthetic patches that are combined to form a complete whole-slide image. It allows us to evaluate whether the model captures important relations and enables hypothesis-driven validation of AI model.

The last complementary part to the global-local relation-aware model is **interpretability**. We emphasize this topic because of its critical importance in the medical domain.

Overall, in our approach, we focus on creating a model that is able to detect different global and local features across a whole-slide image, even if these features are far apart. To validate this model, we aim to create a controlled synthetic validation dataset. We also focus on developing a prediction model in an interpretable way and validating it using heatmaps produced by the same biomarker detection model. All the mentioned parts are further explained in the following sections.

8.1 Approach Overview

We analyzed the pathologist’s point of view and identified the core concepts we want to focus on. Based on this theoretical analysis, we can now propose a practical solution that addresses these concepts.

For a better understanding of our proposed solution, we firstly need to analyze Fig. 8.1, which represents the common pipeline of biomarker detection. This process has multiple problems. In this process are used datasets, which are weakly annotated and not fully interpretable. When it comes to people out of the domain, these datasets are huge bottleneck in terms of understanding and even working with them. The second problem is the AI system, which makes predictions. These systems are black boxes, focusing mostly on processing of local patch level features, and in most cases, the training process is not interpretable. By combination of these datasets and AI system we get not-interpretable results and training, what makes these system not reliable and hard to use in clinical environments.

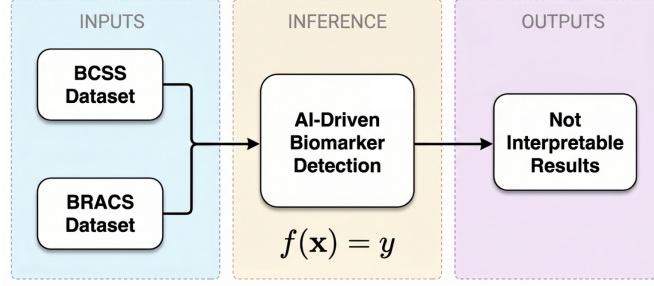


Figure 8.1: Current approach of biomarker prediction

As these problems were mentioned, we propose our solution, which focuses on addressing and resolving these problems. The global perspective of our proposed solution can be seen in Fig. 8.2.

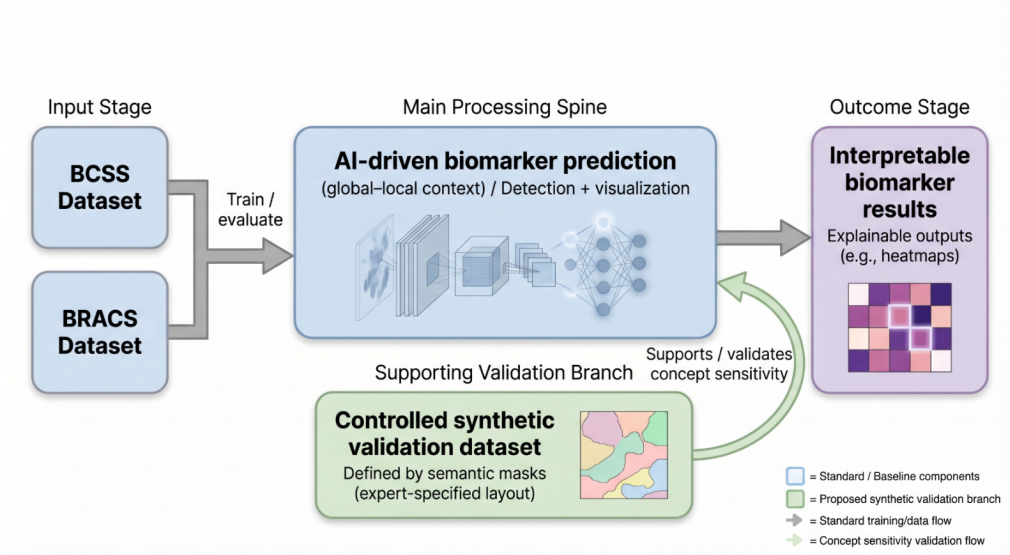


Figure 8.2: Proposed approach for biomarker prediction, where **blue part** represent proposed solution for AI model focused on global-local relations and **green part** represent proposed controlled synthetic validation dataset used for hypothesis-driven validation

The first part consists of an AI model that focuses on the fusion of local and global relations. This model is trained on weakly supervised datasets. We want the model to learn these relations, but since whole-slide images will not be fully

annotated, the use of weak annotations is crucial for applying this model in real-life scenarios.

After the model is trained on a weakly supervised dataset, we need to validate whether it can learn and highlight complex relations. To achieve this, we propose the generation of a controlled synthetic dataset based on defined rules. This dataset is crucial for two purposes: one is model validation, and the second is improving data understanding for AI developers outside the medical domain. By generating datasets according to specific rules, developers can validate their models for multiple tasks. Validation can be challenging in cases with limited data. The proposed generation pipeline aims to address these problems.

The final step of our proposed pipeline is to use the embeddings from our AI model as inputs for various downstream tasks. The first downstream task, which we propose, is the detection and visualization of histopathological biomarkers using heatmaps. This pipeline not only aids in data understanding by incorporating global and local features into the model but also produces interpretable predictions, which are crucial for pathologists.

Chapter 9

AI-driven Biomarker Prediction

The main part of the proposed solution for AI-driven biomarker prediction is the architecture so-called Relation-Aware Hierarchical Image Pyramid Transformer (RA-HIPT). The architecture of the proposed solution can be seen in Fig. 9.2. It processes two resolutions of histological images.

The first global context branch, containing the HIPT architecture, processes images of size 4096x4096. By sequentially processing all patches of the whole-slide image, we obtain an embedding representation of the entire slide. By applying a feature selection and projection, we get smaller embeddings of the most important features of the whole-slide image, which can be further processed in the dual-branch attention module of vision transformer aggregator block. This branch has two options how HIPT can extract embeddings: embeddings of 4096x4096 patches, but also embeddings of smaller 256x256 patches. In experiments we will evaluate both options.

The second local evidence branch processes patches of size 256x256. These are processed by a vision transformer with two distinct modifications. The first

modification is the replacement of multi-head attention with dual-branch attention. This dual-branch attention is further described in Sec. 9.2. The second modification is a custom-optimized loss function, which is further explained in Sec. 9.3. On top of this base model we propose third adjustment, which is graph based regularization, proposed in a Sec. 9.6.

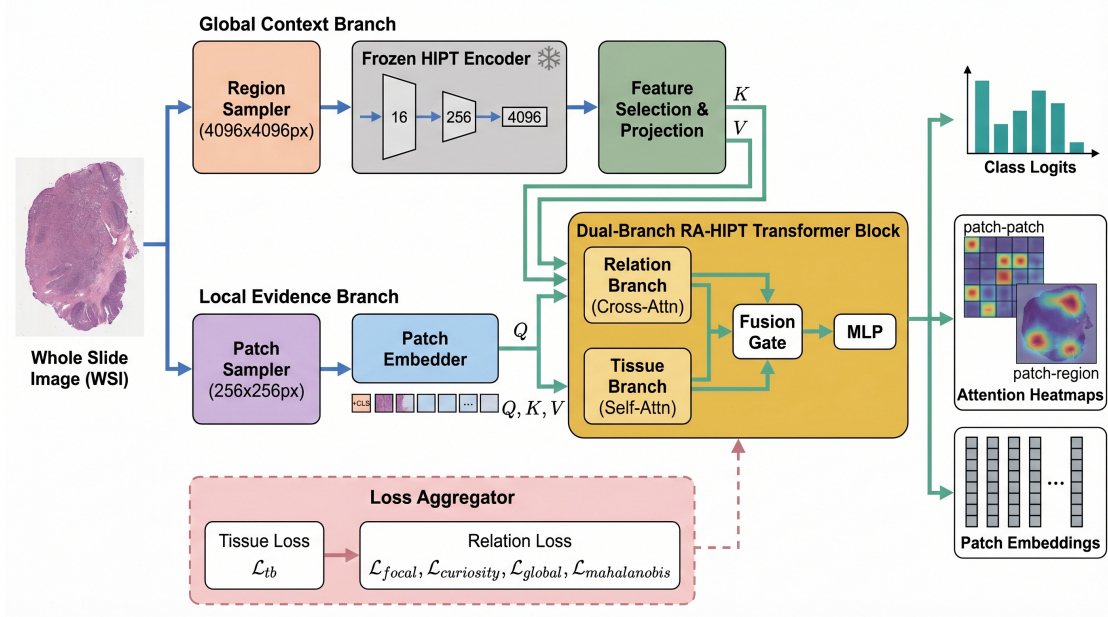


Figure 9.1: AI-driven biomarker detection and visualization pipeline, using RA-HIPT architecture

9.1 Whole Slide Embeddings Extraction

The embedding extractor is an important part of the architecture. Its role is to extract global features of whole-slide images, which are in most cases lost during slide processing.

HIPT architecture is a suitable choice for performing the mentioned global feature extraction. It processes 4096x4096 patches hierarchically from bottom to

top. It starts with 16x16 patches, then processes 256x256 patches, and finally processes the full 4096x4096 resolution. Through this processing, embeddings are aggregated. This process produce embeddings representing both local and global features.

Linear projection of the most important features is important because processing the vast number of embeddings corresponding to all 4096x4096 patches of a whole-slide image creates a tensor of huge size. Working with this tensor directly would be resource- and memory-intensive. Therefore, we use linear projection. Before applying linear projection, embeddings need to be normalized. Multiple normalization techniques, such as simple flattening, mean pooling, or attention pooling, can be used. In our experiments we focus only on mean pooling.

9.2 Vision Transformer with Dual-Branch Attention Mechanism

The whole prediction of the patch is performed by a vision transformer. This transformer does not have any significant modifications besides the replacement of self-attention with dual branch attention module. This module is divided into two brancher: tissue branch and relation branch. The proposed block also connects these two parallel branches with gated fusion to aggregate their attentions.

9.2.1 Tissue Branch

The tissue branch uses ordinary multi-head self-attention, as is used in transformer, over patch tokens. This branch is used to capture structures, relations and features within local patches. Its role is to preserve local structural context of slide. During

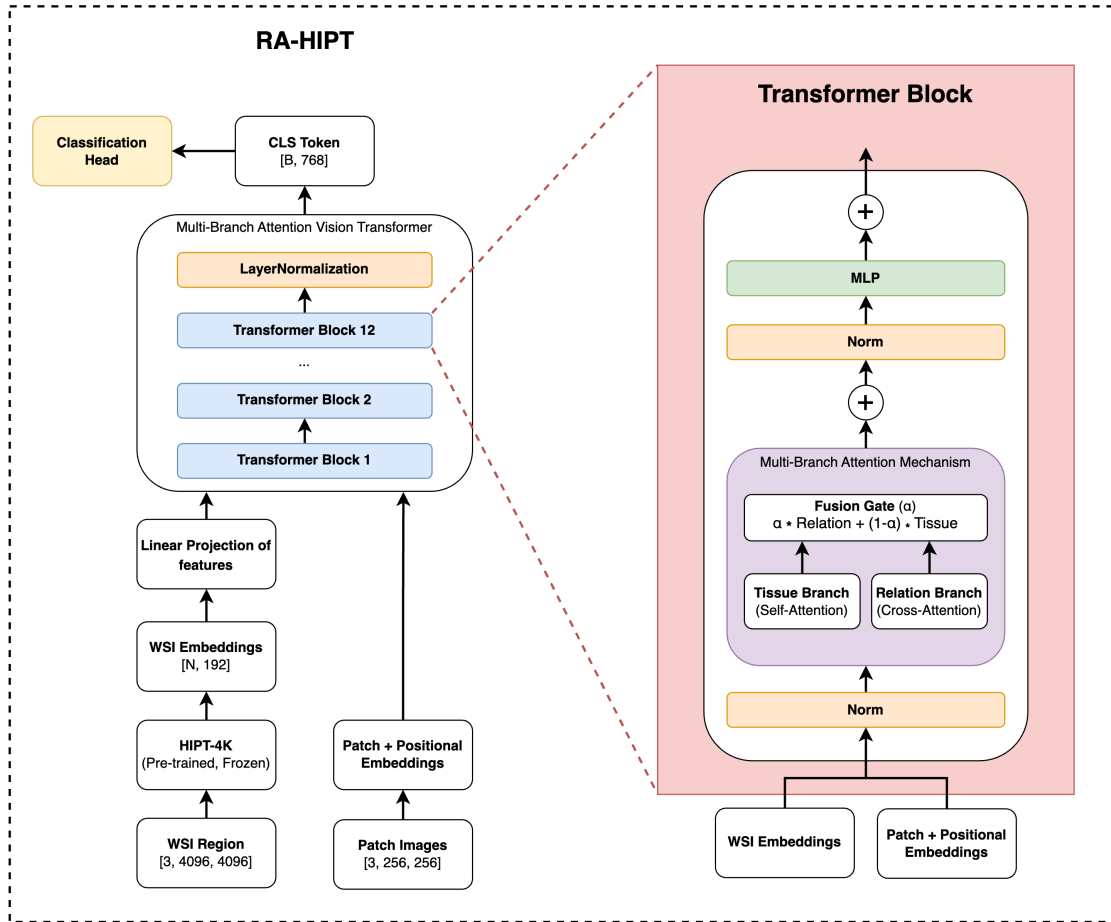


Figure 9.2: RA-HIPT transformer block with relation (cross-attention) and tissue (self-attention) branches

training of this branch we use tissue-attention loss function that focus attention on tissue regions and suppresses attention on background. This objective is done with the help of pseudo-mask supervision.

9.2.2 Relation Branch

The cross-attention, contained in the relation branch, provides the vision transformer with the ability to focus on relations between the currently processed patch and the most important feature embeddings of the whole-slide image. The equa-

tion of the used cross-attention can be seen in Eq. 9.1. An additional advantage provided by the cross-attention is that it enables the creation of heatmaps of distinct feature heads, which addresses the interpretability problem.

The cross-attention mechanism is defined as follows. Given patch queries $\mathbf{q} \in R^{B \times N_q \times d}$ and WSI embeddings $\mathbf{k}, \mathbf{v} \in R^{1 \times N_k \times d_{wsi}}$, the computation proceeds as:

$$\text{CrossAttention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{Proj} \left(\text{Reshape} \left(\text{softmax} \left(\frac{\mathbf{QK}^T}{\sqrt{d_h}} \right) \mathbf{V} \right) \right) \quad (9.1)$$

where the intermediate computations are:

- $\mathbf{Q} = \mathbf{qW}_q$ (patch queries)
- $\mathbf{K} = \text{Expand}_{B, N_q}(\mathbf{kW}_k)$ (WSI keys, expanded)
- $\mathbf{V} = \text{Expand}_{B, N_q}(\mathbf{vW}_v)$ (WSI values, expanded)

9.3 Attention Module based Custom Optimized Loss Function

The loss function represents the main regularization component. The whole loss function with all its regularization terms is shown in Eq. 9.2. By using this custom loss function, we target multiple important latent space features and can adjust the neural network's latent space based on our needs. The loss function is composed of three parts: tissue branch regularization loss, relation branch regularization loss and latent space regularizations. The relation branch regularization loss can be seen more detailed in Eq. 9.3. All of these functions are adapted versions of existing mathematical functions.

$$\mathcal{L}_{\text{total}} = \lambda_t \mathcal{L}_{\text{tissue}} + \lambda_r \mathcal{L}_{\text{rel}} + \lambda_h \mathcal{L}_{\text{head-div}} + \lambda_e \mathcal{L}_{\text{entropy}} \quad (9.2)$$

$$\mathcal{L}_{\text{rel}} = \lambda_f \mathcal{L}_{\text{focal}} + \lambda_c \mathcal{L}_{\text{curiosity}} + \lambda_g \mathcal{L}_{\text{global}} + \lambda_m \mathcal{L}_{\text{mahalanobis}} \quad (9.3)$$

Head-diversity regularization prevent attention-head collapse in the relation branch. We penalize high similarity between raw per-head cross-attention maps. Let $\mathbf{A} \in R^{B \times H \times N_q \times N_r}$ be relation-branch attention weights, where H is the number of heads. For each sample and each head, attention is flattened and ℓ_2 -normalized, and pairwise cosine similarity is computed. Only similarities above margin m are penalized:

$$L_{\text{head-div}} = \frac{1}{|\mathcal{Q}|} \sum_{(h_1, h_2) \in \mathcal{Q}} [\max(0, \cos(\hat{\mathbf{a}}_{h_1}, \hat{\mathbf{a}}_{h_2}) - m)]^2 \quad (9.4)$$

where $\mathcal{Q}$ contains all unordered head pairs and m is the similarity margin.

This encourages heads to attend to different relation patterns and improves interpretability.

Attention-entropy regularization is addition to head diversity. This regularization term focus on the entropy of relation attention and to avoid too much uniform or sharp attention. For relation branch attention $\mathbf{A}$, entropy is computed with the main dimension:

$$H(\mathbf{A}) = - \sum_{k=1}^{N_r} A_k \log(A_k + \epsilon) \quad (9.5)$$

then normalized by $\log(N_r)$ and matched to a target entropy ratio r :

$$L_{\text{entropy}} = \left(\frac{H(\mathbf{A})}{\log(N_r)} - r \right)^2 \quad (9.6)$$

Tissue/Background Loss is supervised using a pseudo-masks, containing binary segmentation of tissue and background. Let $\mathbf{a} \in R^{B \times N}$ be aggregated tissue attention and $\mathbf{m} \in \{0, 1\}^{B \times N}$ the tissue mask. The loss function is defined with two core concept, the background attention decrease and the tissue attention increase, defined as follows:

$$L_{\text{fg}} = -\frac{1}{B} \sum_{b=1}^B \frac{\sum_{i=1}^N a_{b,i} m_{b,i}}{\sum_{i=1}^N m_{b,i} + \epsilon} \quad (9.7)$$

$$L_{\text{bg}} = \frac{1}{B} \sum_{b=1}^B \frac{\sum_{i=1}^N a_{b,i} (1 - m_{b,i})}{\sum_{i=1}^N (1 - m_{b,i}) + \epsilon} \quad (9.8)$$

The tissue loss is then:

$$L_{\text{tissue}} = \lambda_{\text{tissue}} L_{\text{fg}} + \lambda_{\text{bg}} L_{\text{bg}} \quad (9.9)$$

Adapted Focal Loss is a complementary regularization term for curiosity loss. This term focuses on labeled patches. We calculate the focal cross-entropy loss on patch predictions against their labeled classes, which trains our model to align predictions with ground truth labels and provides the ability to classify. This loss is calculated only over the masked subgroup of patches that are labeled. The equation can be seen in Eq. 9.10.

$$L_{\text{focal}} = \frac{1}{N} \sum_{i=1}^N (1 - p_{t_i})^\gamma \cdot \text{CE}(\text{logits}_i, \text{label}_i) \quad (9.10)$$

where:

- $p_{t_i} = p_{label_i}^{(i)}$ is the predicted probability for the true class
- $\gamma = 2.0$ is the focusing parameter
- $CE(logits_i, label_i)$ is the cross-entropy loss
- N is the number of labeled patches

Curiosity Loss focus, in contrast, on the not labeled patches, which is a majority. The regularization term encourages the model to explore and learn from ambiguous or uncertain patches, typically identified by uncertainty or entropy measures. It pushes the model to remain uncertain where it lack knowledge and to explore these regions more. The equation of this regularization term can be seen in Eq. 9.11.

$$L_{curiosity} = -\frac{1}{M} \sum_{j=1}^M H(p^{(j)}) \quad (9.11)$$

where:

- $H(p^{(j)}) = -\sum_{c=1}^C p_c^{(j)} \log(p_c^{(j)} + \epsilon)$ is the entropy of the predicted distribution
- $p^{(j)} = \text{softmax}(logits_j)$ is the probability distribution for unlabeled patch j
- M is the number of unlabeled patches (where $label = 0$)
- $\epsilon = 10^{-6}$ is a small constant for numerical stability
- C is the number of classes

Global Loss focuses, on the other hand, on incorporating whole-slide embeddings into the training process. It enforces consistency between the whole-slide

embedding and patch embeddings, ensuring that patch predictions align well with the global slide representation. This provides slide-level context to patch features. The equation of the global loss can be seen in Eq. 9.14, where the loss combines cosine similarity with an MSE term for feature matching. But the first important step is that the local and global descriptors are mapped into a shared projection space, what can be seen in Eq. 9.13.

$$\mathbf{z}_{\text{local}} = \sum_{i=1}^N \alpha_i \mathbf{x}_i, \quad \alpha_i = \frac{\exp(a_i/\tau)}{\sum_{k=1}^N \exp(a_k/\tau)}, \quad (9.12)$$

where $\mathbf{x}_i$ is the patch embedding, a_i is the aggregated attention score.

$$\mathbf{g} = \text{norm}(\phi_g(\mathbf{e}_{\text{wsi}})), \quad \mathbf{l} = \text{norm}(\phi_l(\mathbf{z}_{\text{local}})), \quad (9.13)$$

where ϕ_g, ϕ_l are learnable projection heads and $\text{norm}(\cdot)$ is ℓ_2 normalization.

$$L_{\text{global}} = \lambda_{\text{cos}} (1 - \cos(\mathbf{g}, \mathbf{l})) + \lambda_{\text{mse}} \|\mathbf{g} - \mathbf{l}\|_2^2. \quad (9.14)$$

Mahalanobis Loss focus on disentanglement of feature heads of cross-attention. It calculates Mahalanobis distance, which can be seen in Eq. 9.15, between each heads and try to maximize the distance. The hypothesis is that Mahalanobis distance is more suitable than Euclidian distance because it accounts for the variance and correlation of the data, making it more precise for pushing of these feature heads apart. Trade-off of Mahalanobis distance is that its computational cost, because of computing of covariance and invariance matrices, so optimizations of its calculation is crucial step. The equation of the Mahalanobis loss can be seen in Eq. 9.16.

$$d_{ij} = \sqrt{(\mathbf{f}_i - \mathbf{f}_j)^T \Sigma^{-1} (\mathbf{f}_i - \mathbf{f}_j)} \quad (9.15)$$

where:

- d_{ij} is the Mahalanobis distance between feature vectors i and j
- $\mathbf{f}_i$ and $\mathbf{f}_j$ are feature vectors (typically from different attention heads)
- Σ is the covariance matrix of the feature space
- Σ^{-1} is the inverse covariance matrix

$$L_{mahalanobis} = -\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} d_{ij} \quad (9.16)$$

where:

- $\mathcal{P}$ is the set of all pairs (i, j) of feature vectors (attention heads)
- $|\mathcal{P}|$ is the number of pairs in the set
- d_{ij} is the Mahalanobis distance
- The negative sign ensures that maximizing distances corresponds to minimizing the loss

9.4 Loss Scheduler

The loss function used for training in RA-HIPT is composed of multiple loss terms, and each term has different goals. To maintain stability of training we use a loss scheduler that changes selected loss weights at selected epochs. This follows a specific evolving strategy: the model first learns a basic representation, and then

stronger relation-focused features.

In implementation, the scheduler reads a list of parameters (epoch, parameter, value) from configuration and applies these updates directly to loss function. This gives us control over optimization without changing model structure. For example, we start with the focus on tissue branch and then after few epoch after the model learns differences between tissue and background we can introduce the relation focused loss functions.

The main advantage is: loss function is easier to tune, less noisy in early epochs, and more robust when combining several complementary losses.

9.5 Self Supervised Learning

Besides weak-supervised mode, RA-HIPT also supports a self-supervised learning (SSL) mode. This mode focus on training class labels and learns from similarities between two augmented views of the same histology patch.

The SSL combines multiple concepts: alignment of views of embeddings, consistency of views of relation and tissue attention, and support regularization terms for diversity and stability. The goal is to learn representations so the similar augmentations of the same image are processed as the same image, and also preserve relevant tissue structure.

Even though SSL is not the main mode in our non-graph experiments, its integration is important because it helps the RA-HIPT architecture to learn structures also with the big amount of labels.

9.6 Graph Structure based RA-HIPT Modification (Graph-RA-HIPT)

In this section we propose modification of the RA-HIPT architecture, focused on graph based normalizations of attention and overall training process. The main idea is to capture with these graph structures long-distance and also local relations between tissue structures. For this reason as an complementary part we used this graph based normalization with the relation branch of our proposed dual branch module. The whole proposed modification of the architecture can be seen in Fig. 9.3.

Graph-RA-HIPT uses three relation levels:

1. Patch-to-region relations from cross-attention (G_{PR})
2. Patch-to-patch relations from tissue self-attention (G_{PP})
3. Region-to-region relations predicted by a learnable module (G_{RR})

These three structured views provide context about the tissue patches and their relations, in a form of graph edge connection.

Compared to the base non-graph version of RA-HIPT, this variant does not only optimize classification focused relation losses, but also structure attention graph made of attentions.

9.6.1 Graph Structure Based Regularization Module

The regularization graph structure is implemented in the graph relation loss pipeline. It first converts attention outputs into adjacency matrices, this adjuncencies represents edges of the individual graphs. So when produced attention weight value on specific point is small, the relation (edge weight) is small and when the value is

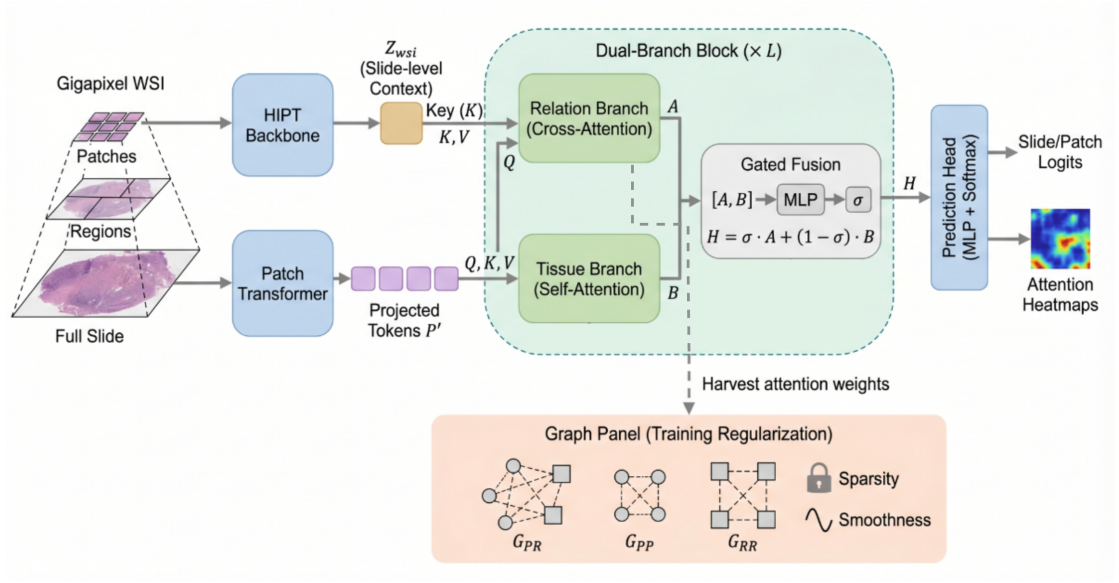


Figure 9.3: Graph RA-HIPT architecture

highlight the relation is more relevant.

For region-to-region edges, we use a kernel-prior plus learnable correction design. Each edge score is built from cosine similarity, negative ℓ_2 distance, and an MLP term. This gives a model ability to start from interpretable geometric structures, but still provide ability to learn more complex relations.

To improve robustness, the module also supports prevention of small random edge noises of attention weights during training and optimized edge construction. This helps to avoid early collapse of latent space and keeps long distance relations active.

9.6.2 Graph Structure Focused Loss functions

The graph branch uses a weighted sum of multiple different loss terms, which optimize the whole process of training. Each loss function term controls a different property of the learned graph.

For further purposes of explanation of the loss functions used let $\mathbf{A}_{pr}$ be patch-to-region adjacency, $\mathbf{A}_{pp}$ patch-to-patch adjacency, $\mathbf{A}_{rr}$ region-to-region adjacency, $\mathbf{X}_p$ patch embeddings, and $\mathbf{X}_r$ region embeddings.

Classification loss focus on a supervised part of the training. It is implemented by a cross-entropy classification term. The model predicts patch-level logits, what provides supervised signal further optimized with the other graph regularization terms explained in this section.

Formally, for logits $\mathbf{z}_i \in R^C$ and class label y_i , the loss is

$$L_{\text{cls}} = \frac{1}{N} \sum_{i=1}^N \text{CE}(\mathbf{z}_i, y_i) \quad (9.17)$$

where N is the number of supervised predictions in the batch and C is the number of classes.

In our implementation, this term is active in graph mode as weakly supervised classification is enabled. The core idea behind this loss function term is weak-supervision of the training so the model know on which features to focus. This makes complementary traing term for shaping regularization of graph structure.

Edge sparsity focus on edges and its connectivity. It push graph structure to make edges less dense and with the higher informativeness. The implementation provide combination of edge mean penalty with entropy targeting:

$$L_{\text{sparsity}}(\mathbf{A}_{pr}) = \underbrace{\text{mean}(\mathbf{A}_{pr})}_{\ell_1\text{-like sparsity}} + \beta \underbrace{(H(\mathbf{A}_{pr}) - H_t)^2}_{\text{entropy target}} \quad (9.18)$$

Graph diversity focus on a problem, which occurred very often while training of our model. Different patches collapsed to identical connection profiles, what made the representation of graph structure not accurate. Because of this we proposed loss function term, which aims exactly this problem:

$$L_{\text{div}} = L_{\text{cos-div}}(\mathbf{A}_{pr}) + L_{\text{var-div}}(\mathbf{A}_{pr}) \quad (9.19)$$

Embedding variance losses prevent collapse. It is done by regularization of patch and region embeddings with VICReg-style variance/covariance constraints:

$$L_{\text{emb-var}} = L_{\text{var}}(\mathbf{X}_p) + \gamma L_{\text{cov}}(\mathbf{X}_p), \quad L_{\text{reg-var}} = L_{\text{var}}(\mathbf{X}_r) + \gamma_r L_{\text{cov}}(\mathbf{X}_r) \quad (9.20)$$

Head diversity force cross-attention heads to learn different patterns, so we can also visualize these separate head for further explainability:

$$L_{\text{head-div}} = \frac{1}{|\mathcal{P}|} \sum_{(h_1, h_2) \in \mathcal{P}} [\max(0, \cos(\hat{\mathbf{a}}_{h_1}, \hat{\mathbf{a}}_{h_2}) - m)]^2 \quad (9.21)$$

Disentanglement into subspaces is a bit more complex concept. The idea behind this is that the model will be able to learn more complex, non visible features and relations, when we will be able to provide to model more separated embeddings. Because of this we disentangled each embedding to k different views, where each represent different features of the patch. We selected 3 views by the default, because of time constraints we were not able to experiment with different numbers of views. Further, graph structure can form edges also between these views and capture more fine grained relations between patches. These views should be also

controlled by their own regularization terms, where we proposed two constrains: within-view variance/covariance regularization (anti-collapse), and cross-subspace decorrelation.

Let $\mathbf{X}_p \in R^{B \times N \times D}$, and $\mathbf{V}_k \in R^{B \times N \times S}$ be the k -th view. The loss is:

$$L_{\text{disentangled}} = w_{\text{within}}(L_{\text{var-within}} + L_{\text{cov-within}}) + w_{\text{cross}} L_{\text{cross-cov}} \quad (9.22)$$

The term within views keeps each view informative, while the cross term penalizes covariance between different views, so each subspace captures different structure.

Total graph objective The final graph objective is:

$$L_{\text{graph-total}} = L_{\text{cls}} + L_{\text{sparsity}} + L_{\text{div}} + L_{\text{emb-var}} + L_{\text{reg-var}} + L_{\text{head-div}} + L_{\text{disentangled}} \quad (9.23)$$

In experiments, not all components are enabled at the same time. We tune active terms and their weights based on stability and collapse diagnostics, with the help of the loss scheduler explained in a previous section.

Chapter 10

Controlled Synthetic Validation Dataset

From the perspective of synthetic data generation, we need to focus on five main concepts. These concepts can be seen in Fig. 10.1. The first is the specific morphological concept. This concept is chosen by a domain or non-domain expert. Based on these structures and data features, the user is able to create concept-based data. These concepts are crucial to integrate for further AI model validations, which will be done with the synthetic dataset. Starting with the first step of the pipeline, we need to train a latent space, using encoder/decoder architecture, which is trained on a reconstruction task. This model is crucial for the second step, which is the training of the semantic-mask-conditioned diffusion model. By combining the diffusion model and the provided morphological concepts, we can generate interpretable synthetic patches. Therefore, these patches are generated they are not stitched together to form one continuous region, because of that we introduce inpainting as a fourth step. Resulting these patches, we get controlled synthetic interpretable WSIs used as a validation dataset with integrated concepts.

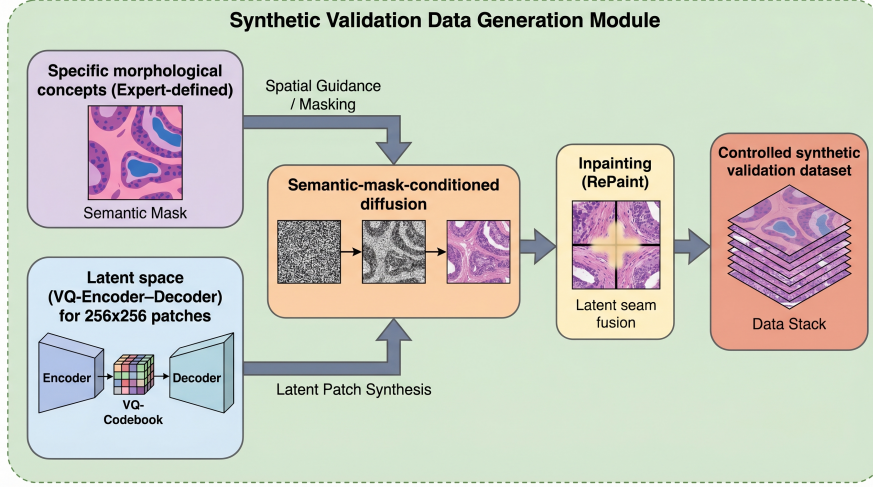


Figure 10.1: Pipeline of controlled synthetic validation dataset generation

10.1 Overview and Pipeline

The pipeline shown in Fig 10.2 contains three parts.

First, an encoder–decoder provides a low-dimensional latent representation of histology images.

Second is a diffusion model, which use latent space and conditioning with a semantic mask. The random noise is sampled by DDIM sampler and mask guidance, producing a new synthetic patch consistent with the specified tissue classes.

Third, procedural region generation arranges multiple patches on a grid, defines a binary mask, and runs RePaint inpainting in latent space so seams are filled by the same diffusion model and semantic conditioning. This process creates a seamless synthetic conditioned region. Local structures comes from mask-conditioned diffusion and global patch connections from inpainting at patch boundaries.

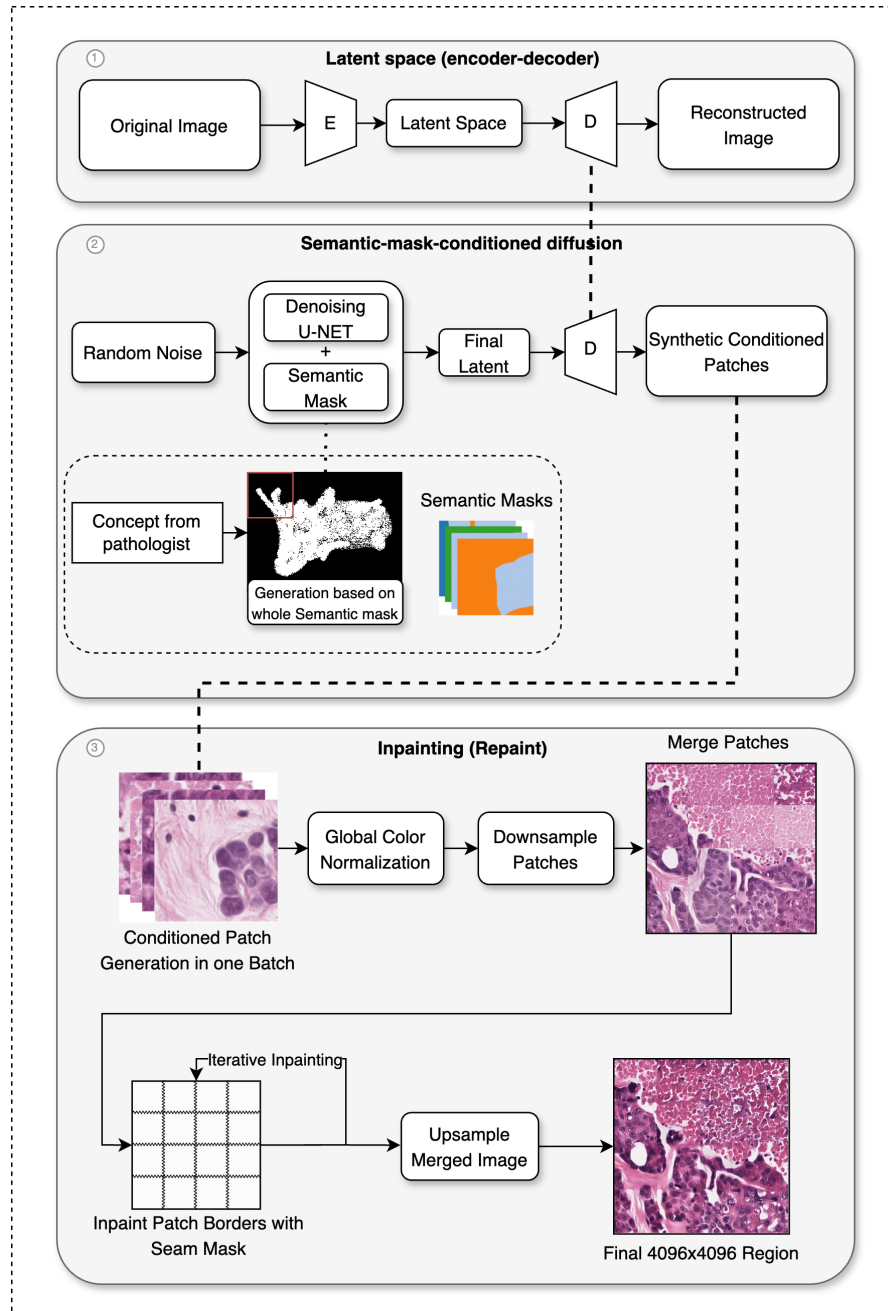


Figure 10.2: **Pipeline of controlled synthetic validation dataset generation.** (1) Latent autoencoder (2) Semantic-mask-conditioned diffusion (3) Patch stitching (Inpainting - RePaint)

10.2 Latent Space (encoder–decoder)

The first step, shown in Fig. 10.2, represents the encoder and decoder architecture. This architecture is the starting point of our pipeline. It is trained to create reconstructed images of real histological patches. The most important parts of the architecture are the latent space and the decoder itself. With this decoder applied to the latent space, we are able to generate image reconstructions from latent representations. The architecture used in this step is VQ-VAE, which is suitable for this task in terms of training time and reconstruction quality.

10.3 Semantic-mask-conditioned Diffusion

The second part of the architecture represents the semantic-mask-conditioned diffusion model. The function of this model is to create synthetic histological patches, which can be used in step three for further processing. Here, the focus is not image quality but the correctness of generating structures in exact locations. We used UNET architectures in our latent diffusion model for synthetic data generation. A pathology concepts are represented as a semantic mask at image resolution. One-hot encoding of labels was created and downsampled to the latent resolution, creating conditioning c that enters the UNet together with the noisy latent z_t and timestep t .

10.3.1 Classifier-free Guidance

To strengthen impact of the semantic map, we apply classifier-free guidance when forming the noise prediction. Let $\varepsilon_\theta(z_t, t,)$ denote the model with conditioning dropped and $\varepsilon_\theta(z_t, t, c)$ the fully conditioned prediction. The guided estimate is

$$\tilde{\varepsilon}_t = \varepsilon_\theta(z_t, t,) + w (\varepsilon_\theta(z_t, t, c) - \varepsilon_\theta(z_t, t,)), \quad (10.1)$$

with guidance scale $w \geq 1$.

Only sampling uses Eq. 10.1. The training uses the standard conditional objective with random dropout of c .

10.4 Inpainting (Repaint)

The final stage merges a grid of semantically conditioned patches into one synthetic conditioned region and removes visible seams by inpainting. Patches from stage 2 are placed on a canvas in a regular grid. A binary inpainting mask m marks patch interiors as known ($m = 1$) and bands along the grid lines as unknown ($m = 0$). RePaint-style inpainting is applied in latent space with the following steps:

1. The canvas is encoded to latent z_0
2. The binary mask is downsampled to the latent grid
3. Each denoising step known regions are replaced with the noised known latent while unknown (seam) regions are denoised

The denoiser acts only on the unknown region, so we do not change areas which are correct. Optionally, at each t we run K resampling iterations, such as original Repaint implementation do. The same diffusion model and semantic conditioning are used throughout, so inpainted seams match both the pre-defined tissue classes and the surrounding patches.

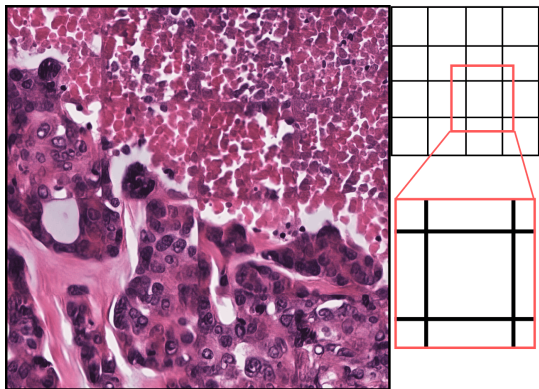


Figure 10.3: Straight mask example

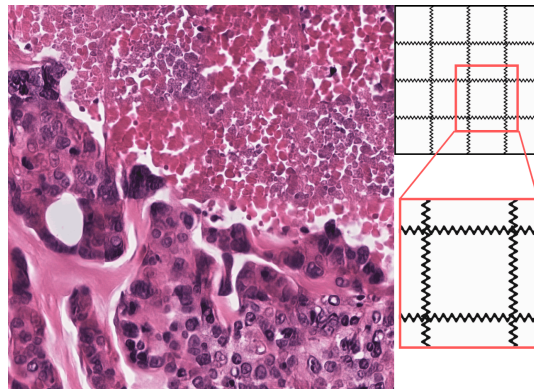


Figure 10.4: Wavy mask example

10.4.1 Grid Mask Inpainting

The binary mask that define seams, can be used in a multiple different styles. In our thesis we test two different: straight and wavy.

The seam is a fixed band along each grid line: at vertical boundaries between columns and horizontal boundaries between rows.

Straight Mask The straight masks are just normal straight lines alongside patch borders. The example can be seen in Fig. 10.3.

Wavy Mask The wavy masks are shifted along each grid line by a sinusoidal offset so that the inpainting boundary is wavy rather than a straight line. This technique helps to inpaint in a more irregular manner and helps to prevent border artefacts. The example can be seen in Fig. 10.4.

10.4.2 Higher Resolution Region Generation with In-painting

To produce large regions, such as 4096x4096 pixels, without running memory heavy full-resolution inpainting over the entire canvas, we use a downsample–inpaint–upsample

strategy, shown in third step of Fig. 10.2. We first generate a grid of 256x256 patches that compose the whole generated region. The global color normalization is applied so that all patches share similar colors. Each patch and its semantic mask are then downsampled with bilinear interpolation by a factor 2. The inpainting seam width in pixels is reduced in proportion so the seam thickness is preserved in low resolution. RePaint then combines the entire low-res grid in a **single** global inpainting pass, producing one stitched region whose size is the low-res of the final region. Finally, we apply a **deterministic upsampling** of the stitched image to get 4096x4096 region using bilinear interpolation. This way, seam blending is performed once at lower resolution, reducing compute while still creating a large image. By procedurally generating regions in this manner, we can merge them to create a complete whole slide image. This image is guided, interpretable, and suitable for hypothesis-driven validation.

10.5 Color Normalization

To reduce color differences between patches and make seams less visible, three normalization strategies are used. Example of the image before and after normalization can be seen in Fig.10.5.

10.5.1 Reinhard Normalization

Reinhard (LAB) normalization [83] matches the colour statistics of generated patches to a reference image in CIE Lab color space. The reference mean μ_{ref} and standard deviation σ_{ref} are computed from a reference image and each patch is transformed so that its LAB mean and std match the reference. This is standard for H&E-like histology images and improves color consistency of all patches. This normalization helps with much smoother inpainting results.

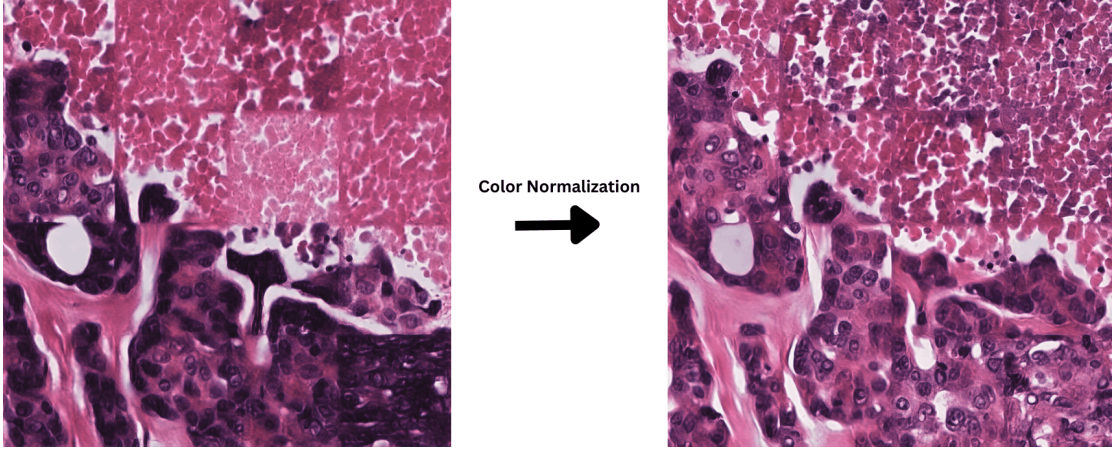


Figure 10.5: Combination of Reinhard and edge normalization without inpainting

$$\text{LAB}'(p) = \frac{\text{LAB}(I(p)) - \mu_{\text{src}}}{\sigma_{\text{src}}} \odot \sigma_{\text{ref}} + \mu_{\text{ref}} \quad (10.2)$$

10.5.2 Edge Normalization

Edge normalization process each patch so that the colour of its border matches the edge of the neighbouring patch. For each patch all edges are considered. The mean colour of current patch edge $\bar{\mu}_e$ is compared to the mean colour of the neighbour's edge $\bar{\mu}'_e$, and a color correction is applied in the way that edges blend toward the neighbour's colour $P_{\text{corrected}}(p)$. The result reduces visible seams at patch boundaries before inpainting process is performed. It is important to use this as a complementary part to Reinhard normalization because Reinhard does not controll neighborhood colors, so color inconsistency is still present.

$$\bar{\mu}_e = \frac{1}{|\Omega_e|} \sum_{p \in \Omega_e} P(p), \quad \bar{\mu}'_e = \frac{1}{|\Omega'_e|} \sum_{p' \in \Omega'_e} P'(p') \quad (10.3)$$

$$P_{\text{corrected}}(p) = P(p) + \lambda \sum_e w_e(p) (\bar{\mu}'_e - \bar{\mu}_e) \quad (10.4)$$

10.5.3 Latent Noise Initialization Normalization

To have consistent appearance across patches, the initial latent noise x_T is shared across patches. A single initial noise tensor ε_0 is drawn and broadcast to all patches so that every patch starts from the same z_T . The z_T is a combination of this shared noise and independent per-patch noise. Shared or correlated z_T tends to reduce colour and texture drift across the grid.

$$z_T^{(i)} = \varepsilon_0, \quad i = 1, \dots, N \quad (10.5)$$

$$z_T^{(i)} = \tau \varepsilon_0 + (1 - \tau) \varepsilon_i, \quad \varepsilon_0, \varepsilon_i \sim \mathcal{N}(0, I) \quad (10.6)$$

Chapter 11

Experiments

This chapter contains experiments, which can be divided into four separate parts. The first part is dataset preparation, which includes publicly available datasets as well as one synthetically created datasets. The second part shows metrics used for experiment validation. The third part presents experiments with controlled synthetic validation dataset. Further experiments with RA-HIPT are provided in the fourth section.

For each best setup of experiment provided in this section, we propose a separate detailed training report, which can be found in complementary assets of this thesis.

11.1 Datasets

For experiments, three different datasets were used. The first two datasets were real-world datasets representing clinical data: BRACS and BCSS. The last custom datasets was synthetically created with the help of generative neural networks.

11.1.1 BReAst Carcinoma Subtyping Dataset (BRACS) [15]

This dataset is a publicly available dataset focused on breast cancer histopathological images. It contains seven distinct subtypes of breast carcinoma, which can be seen in Fig. 11.1. The dataset contains region-level patches and whole-slide images with bounding box annotations. All images were scanned at 40x magnification, which makes the images highly detailed. Three levels of detail can be seen in Fig. 11.2. The dataset itself is weakly annotated, which makes it challenging to use but perfectly reflects a real clinical environment.

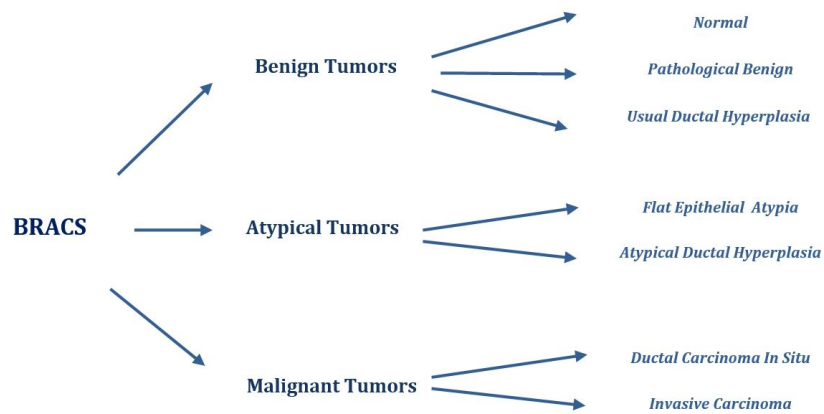


Figure 11.1: Class represented in BRACS dataset

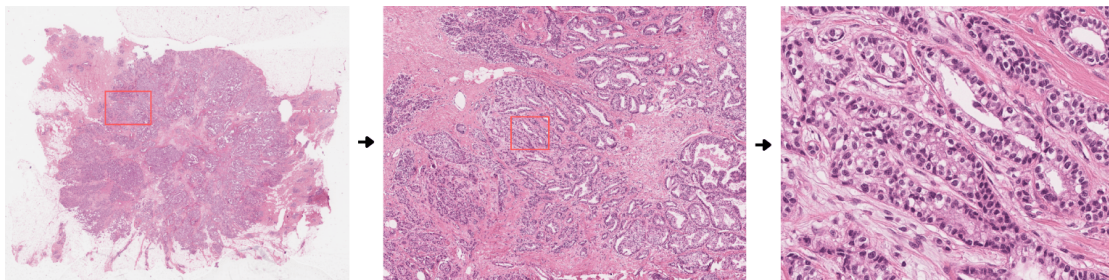


Figure 11.2: Examples of BRACS dataset in different resolutions

BRACS contains 547 whole-slide images from 189 patients. Overall, the dataset contains 4539 annotated regions of interest. An example of the anno-

tations can be seen in Fig. 11.3. Another important feature of the dataset is its class imbalance, which needs to be considered during model training and validation. Exact class numbers are 484 N, 836 PB, 517 UDH, 756 FEA, 507 ADH, 790 DCIS, and 649 IC. This dataset is used for training the RA-HIPT architecture, and it is also a suitable choice for validating heatmaps.

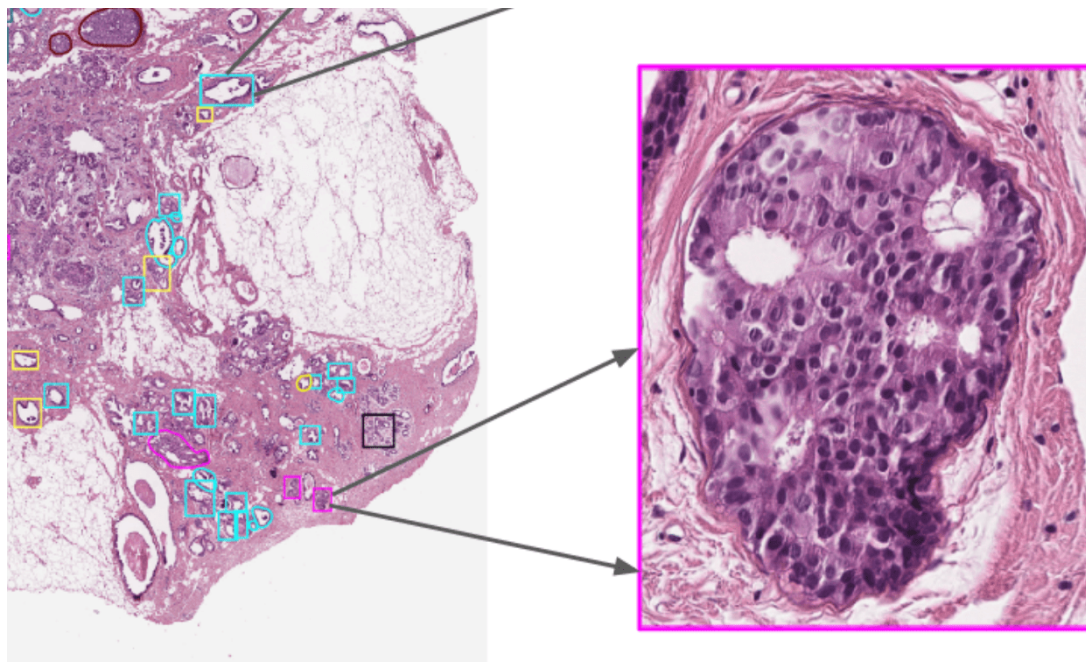


Figure 11.3: Examples of BRACS dataset and its annotations

11.1.2 Breast Cancer Semantic Segmentation Dataset (BCSS) [9]

The second dataset is also a histopathological dataset. It focuses on multi-class tissue segmentation of breast cancer areas. Images in the dataset are H&E stained and focus specifically on triple-negative breast cancer cases, which are typically high-grade, poorly differentiated tumors.

The dataset contains 150 whole-slide images. Regions of interest are selected

from these whole-slide images, and around 20 000 annotations were made over these areas. An example of an annotated area can be seen in Fig. 11.3. The dataset contains 22 classes used for annotations. Some of the classes are tumor epithelium, in situ or non-invasive epithelium, stroma, inflammation, necrosis, and even the background itself. This dataset is used primarily for training generative neural networks, especially for the conditional diffusion model.

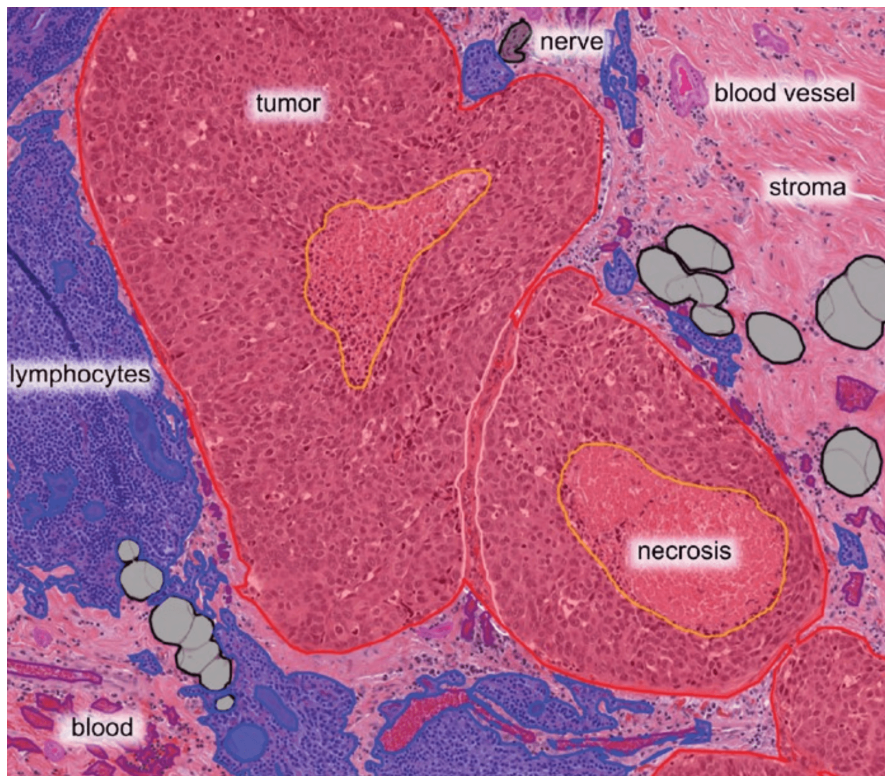


Figure 11.4: Example of BCSS dataset with segmentation masks

11.1.3 Controlled Synthetic Validation Dataset

The third dataset is a synthetic histopathological dataset, created with our proposed controlled synthesis image generation pipeline. It focuses on the controlled validation of the proposed RA-HIPT architecture. The dataset contains two types of patches: regions of size 4096x4096 pixels, and patches of size 256x256 pixels.

For each patch, a label classification is also provided. The example of the synthetic whole slide image can be seen in Fig. 11.5.

The dataset contains 38 whole slide images of resolution around 13000x13000 pixels. These images were generated based on 38 multi-class semantic masks from the Histo-Seg dataset [2]. This dataset contain segmentations of skin cancer tissue, containing 12 different classes such as epidermis, keratin, glands, inflammation or basal cell carcinoma.

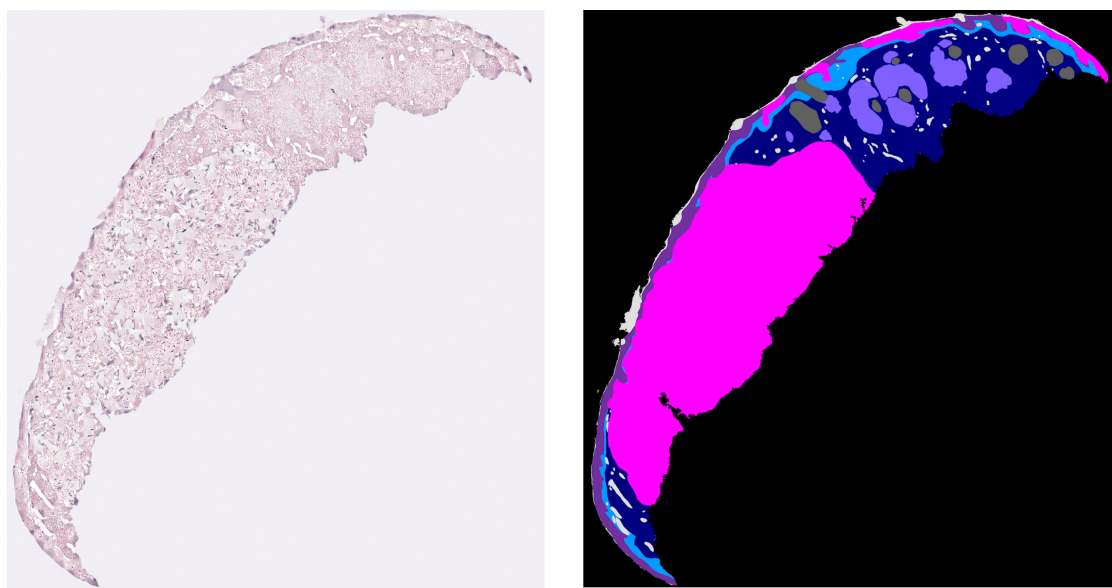


Figure 11.5: Example of controlled synthetic whole slide image with segmentation masks (made with our proposed synthetic image generation pipeline)

11.2 Evaluation Metrics

Evaluation metrics used in the following experiments can be divided into two categories. The first category includes metrics used in synthetic data generation. The second category includes metrics used for AI-driven biomarker detection architecture validation.

11.2.1 Synthetic Data Generation Metrics

The Inception Score (IS) measures how well generated images correspond to real ImageNet classes, which are used for pre-training the Inception v3 model. This model is a core component for IS calculation. The metric is maximized when images are easily classifiable and also diverse across categories. It also reports mean and standard deviation to reduce variance. Regarding limitations, the metric cannot compare to real data, as it relies solely on ImageNet classes, and it also exhibits high bias on small samples. The equation for this metric can be seen in Eq. 11.1.[12]

$$\text{IS} = \exp \left(E_{x \sim p_g} [D_{\text{KL}}(p(y|x) || p(y))] \right) \quad (11.1)$$

where:

- $x \sim p_g$ denotes samples x drawn from the generator's distribution p_g
- $p(y|x)$ is the conditional class distribution, representing the probability distribution over classes y given a generated image x , typically obtained from a pre-trained Inception network
- $p(y)$ is the marginal class distribution, representing the overall probability distribution over classes y across all generated images, obtained by averaging $p(y|x)$ over many generated samples
- $D_{\text{KL}}(p(y|x) || p(y))$ is the Kullback-Leibler divergence, measuring how the conditional distribution diverges from the marginal distribution
- $E_{x \sim p_g}$ denotes the expectation (average) over generated samples

Fréchet Inception Distance (FID) is a metric that calculates the distance between feature vectors of real and generated images. This score measures the similarity of two groups using a pre-trained Inception v3 model. A lower FID indicates more similar images, meaning that if FID equals 0, the two image groups are identical. Compared to IS, FID combines confidence in both quality and diversity. It evaluates synthetic images based on the statistical distribution of synthetic images compared to the distribution of real images. The equation for this metric can be seen in Eq. 11.2. This metric is one of the most relevant for generative tasks.[119]

$$\text{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|^2 + \text{Tr}(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{1/2}) \quad (11.2)$$

where:

- $\boldsymbol{\mu}_r$ and $\boldsymbol{\mu}_g$ are the mean feature vectors of real and generated images respectively, generated with inception model
- $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_g$ are their covariance matrices
- $\|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2$ represents the difference in average feature activations
- $\text{Tr}(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g)$ represents the total variance of both distributions
- $-2 \text{Tr}((\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{1/2})$ measures the overlap in feature covariance structure

SSIM (Structural Similarity Index) is a metric that measures similarity between two images focusing on luminance, contrast, and structural information. This metric process local features and is designed to imitate human vision when comparing image quality. In our experiements, SSIM is used to quantify how closely synthetic histology patches preserve the structural organisation of tissue relative to real reference images. Higher SSIM values indicate that local colors

and global patterns are well captured, which is important for downstream tasks where the shape and arrangement of tissue structures matter. The equation of the SSIM, can be seen in Eq. 11.3.[111]

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (11.3)$$

where:

- μ_x and μ_y are the mean intensities of x and y
- σ_x^2 and σ_y^2 are the variances of x and y
- σ_{xy} is the covariance between x and y
- C_1 and C_2 are small constants for numerical stability

PSNR (Peak Signal-to-Noise Ratio) is a reference metric for measuring the fidelity of reconstructed or generated images. In our case we focus mostly on the image generation task. It use calculation of the Mean Squared Error between a synthetic image and the real image. The MSE represents the cumulative squared error between the compressed and the original image, where PSNR represent a measure of the peak error. In our experiments, PSNR is applied to evaluate the similarity of values between real and synthetic histology patches. Higher PSNR values represent the lower reconstruction error. The equation of the PSNR, can be seen in Eq. 11.5.[45]

$$\text{MSE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2 \quad (11.4)$$

$$\text{PSNR}(x, \hat{x}) = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(x, \hat{x})} \right) \quad (11.5)$$

where:

- N is the number of pixels
- x_i and $\hat{x}_i$ are pixel intensities at position i
- MAX is the maximum possible pixel value (e.g. 255 for 8-bit images)

Precision in our seam analysis is defined as boundary-on-seam precision and measures how often superpixel or SAM boundaries, depends on method used, match location of the seam. It was complex to quantify visibility of the seam, because it is more visual task rather than measurable one. Because of that we tried to propose our metric to quantify seam visibility by the fraction of detected tissue boundaries that are aligned with tiling seams instead of following real histology structure. A lower value is better, as it means that most boundaries segmented by the segmentation method lie inside tissue regions rather than along patch borders. In our experiments, we use boundary-on-seam precision to show that the proposed inpainting-based stitching produces seams that are not much visible.

Recall is used as seam recall and measures how many seam pixels are close to any SAM or superpixel boundary. This metric focus on theory that if the seam match object segmentation border, the seam is visible. In this metric we also focus on lower values, because they measure that only a small proportion of seam pixels align contour-like segmented structures. In our evaluation, seam recall is used together with boundary-on-seam precision for ablation studies.

11.2.2 AI-driven Biomarker Detection Methods for Solution Analysis

Methods focused on AI-driven biomarker detection need to be divided into two groups. These groups come from the complexity of task, on which we are focusing. Task's challenge is that we can not really validate results of the attention heatmaps without the domain expert, so we need to focus more on embeddings and collapse analysis methods rather than some quantitative metrics. Because of this we divided analysis into two groups of methods: collapse analysis methods and attention analysis methods. As a complementary metric we used accuracy metric for weak-supervised learning mode.

Collapse analysis methods In biomarker detection experiments, collapse representation are risky: patch tokens, region tokens, or attention maps can become almost constant, because of this problem. In this case the model looks trained but does not keep useful diversity in the latent space. Because we have just weak-supervised dataset it is difficult to obtain the ground-truth explanations of attentions. In this case we complement qualitative plots with latent space collapse diagnostics, which are monitored during training.

In our implementation, collapse analysis aggregates several statistics over batches, which can be logged per epoch. These statistics were selected as follows:

- **Dispersion of embeddings**, tracking the mean standard deviation of patch embeddings across tokens and feature dimensions. Small values over long training time suggest that tokens or regions collapse to similar vectors.
- **Mean L2 norm of patch embeddings**, helps to detect unstable vanishing or growing norms

- **Variability across query tokens** are averaged, what analyze variability because uniform attention across patches is sign that the model does not sperate local structures. With this metric we analyze impact of relations between patches and regions.

Together, these methods give a sufficient analysis about training health, and provide insight that the loss weights do not silently remove useful features from the latent space.

Attention weights (Embeddings) analysis methods For the AI-driven biomarker pipeline we rely on interpretable views of embeddings and attention, because pixel-level heatmaps alone are not interpretable without domain review. Because of this we use a set of complementary visualization tools: cross-attention between patches and WSI regions, tissue self-attention inside the patch, and optional learned region-region structure, used only with graph mode of proposed architecture.

Principal Component Analysis (PCA) is applied directly on tensors, mainly in two dimensions. We use PCA of patch embeddings coloured by the strongest region connection, PCA of region embeddings, and PCA of patch-to-region attention vectors. This is useful to see if the patches develop different attention strategies or they all map to the same pattern. We additionally plot PCA per attention head, so we can check if specific heads capture the same features. When a learned region adjacency exists, we visualise PCA of region edge vectors and a small bar-chart panel of explained variance for embeddings, attention patterns, and region features, indicating impact of the signal in the components.

Besides PCA, we generate **graph-style figures** for a representative sample: heatmaps of patch-region weights and region-region similarity, degree distributions, top- K connection bars, and a bipartite-style graph layout. These plots make global

connectivity easy to inspect.

Finally, for patch- and region-level illustration on real RGB tiles we overlay **attention weight heatmaps** on the image. We visualize separately tissue-branch and relation-branch maps. The another important visualization is separately visualized attention heads, for better analysis of different features captured.

In summary, PCA answers whether structure in embedding and attention space is separated. Graph figures answer how patches and regions are linked and heatmap overlays answer where the model focus attention within the tissue image.

Accuracy is complementary metric calculated only when weak supervision is enabled. Because it is hard to decide whether the heatmaps capture valid features, we also added classification layer on top of our pretrained embeddings. With this approach we could calculate accuracy of our model trained embeddings. When enough examples exist per class, we also report per-class accuracy and a confusion matrix, because overall accuracy alone can be biased because of morphological similarities between tissue classes. Accuracy shows whether the representation supports the weakly supervised task and whether changes in the architecture or losses preserve ability to classify. Generally the accuracy is calculated by simple equation where all correct predictions are divided with total number of predictions.

11.3 Synthetic Dataset Experiments

In this section, we performed experiments with synthetic data generation models. All experiments can be divided into three parts: the first is autoencoder based reconstruction with the VQ-VAE architecture, the second is synthetic patch generation with a diffusion model and the last one is application of the inpainting

method. Additionally are also provided ablation studies of few components such as seams visibility validation with the help of segmentation and superpixels and comparison of wavy vs straight mask. These experiments were complementary to the qualitative analysis with laiks.

The goal of this part of the experiments is slightly different from similar tasks. Here, we focus on validating the whole pipeline of whole-slide image generation. In the first step, we do not focus on patch quality. Rather, we focus on the pipeline as a whole. The core requirements are artifacts minimalization and ensuring that patch color schemes are as consistent as possible, since our goal is to merge these individual patches.

11.3.1 VQVAE Autoencoder Experiment

The following experiments with the autoencoder were the first experiments performed. The autoencoder architecture is trained on an image reconstruction task, focusing on reconstruction quality. Multiple experiments with fine-tuning of the parameters were conducted, all using the VQVAE architecture. Experiments included differences in the number of network layers, different loss function parameters, and different downsampling and upsampling techniques.

Challenges we encountered occurred, when using publicly available repository as a framework for our approach. It contained multiple problems that needed to be resolved. The first problem was artifacts in the reconstructed image corners. This was resolved by replacing nearest-neighbor interpolation in the upsampling block with bilinear interpolation. In the downsampling block, constant padding was replaced with reflected padding. The second problem was inconsistent shades in the reconstructed images, which is crucial for our further use, where patches need to be connected. This was solved by performing patch normalization in the

range from -1 to 1.

Architecture of the model used for the best training run is VQVAE. This architecture can be divided into five main parts: encoder, quantization convolution, vector quantizer, post-quantization convolution and decoder. Parameters used to configurate this architecture were 4 z channels, representing latent space, 16 384 embeddings representing size of the codebook and 4 embedding dimensions.

From a data point of view for this experiment BRACS dataset was chosen. From whole slide images were extracted 256x256 patches. The dataset was splited into 95 000 training and 5 000 validation patches. All pacthes needed to be normalized in pixel range from -1 to 1.

Exact configuration of the best run, can be seen in complementary files of the thesis. The most important parts of the configuration were 4.5e-06 learning rate, Adam optimizer, 4 batch size, 25 epochs, LambdaLR scheduler. The loss function used in this run was VQLPIPSWithDiscriminator, which consists of four components: pixel loss, perceptual loss, codebook loss and adversarial loss.

Graph visualization of loss functions and metrics can be seen in Fig. 11.7. Validation curves show stable convergence of the VQ-VAE. Reconstruction, commitment, and quantization losses decrease gradually, while both codebook perplexity and cluster usage increase slightly, indicating effective and diverse utilization of discrete latent codes.

Looking at the results , shown in Fig. 11.7, it can be seen that the model is able to reconstruct patches with optimal quality. It is not able to capture the finest details, but this is not the focus of our work. We also calculated quantitative metrics over 10 000 samples, which can be seen in Table. 11.1. These

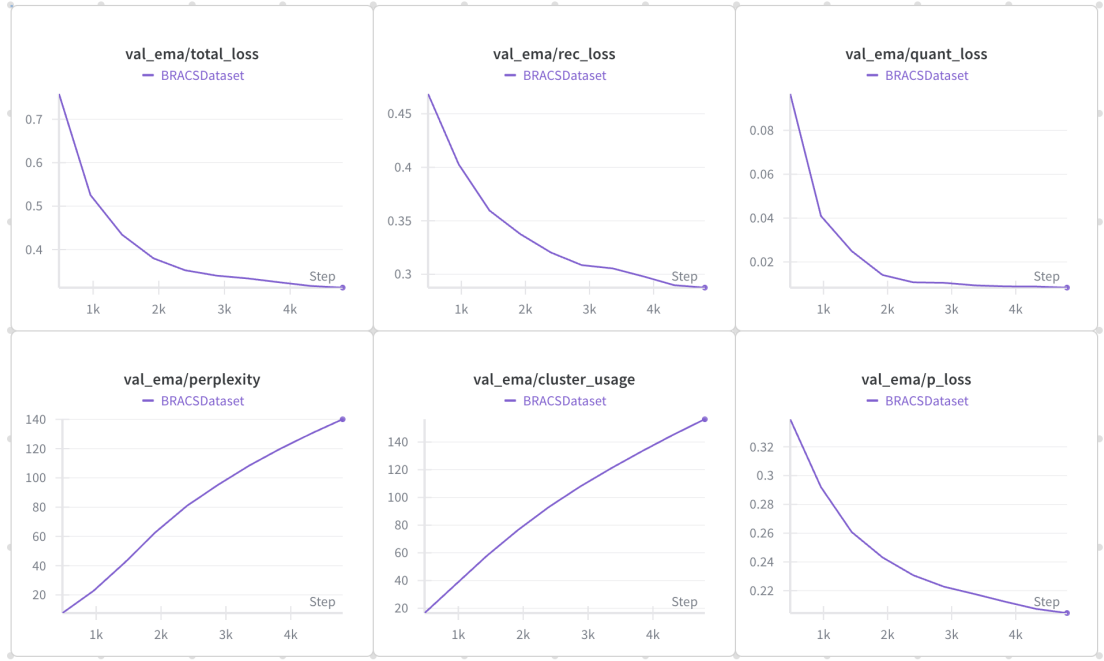


Figure 11.6: Graphs of VQVAE training

results demonstrate the model’s ability to reconstruct original images with good quality.

Architecture	Samples	FID ↓	IS Mean↑	IS Mean↑	Time (min)↓
VQVAE	10 000	13.14	1.33	0.02	5.14

Table 11.1: Comparison of sampling techniques for diffusion model generation

11.3.2 Synthetic Image Generation Experiments

These experiments focused on synthetic sample generation, which is the core step of our synthetic dataset generation pipeline. During training of this denoising diffusion model, multiple fine-tuning experiments were performed with the learning rate and UNET architecture parameters. We also experimented with real down-sampled patch conditioning, which did not perform well and resulted in generating very similar patches of random noise. The best training run is explained in more

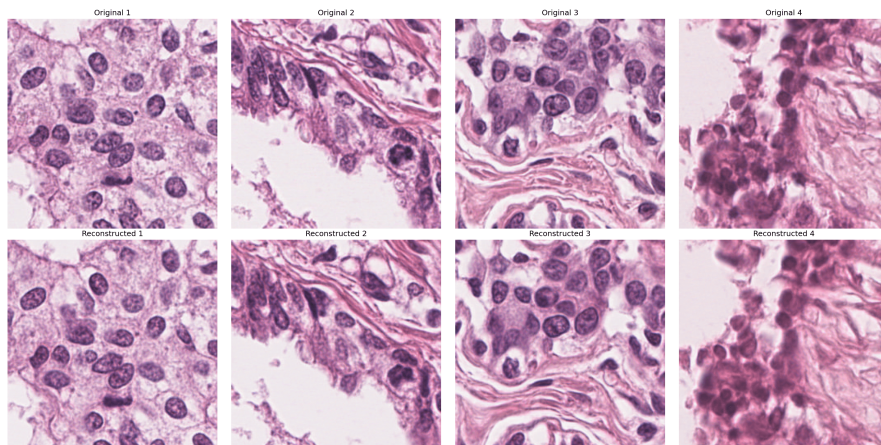


Figure 11.7: Reconstructed image patches examples

detail in the following paragraphs.

Architecture in these experiments is composed of three stages. The first stage is the VQVAE encoder, which is frozen and used only for training to generate latent representations needed for denoiser training. The second stage, the denoiser component, uses the UNET architecture, which goal is to learn to denoise. Once the latent space is denoised, the frozen VQVAE decoder is applied to retrieve generated images in the required size of 256x256 pixels.

Data used in the thesis experiments were also from the BRACS dataset. The same data preparation as in the VQVAE training was applied. We used 95 000 training and 5 000 testing patches of 256x256 pixels. All images were scaled to the range from -1 to 1.

Configuration of the best run can be seen in complementary files of the thesis. The most important parts of the configuration were a 0.0001 learning rate, Adam optimizer, 64 batch size, 15 epochs, and LambdaLR scheduler. The loss function used in this experiment was mean squared error loss.

Results from the generation of a few synthetic patch samples, were able to create multiple variants of the patches, which can be seen in Fig. 11.8.

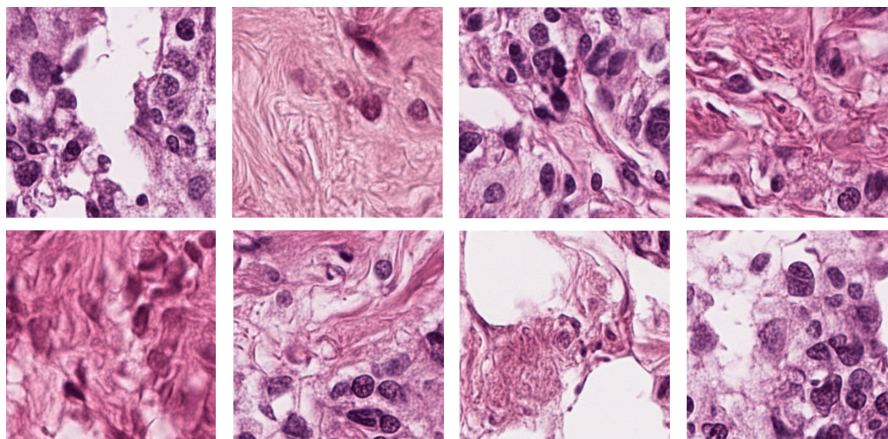


Figure 11.8: Synthetic data patch examples

We also experimented with multiple types of sampling techniques, such as DDPM, DDIM, and PLMS. For all of these techniques, we calculated FID and IS and also measured the sampling time. Sampling was performed on 10 000 synthetic patch images. Results of these quantitative metrics can be seen in Table. 11.2.

Sampling Technique	Steps	FID ↓	IS ↑	Time (min) ↓
DDPM	1000	35.54	1.36 ± 0.02	92.09
DDIM	50	33.37	1.28 ± 0.02	7.07
DDIM	200	36.11	1.23 ± 0.03	20.52
DDIM	500	38.11	1.26 ± 0.03	47.45
PLMS	50	38.65	1.30 ± 0.01	7.17
PLMS	200	39.37	1.25 ± 0.03	20.65
PLMS	500	39.75	1.27 ± 0.02	47.57

Table 11.2: Comparison of sampling techniques for diffusion model generation

11.3.3 Semantic Mask Conditioned Synthetic Image Generation Experiments

This experiment focused on semantic mask conditioned synthetic image generation, where the goal was to generate histology patches that follow predefined rules by semantic classes in the conditioning mask. In this setup, the model was trained to keep structures while also respecting class boundaries and region placement from the mask.

Architecture in these experiments also consisted of three stages. The first stage is a frozen VQ-VAE encoder, from the first experiment, used to map image patches into latent representations. The second stage is a semantic conditioned denoiser, the UNET architecture-based, where the model receives class information from one-hot encoded masks downsampled to latent resolution. The conditioning was implemented with the `adm` conditioning mode and classifier-free guidance. After denoising in latent space, a frozen VQ-VAE decoder reconstructed output images of size 256x256 pixels.

Data used in this experiment were from the BCSS dataset containing image patches and semantic masks. We used the prepared ROI patches and corresponding ROI mask annotations. The training setup followed the dataset split from configuration, which was 41 731 training and 100 validation samples. All images were normalized to the expected input range for latent diffusion training.

Configuration of the best semantic conditioned run can be found in complementary files of the thesis. The important parameters were a 0.0001 learning rate, Adam optimizer, 64 batch size, and LambdaLR scheduler. The diffusion setup used 1 000 timesteps, L2 (MSE) objective, 21 semantic classes, mask latent size

64, and guidance scale 5.0 for conditioned sampling. The UNET backbone used 160 base channels, two residual blocks, and attention at resolutions 16 and 8.

Results showed that the model is able to generate synthetic patches that follow the semantic mask structure. Compared to unconditional generation, this setup provided better control of tissue regions and more interpretable outputs for downstream use. For evaluation we also performed tests with sampling quality and efficiency with FID, PSNR, SSIM and sampling time for different samplers (DDPM, DDIM, and PLMS), which table can be seen in Table. 11.3. Overall results show sufficient quality of results with artefacts, because our goal is not best quality of images as possible the small differences between real and synthetic structures are acceptable. The example of the conditioned synthetic images can be seen in Fig. 11.9.

Sampler	Steps	FID ↓	PSNR (dB) ↑	SSIM ↑	Time (min) ↓
DDPM	1000	161.47	10.10 ± 2.46	0.10 ± 0.05	152.54
DDIM	50	122.84	10.29 ± 3.14	0.13 ± 0.12	9.79
DDIM	200	122.29	10.46 ± 3.04	0.14 ± 0.12	32.41
DDIM	500	121.94	10.51 ± 3.04	0.13 ± 0.12	77.74
PLMS	50	121.65	10.32 ± 2.97	0.13 ± 0.11	9.96
PLMS	200	121.42	10.28 ± 3.04	0.13 ± 0.12	32.59
PLMS	500	120.41	10.53 ± 3.09	0.14 ± 0.12	77.88

Table 11.3: Comparison of sampling techniques for conditioned diffusion model generation over 10 000 patches

11.3.4 Inpainting Grid Mask Experiments

In this experiment, we focused on the seam masks used during the inpainting. We compared straight seam masks and wavy seam masks, while keeping the rest of the pipeline the same. The goal of the experiment was to find out which mask shape makes less visible edges between patches. Looking at our experiments, wavy mask

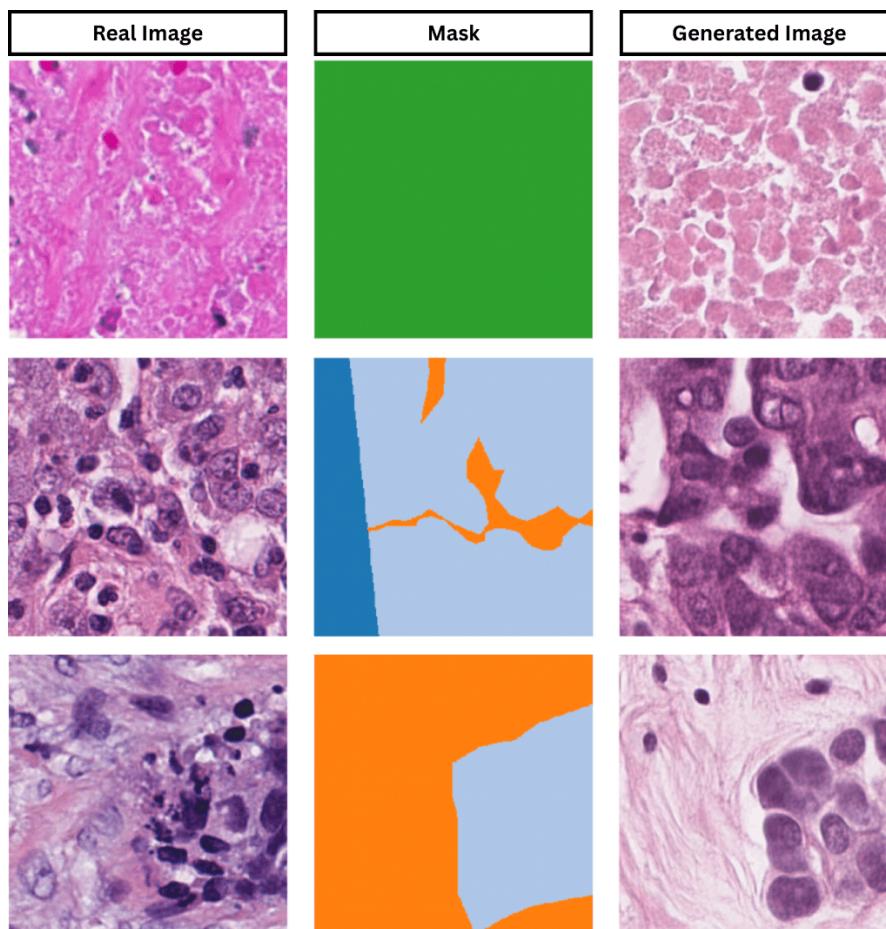


Figure 11.9: Conditioned synthetic data patch examples

worked better in most cases. This is because they break the regular grid pattern and make edges look more natural. Straight masks were sometimes working well with the structures without much textures, but they leave in most of the cases visible line artifact along edges.

11.3.5 Guided Region Generation with Inpainting of 1024x1024 regions

These experiments were focused on generating full 1024x1024 regions from semantic masks. The region was built from 256x256 conditioned patches and then merged with inpainting to reduce border artifacts. These patches were generated in a batches of 16 patches at once. The generated regions followed the target mask structures, with the most edges between patches not visible. In some tissue dense areas, slight seams were still visible, but overall structures were clearly better than with simple naive patch tiling. This confirmed that inpainting-based merging is a useful step for preserving both semantic guidance and image quality. The example of 1024x1024 generated image can be seen in Fig. 11.10.

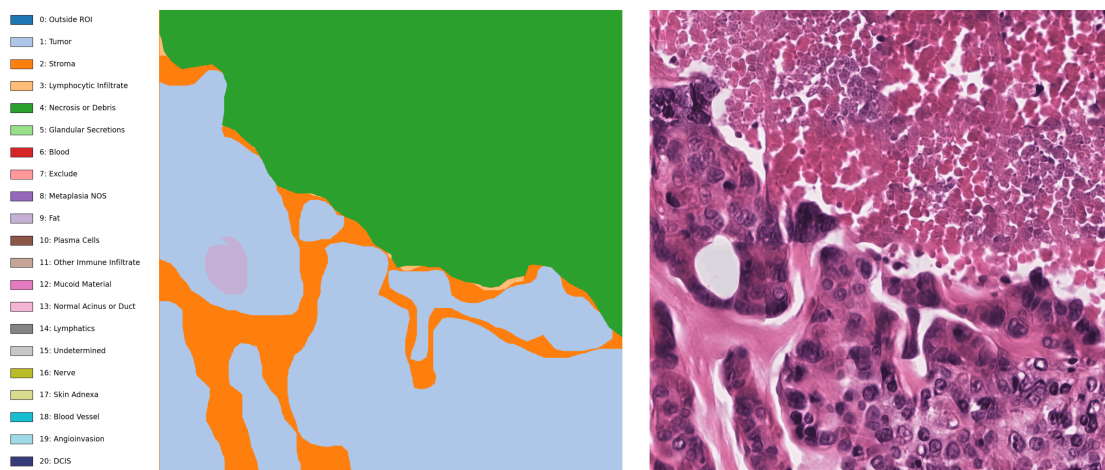


Figure 11.10: Example 1024 x 1024 synthetic region (left: mask legend, center: semantic mask, right: synthetic image)

11.3.6 Guided Region Generation of 4096x4096 regions via downsample-inpaint-upsample

For larger 4096x4096 regions we were not able to use direct full-resolution inpainting because of expensive GPU memory constrains and runtime. So we used a downsample-inpaint-upsample approach. First, full patch content was generated with semantic conditioning, then the grid was downsampled and merged in low resolution, and final upsampling back to 4096x4096. This made the generation possible and kept sufficient global consistency of histological structures. This approach gave a good balance between quality of structures, which we focus on, and computational cost. The approach was also suitable for generating large synthetic regions for further validation experiments. The example of 4096x4096 image can be seen in Fig. 11.11.

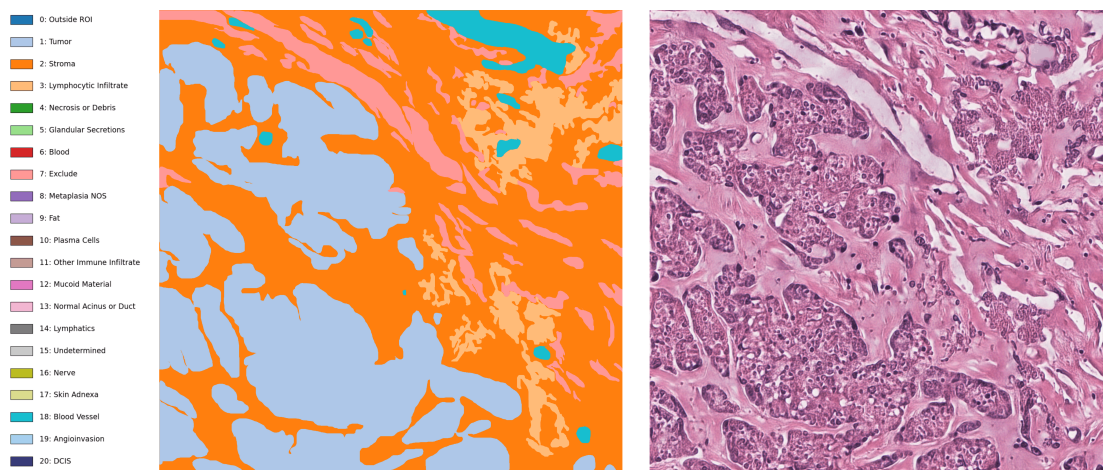


Figure 11.11: Example 4096 x 4096 synthetic region (left: mask legend, center: semantic mask, right: synthetic image)

11.3.7 Guided Whole Slide Image Generation via downsample-inpaint-upsample

After validating the approach on 1024x1024 and 4096x4096 regions, we applied the same strategy on full whole slide image generation such as with 4096x4096 regions. In the downsample-inpaint-upsample manner we samples 4096x4096 regions based on the semantic masks of the whole slide image. After all regions are generated we performed global color normalization over all 4096x4096 regions. This enabled generation of very large images without exceeding memory limits. Visibility of the borders between 4096x4096 regions is not problem in current setup, because we processed in our HIPT layer regions maximally in a size of 4096x4096 regions. This methodology constraint will be targeted in a future research. As shown in Fig. 11.12, the synthetic WSIs preserve the global tissue structure layout given by the masks and maintain local histological textures with no critical detail loss. Minor seam artifacts can still appear in highly complex regions, but they are overall less visible than with direct patch tiling. Overall, the method proved suitable for scalable WSI synthesis with a good trade-off between structural quality and computational efficiency. Even with all of these constraints we can say that the method and results are sufficient for downstream model validation.

11.3.8 Ablation Studies

In the ablation studies, we evaluated how setup and configuration choices in the stitching pipeline affect visibility of patch seams. The main focus was to verify whether seams remain visible by generic segmentation methods. Even when the human eye can not see some seams it can be clearly stated that model can. We compared multiple variants, from naive patch tiling to progressively improved settings with inpainting, wavy seam masks, shared latent noise initialization, and

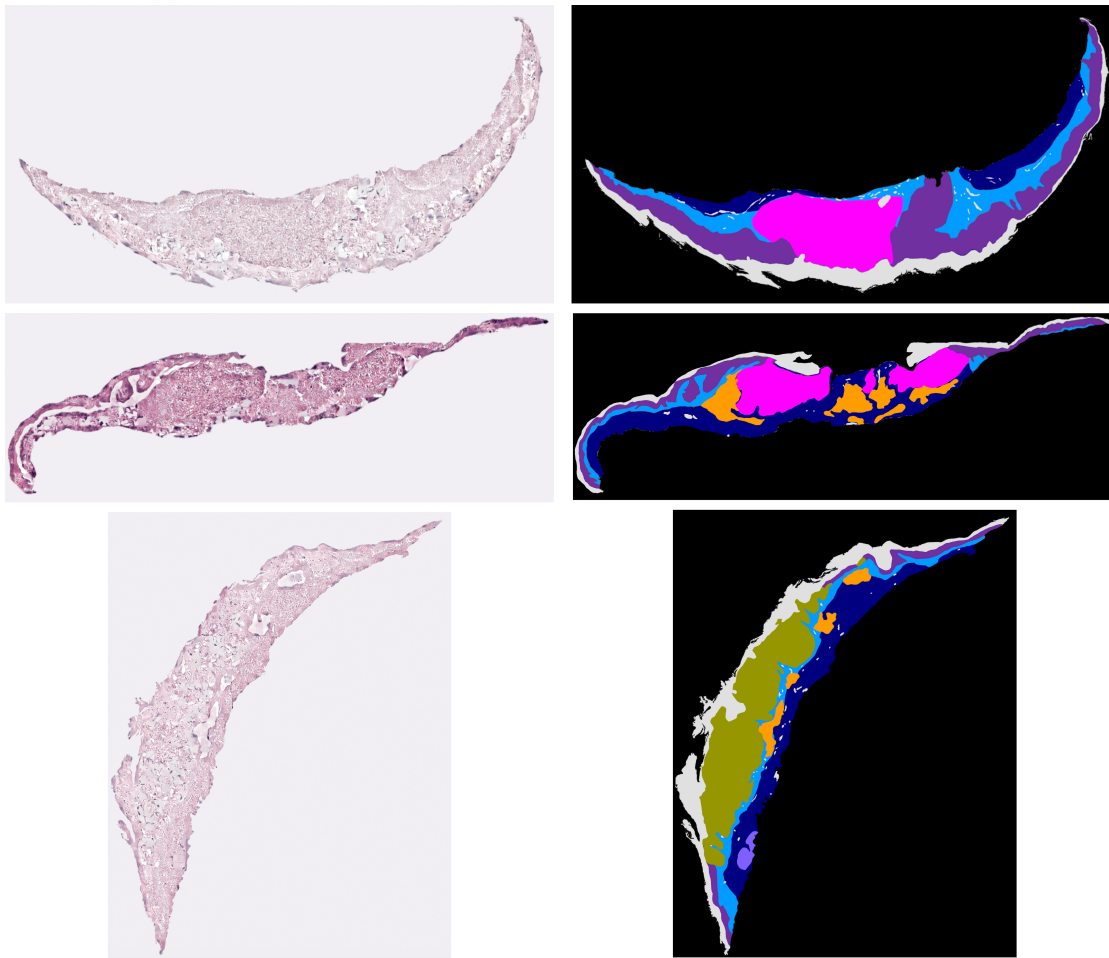


Figure 11.12: Examples of synthetic whole slide images with their semantic masks

color normalization techniques. This gave us a view, which components contribute most to seamless generation. Overall, the strongest setup combined inpainting with wavy seams and color normalization producing the least seam-like artifacts.

Segment Anything (SAM) based Seam Visibility Validation To quantify seam visibility, we used Segment Anything model and overlaid predicted segmentations with the patch grid. The idea was following: if seams are still visible, SAM boundaries often follow grid lines because seams look like object borders. In the naive tiling baseline, this effect appeared, and grid was clearly visible. This case

can be seen in the Fig. 11.13. In the improved configurations, such as inpainting and wavy seams, SAM boundaries matched seam locations more and were less tied to the grid. The lowest seam-related overlap (overlap of segmentation boundaries and the seam) was achieved by the strongest variant, which indicates that its seams are less confusable with true tissue boundaries. The results of these quantitative evaluations of experiments can be seen in the Fig. 11.14.

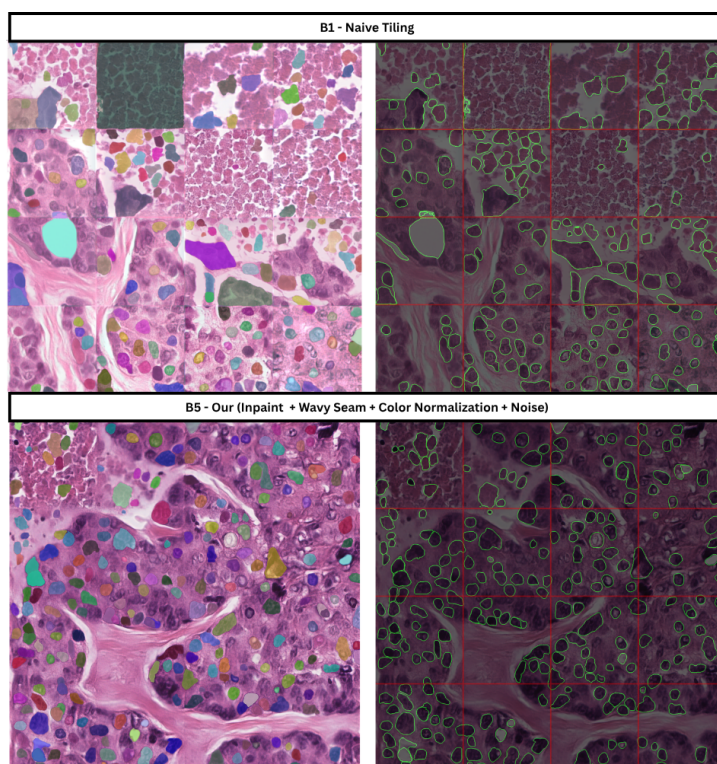


Figure 11.13: SAM seam visibility (left: instance map; right: SAM boundaries + tile grid).

Felzenszwalb based Seam Visibility Validation As a second experiment, we applied Felzenszwalb superpixel segmentation as a lightweight method, which does not need training. This removes bias of training on specific data. The method tends to follow strong local features, so visible seams usually create superpixel borders aligned with patch edges. In the naive tiling case, superpixels matched

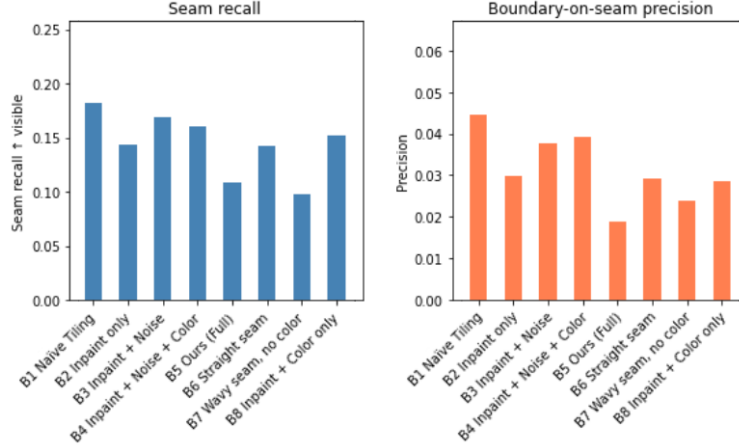


Figure 11.14: SAM seam recall and boundary-on-seam precision (B1–B8).

the grid pattern in some regions, which confirmed strong seam visibility. After applying the full inpainting-based pipeline, superpixel boundaries became more irregular and better followed tissue structure not tile borders. The example of this can be seen in the Fig. 11.15. The quantitative trends, shown in the Fig. 11.16, were consistent with the SAM analysis and again showed clear improvement over naive tiling.

Wavy vs Straight mask comparison We also compared straight and wavy seam masks directly using LPIPS measured in boundary bands around patch borders. The results, shown in the Fig. 11.17, showed that straight seams are better only in very thin bands, but this statement is not true as the band width increases. For medium and wider seam, wavy masks performed better as the gap became larger. This behavior matches the visual analysis of the image, where wavy masks avoid border aligned transitions. Based on both quantitative and qualitative results, wavy seams were selected as the better alternative to straight seams.

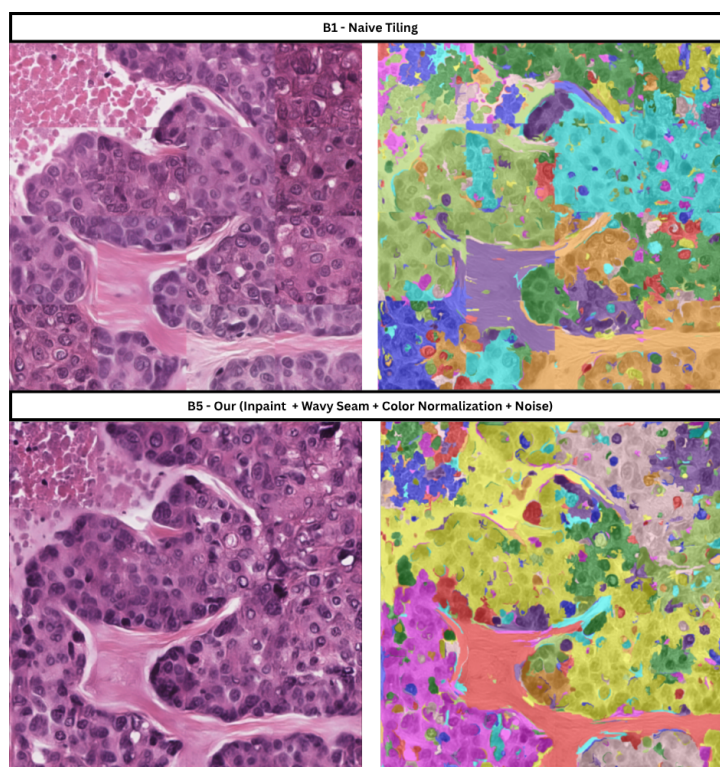


Figure 11.15: Felzenszwalb superpixels for B1 (tiling) vs. B5 (ours).

11.3.9 Qualitative Analysis with Lay Users

In this experiment, a small group of lay users evaluated synthetic histology images using our custom HistoView application. The task, which they performed was the structure search. Participants were asked to localize a tissue structure in a large synthetic image based on a reference patch and then annotate at least one matching region. Overall, most participants completed the task successfully in a few minutes. The reports show that the task was understandable after a short introduction. It was also reported that the reference image was sufficient. Only a few users ended with supported success. In most of the cases participant selected a visually similar structure to the target or hesitated between several very similar areas. In our experiments we performed analysis of only small subset

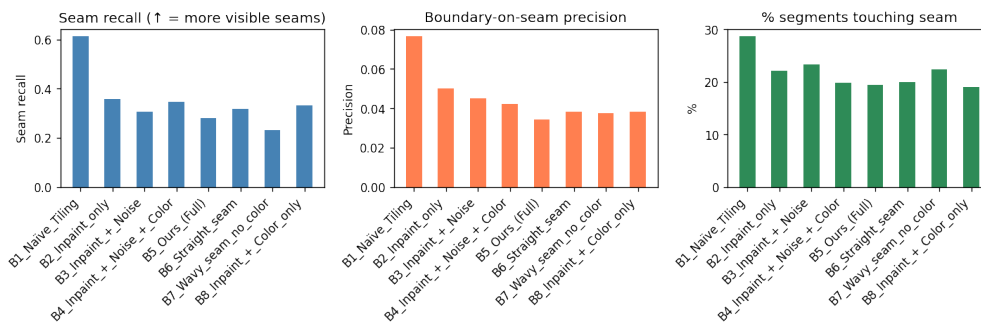


Figure 11.16: Felzenszwalb seam metrics (B1–B8).

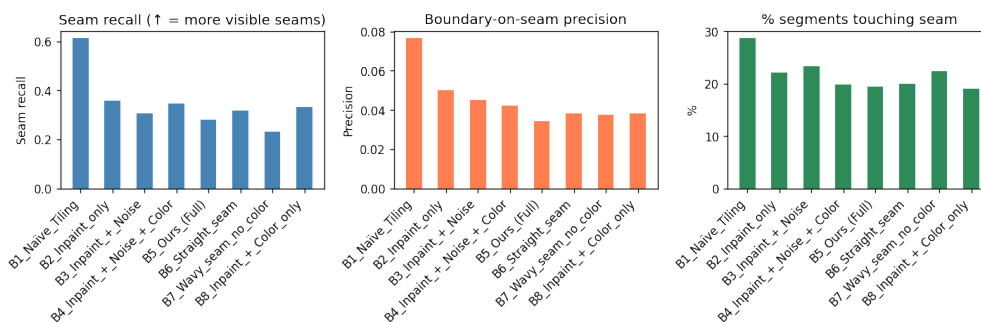


Figure 11.17: LPIPS comparison of wavy vs. straight seam masks

of tissue classes, which contained majority of annotations of the dataset. These classes were: stroma, lymphocytic infiltrate, tumor, blood vessel and fat tissue. For our research we performed this analysis on one pilot participant and 15 other participants.

Differences between Structures were generally visible, with users describing the tissue structures as visually clear enough for the task. Fat tissue regions were often reported as easiest to find, because they are mostly white and correspond the most to background. Structure such as this was in our dataset only one. Tumor clusters were also recognisable when they were bordered with other type of tissue, but some users mentioned that neighbouring regions looked similar, especially at lower magnifications. Stromal regions were perceived as more problematic, because

their boundaries were less sharp and their texture blended more smoothly into surrounding tissue, which led to longer search times and occasional uncertainty. The blood vessels were on the other hand very straight forward, because of their circular shape and hollow center.

User Experience session with the viewer and reference patches confirmed that the current setup is suitable for lay users. But it also revealed several usability issues. The short practice task at the beginning helped participants to understand zoom, pan, and annotation tools, what helped users to understand tool. Many participants said that the reference patch was helpful and gave them a great visual target, but some suggested that a text explanation of typical regions features could help further. Reported difficulties were mostly related to annotation/viewer tool, such as knowing when an annotation was successfully placed, zoom control, rather than problems with the core search task. Overall, the qualitative feedback suggests that the synthetic images and viewer are usable for non-experts, while still exposing meaningful edge cases where structure boundaries and texture similarity may challenge human perception. The more concrete results of the testing can be seen in Fig. 11.18.

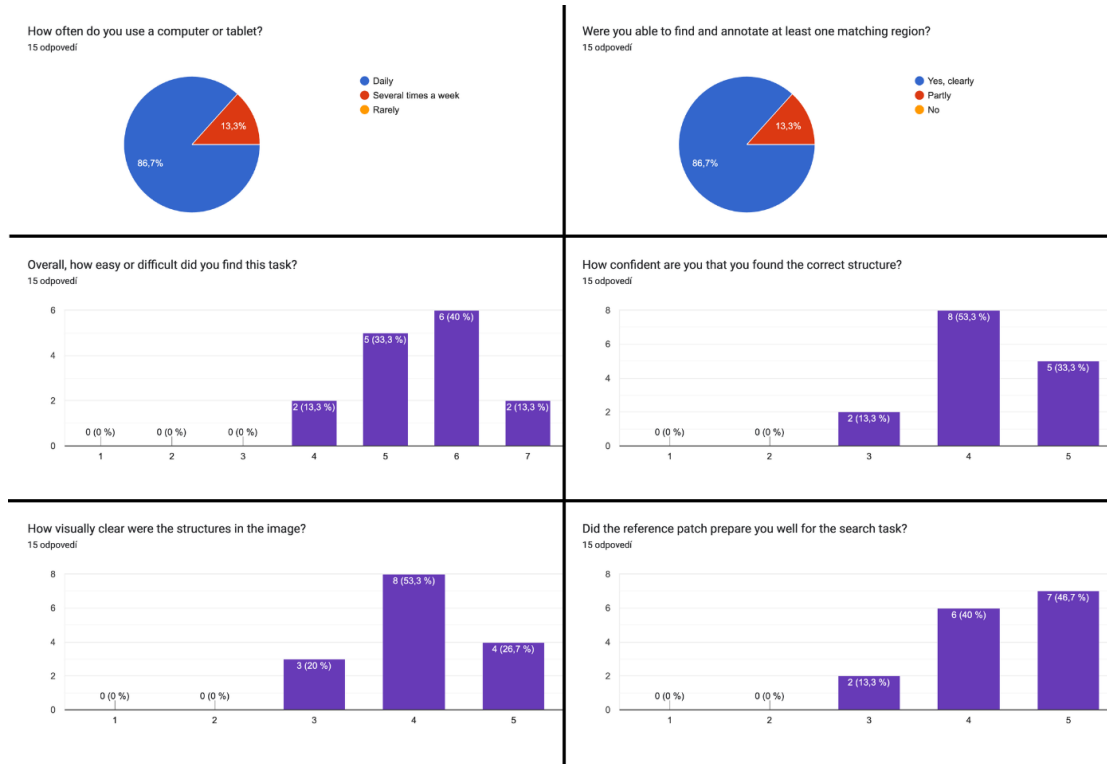


Figure 11.18: Selected Results of the Structure Search Task

11.4 AI-driven Biomarker Detection Preliminary Experiments

In this section of our proposed work, we performed preliminary experiments, which focus on verification if the proposed RA-HIPT architecture can learn useful relations within tissue structures. These relation can be in other words stated as a biomarkers. The main goal is to understand model behaviour, training stability, and interpretability of learned attention weights inside our custom dual-branch attention.

Performed experiments were composed of different architecture sizes, schedulers, self-supervised and weak-supervised modes, two outputs of HIPT architecture em-

beddings (4096x4096 and 256x256), graph mode on/off, and different loss-function configurations.

11.4.1 RA-HIPT Experiments (self-supervised, 4096x4096 regions, BCSS Dataset)

Configuration and hyperparameters used in this supervised RA-HIPT on 4096×4096 region-level inputs baseline, are following. The model used AdamW optimizer (learning rate 1e−5, weight decay 1e−4), CosineAnnealingLR scheduler ($T_{max} = 25$, $\eta_{min} = 1e-6$), batch size 16, and 25 epochs. Cross-attention depth was 12 with 12 heads and embedding dimension 192. The loss was in SSL mode with diversity, consistency, contrastive, reconstruction, curiosity, and Mahalanobis terms, with scheduled increase of relation-attention weight after warmup.

Training curves , in the Fig. 11.19, shows the training process of the experiment in a form of curves. These curves are generated from logs collected during training through training epochs. From this figure, we can see that training is at first very unstable and even looks like it is not learning at all, but after few training steps model is able to stabilize and train based on our proposed loss functions.

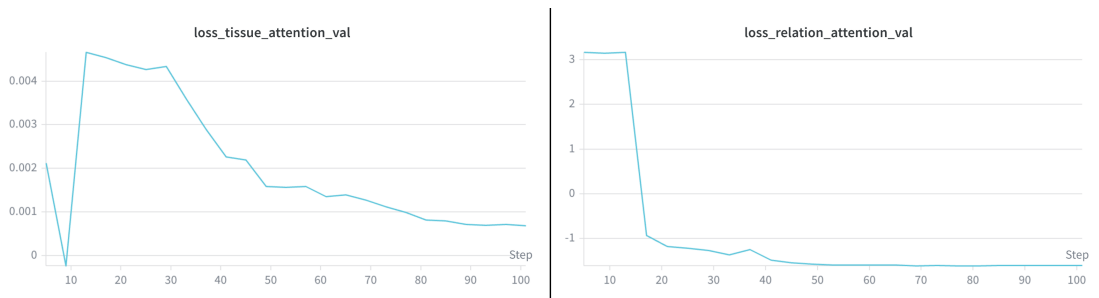


Figure 11.19: Graphs of RA-HIPT training

Patch localizations , in the Fig. 11.20, show the localization of the patch used for the inference within the whole slide image. It depends on the mode of the training but in some cases the region of 4096 is used and in the other cases is used 256 patches.

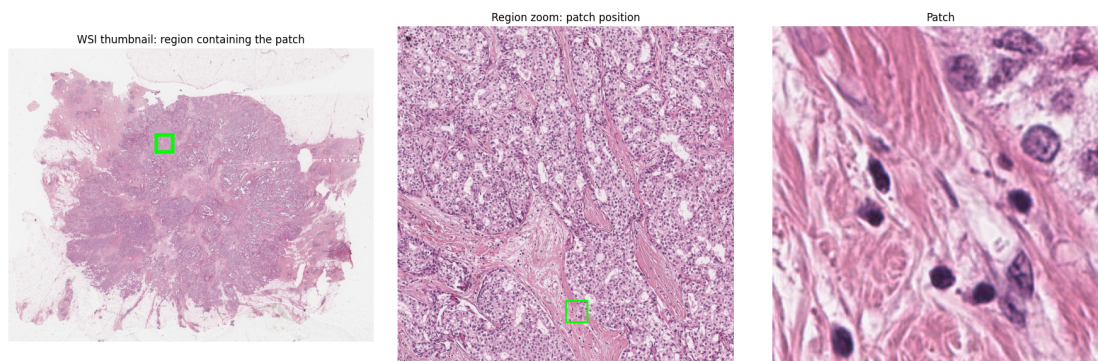


Figure 11.20: Example of patches used for inference of the model

Per-head attention , in the Fig. 11.21 introduces the per-head attention visualization of weights. It is generated by extracting attention maps from each transformer head and displaying them separately. This helps to check whether heads learn complementary patterns or collapse to the same behaviour. From the figure can be seen that the model is able to learn in each head different structures. It can be seen that it distinguish background, tissue and even the fat tissue in the upper part of the image, where seems to be background. It also found other structures but because of the nature of dataset and too little annotated data we were not able to tell without the domain expert if it found any valid structures. The diversity of the heads in the attention was problematic and all heads heatmaps were the same, because of this problem we proposed the diversity loss term. This regularization term ended up to fix this problem and resulted visualization, which can be seen in the figure.

Because the baseline still struggled to enforce direct and much more explainable

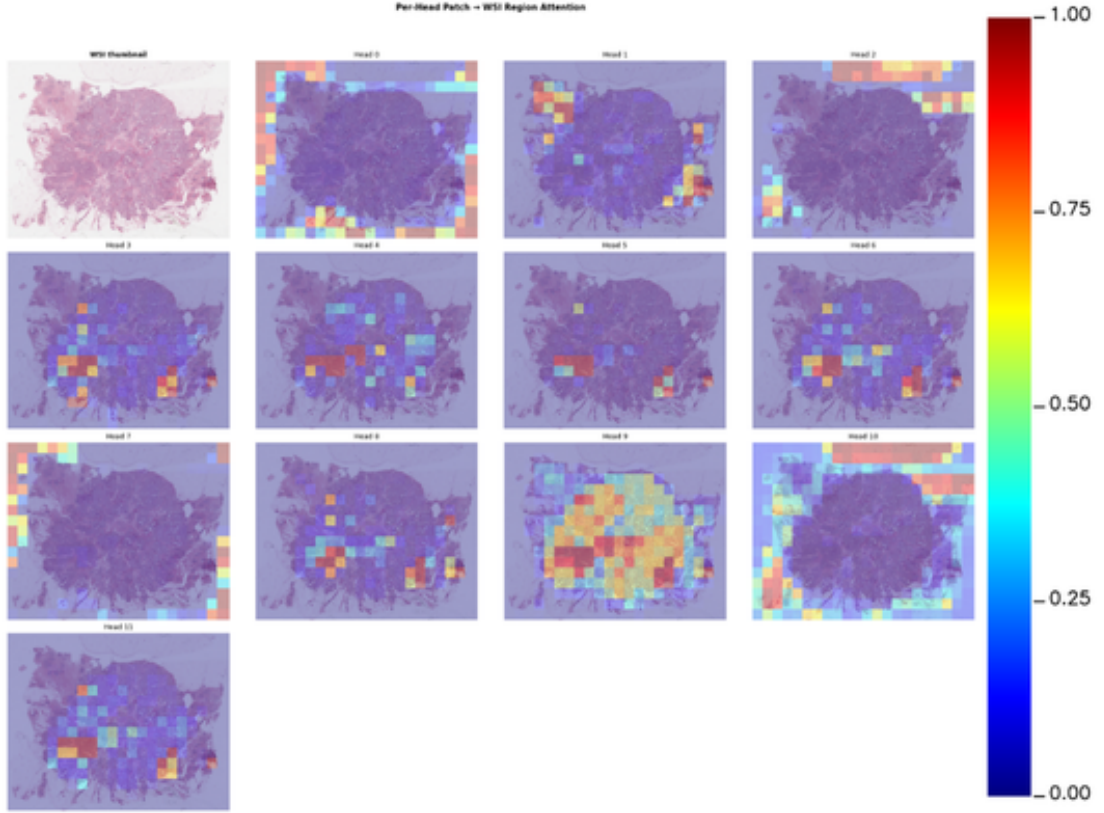


Figure 11.21: Visualization of per head attention module of trained RA-HIPT

long-range dependencies, we extended the model with graph regularization in the next experiment.

11.4.2 Graph RA-HIPT Experiments (self-supervised, 4096x4096 regions, BCSS Dataset)

In configuration and hyperparameters were enabled graph mode and used a graph-aware setup for region relation learning. The run used AdamW (learning rate $1e-5$, weight decay $1e-4$), CosineAnnealingLR, batch size 16, and 15 epochs. The cross-attention module used depth 6, 6 heads, and embedding dimension 384. Important graph terms included diversity, sparsity, smoothness, prediction, coher-

ence, and region-graph constraints, with part of these terms activated progressively during training with scheduled loss weights.

Attention heatmap , in the Fig. 11.22 introduces the attention heatmap visualization. It is computed by aggregating attention weights over attention heads and visualizing it over the histology image. This figure shows where the model concentrates relation-aware attention and gives fast qualitative feedback about whether global structure is being captured. It can be seen that the model focused on specific parts of the image, mostly within the tissue area. But in this case we are still not able to tell if the model captured relevant features, because of its complexity of representation for non-domain user. This problem is further resolved in the next experiment by using of synthetic dataset and weak-supervision.

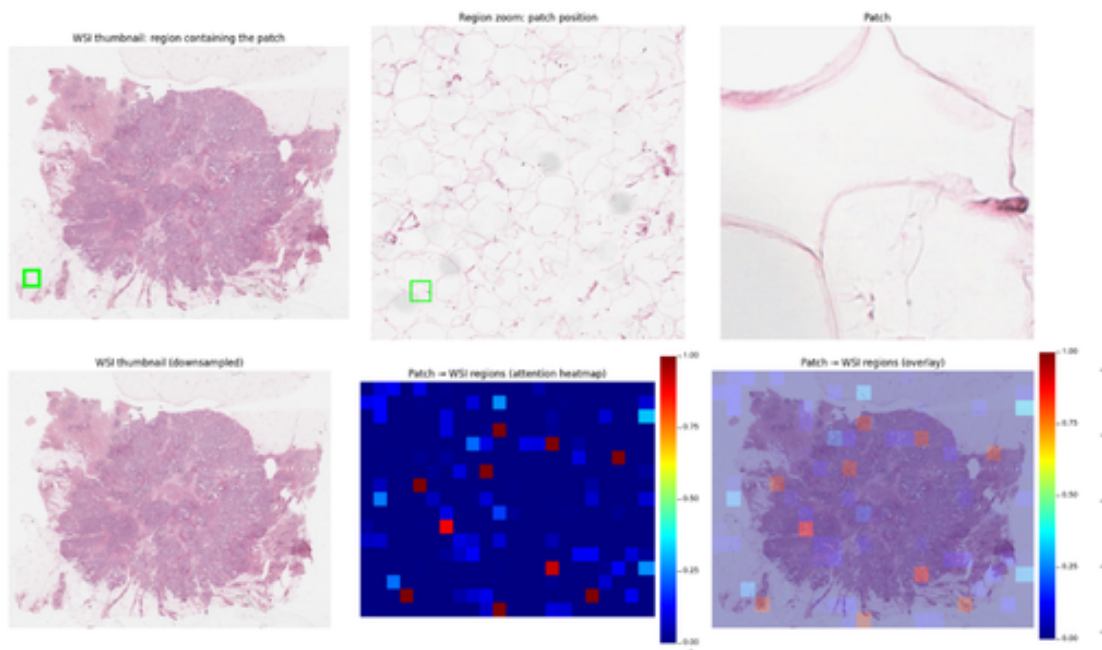


Figure 11.22: Visualization of heatmap made of attention weights from Graph RA-HIPT

Region-to-region relation graphs , in the Fig. 11.23, introduces region-to-region relation between the individual regions. These are derived from the regularization graph by converting region similarities/edges into a visual connectivity map. The figure suggests that graph mode increases explicit long-range relational structure compared to the non-graph baseline. The problem is with relation edges quantity, too few regions relate to each other, what makes the question, why does this happen and how can be this resolved. This is also aimed in the following experiment.

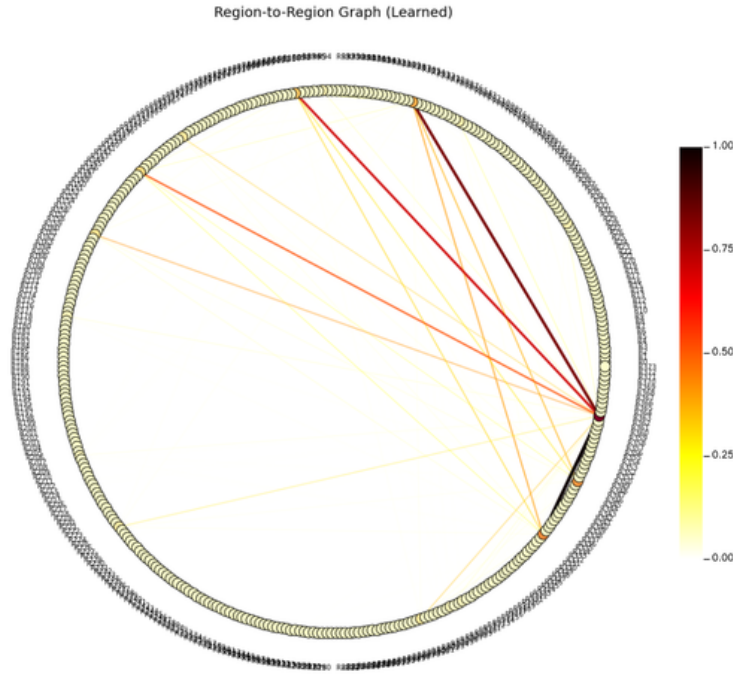


Figure 11.23: Visualization region to region relations derived from the regularization graph structure

Collapse diagnostics , in the Fig. 11.24 introduces analysis of the embedding collapse through the training. These plots are computed from embedding statistics during training and are used to detect latent-space and representation collapse. We observed phases where collapse risk increased. After multiple experiments we

performed optimization with hyperparameters tuning to resolve problem of the collapse, which results can be seen in the plots. These plots were used as an main part and core background analysis. Because of them we proposed into our architecture loss function regularization terms such as embedding variance loss, head diversity loss and edge sparsity loss.

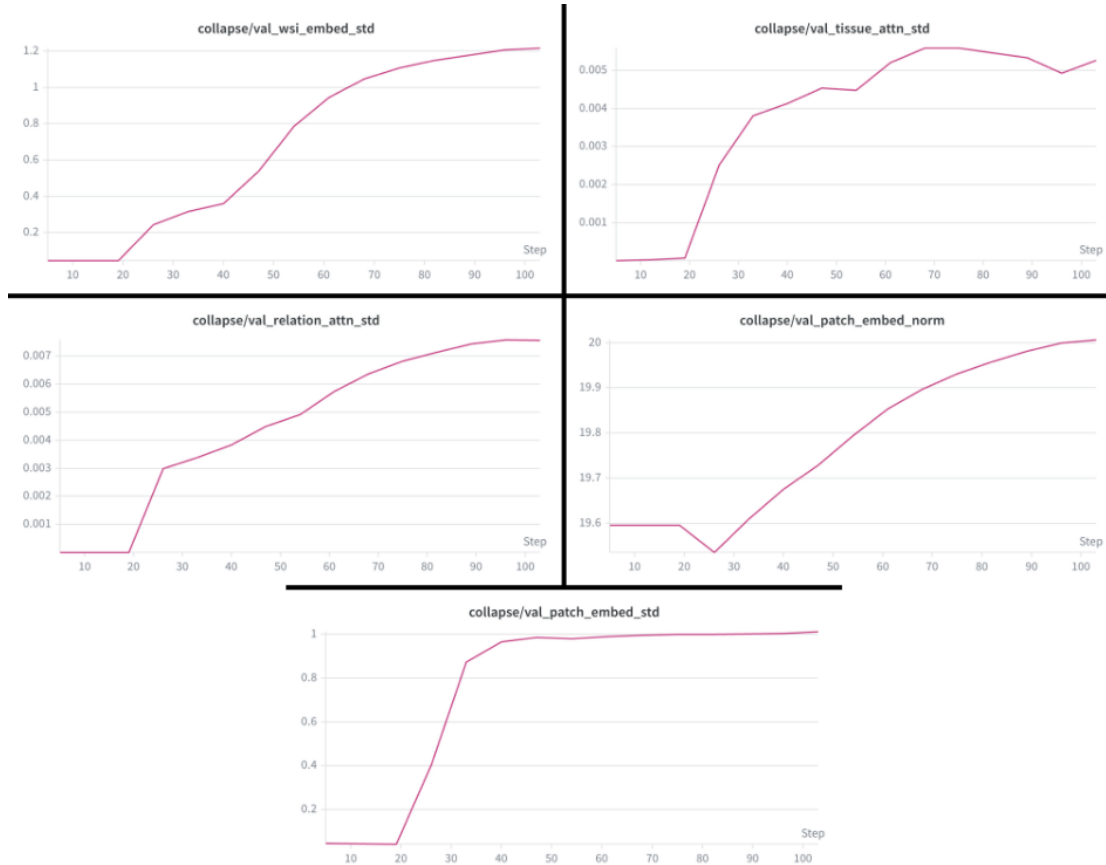


Figure 11.24: Analysis of embeddings collapse through the training

Principle component analysis , in the Fig. 11.25 introduces the PCA embedding view together with graph degree distribution. PCA is computed by projecting high-dimensional region embeddings into 2D for cluster inspection, while degree distribution summarizes how strongly nodes are connected in the learned relation

graph. The figure shows partial structuring, but still insufficient separation and signs of unstable representation geometry in this self-supervised setup.

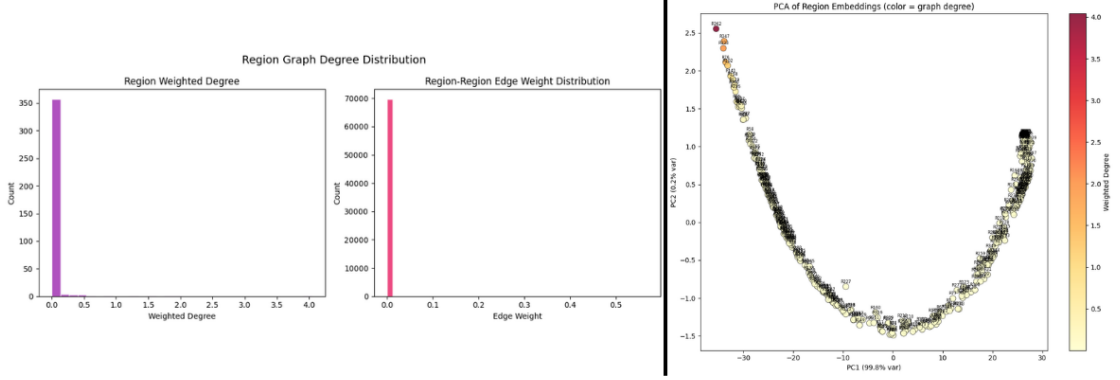


Figure 11.25: Visualization of PCA of region embeddings and Region graph degree distribution

11.4.3 Graph RA-HIPT Experiments (weak-supervised, 256x256 patches, Controlled Synthetic Dataset)

Configuration and hyperparameters of this experiment were switched to weak supervision and 256x256 patch-level WSI embeddings. The dataset used in this experiment was a controlled synthetic dataset. The core idea was to use a dataset with interpretable structure and a stronger focus on relation training. We used AdamW (learning rate $5e-5$, weight decay $1e-4$), CosineAnnealingLR, batch size 16, and 15 epochs. The cross-attention block was smaller in depth (2), with 8 heads and embedding dimension 512. For regularization and the training process, weak supervision and graph regularization were enabled, using diversity/sparsity and embedding variance regularization subterms for the core loss function.

Attention heatmap, in Fig. 11.26, uses the same heatmap visualization as in the previous experiment. In this experiment, we wanted to focus more on testing whole slide images, which were fully interpreted with their corresponding

semantic masks used for generation. The first row represents epoch 0, which reflects the model’s ability to detect multiple different structures, even though the image was synthetically generated and the pretrained HIPT model was trained on real histological images. The second row represents our trained proposed Graph RA-HIPT. In these heatmaps, it can be seen that the model focused on specific regions. These regions are inside the tissue, which is one of the evidences that these heatmaps might refer to real relations. The attention regions also accumulate in one area. However, any definitive statement from these heatmaps can be made only after consultation with a domain expert. The third row represents disentanglement of the embedding into three different views. This also shows that each view detects a different part of the tissue. From these heatmaps alone, no clear conclusions can be stated. Therefore, we performed other visualizations to support these findings.

Semantic mask , in Fig. 11.27, introduces the control parameter in the form of a mask used in synthetic data generation. This mask encodes predefined tissue structures and spatial arrangement. The mask also has a second role as a validation reference, used to check whether the same or related structures are visualized in the heatmap. This allows more reliable validation whether the model learns the expected relations or not.

Region-to-region relation graphs , in Fig. 11.28, shows the corresponding learned region-to-region graph. In comparison with the relation graph in Fig. 11.23, big differences can be seen. The graph from the experiment with weak-supervised mode captures many more relations between the regions. This graph shows an interesting result: multiple darkest edges are connected through the regions, and most of these regions are located close to each other. That means this graph, visualizing the relations detected by our graph-structured regularization, captures

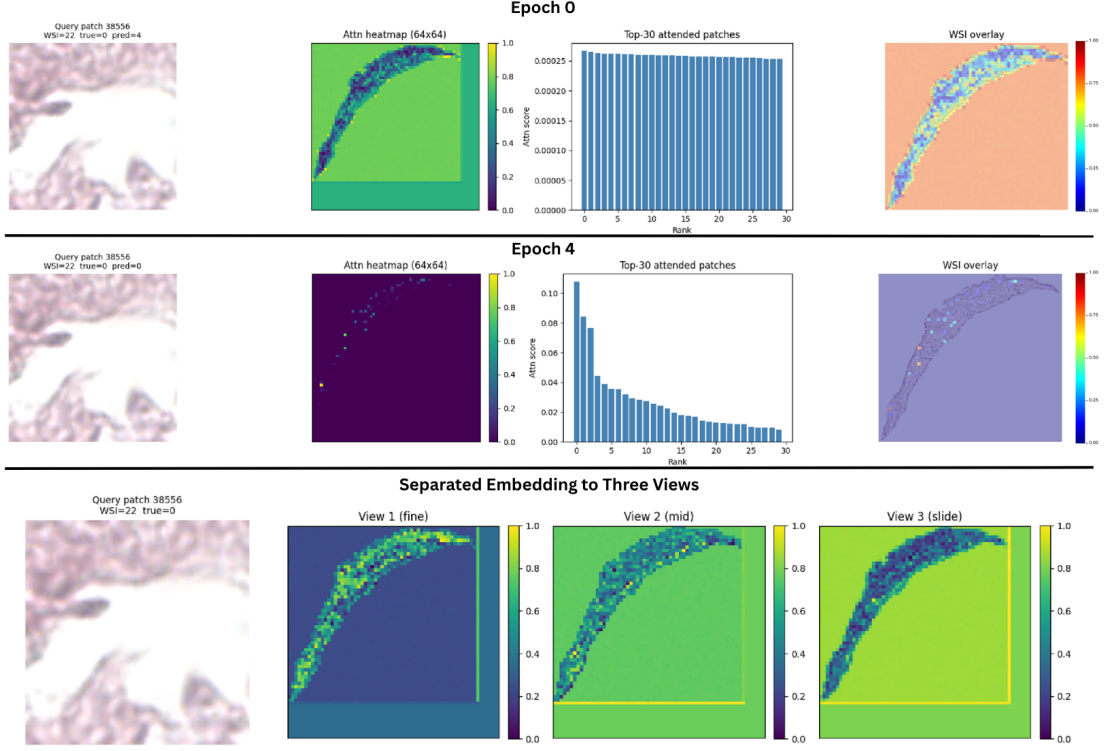


Figure 11.26: Visualization of heatmap made of attention weights from Graph RA-HIPT on Controlled synth dataset

another relevant feature. This feature is the relation between neighbouring regions and larger areas. This confirms that the model captures meaningful synthetic long-distance relations.

Confusion matrix , in Fig. 11.29, introduces the weak-supervised classification analysis. It is computed from predicted vs. target classes, with the help of a classification layer plugged on top of our model output embeddings. The confusion matrix clearly shows which classes the model struggles with the most and which classes it is most confident in predicting. This confusion matrix is a complementary part of the previous visualizations, providing information about the model's ability to classify based on these output embeddings. The results show that the model struggles with some classes such as necroses and debris, blood, and metaplasia.

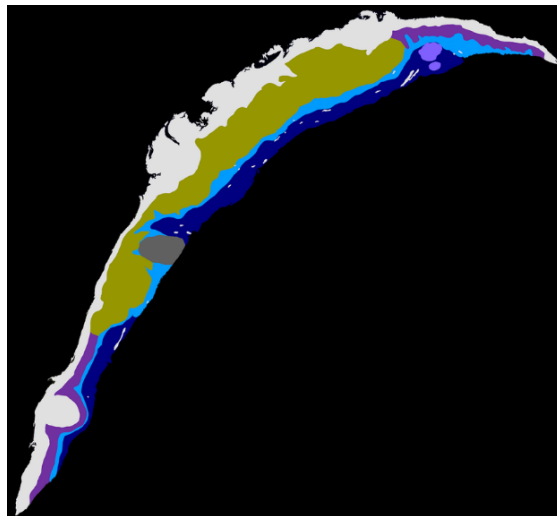


Figure 11.27: Semantic mask used for generation of wsi

These results were caused by the extremely non-balanced dataset. However, this problem with accuracy was not needed to be analyzed further, because we only wanted to analyze the model’s ability to classify, which we validated and confirmed based on the results.

From the analysis of all these experimental results, it can be stated that the model detects both local and global relations of whole slide image regions with the help of our proposed graph-structured regularization. However, one question remains, which can be answered only by a domain expert, and that is the correctness and reasonability of these highlighted relations. This creates a clear path for our future experiments and validation.

11.4.4 Graph RA-HIPT Experiments (weak-supervised, 256x256 patches, BRACS Dataset)

Configuration and hyperparameters were in a final experiment the same as in a weakly supervised patch-level setup for BRACS. The run kept AdamW

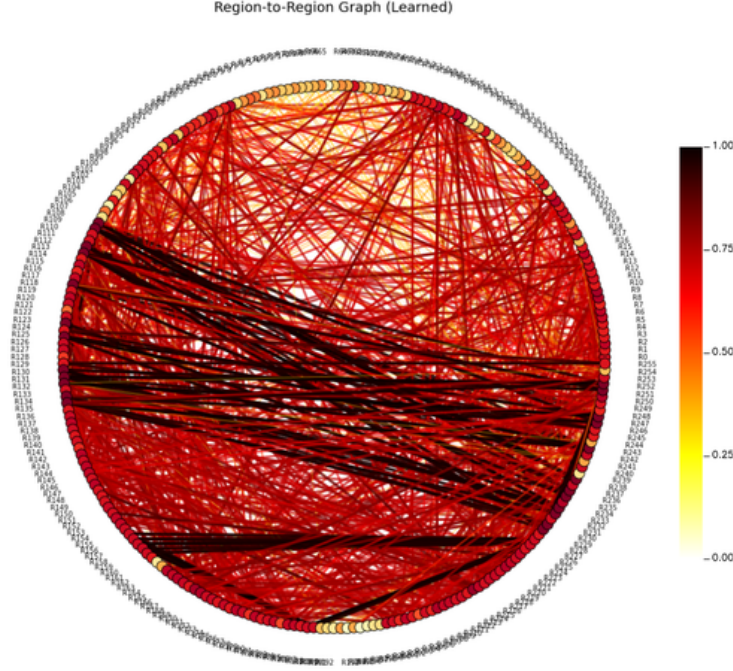


Figure 11.28: Visualization region to region relations derived from the regularization graph structure

(learning rate $5e-5$, weight decay $1e-4$), CosineAnnealingLR, batch size 16, and 15 epochs, with cross-attention depth 2, 8 heads, and embedding dimension 512. Graph mode remained active with the same family of diversity/sparsity/variance controls, and weak supervision was applied with BRACS labels ($n_{classes} = 7$).

Attention heatmap , Fig. 11.30 shows attention heatmaps overlaid over the real BRACS WSI. The maps are less clean than in the synthetic data because of its much higher resolution. Looking at epoch 0, we can see that the model is able to detect relevant areas, where tissue is present and absent. Looking at the heatmaps after training for 4 epochs, it can be seen in the zoomed heatmap that the model is able to detect individual areas. This confirms that the model is also able to work on real-world data and is able to highlight biological structures.

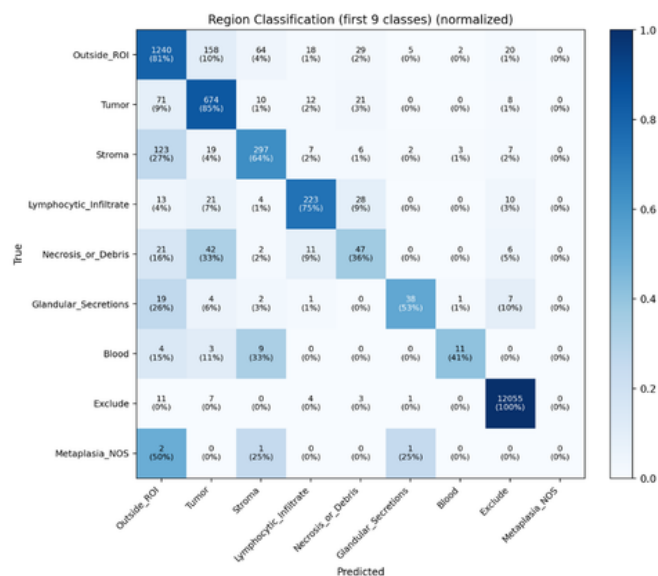


Figure 11.29: Confusion matrix of classification over synthetic dataset

Principal component analysis , Fig. 11.31 shows improved results comparable to the experiment from Sec. 11.4.2. Looking at the differences between PCA and degree distribution, we can see great improvements in a terms of the collapse of latent space. In this experiment, the latent space did not collapse into the curve, and the degree of the region weights stayed consistent and was not zeroed. The embedding projection shows partial class structuring with expected overlap for difficult histopathology patterns. This confirms the healthiness of our regularization graph structure.

Region-to-region relation graphs , Fig. 11.32 shows results similar to the experiment in Sec. 11.4.2 with the synthetic dataset, where some regions are more related to other regions. This graph confirms that graph regularization still learns interpretable region-level relations under real data conditions.

Class relation visualization graphs , Fig. 11.32 shows the most interesting results. From this graph, it can be found that our approach, in a weakly supervised

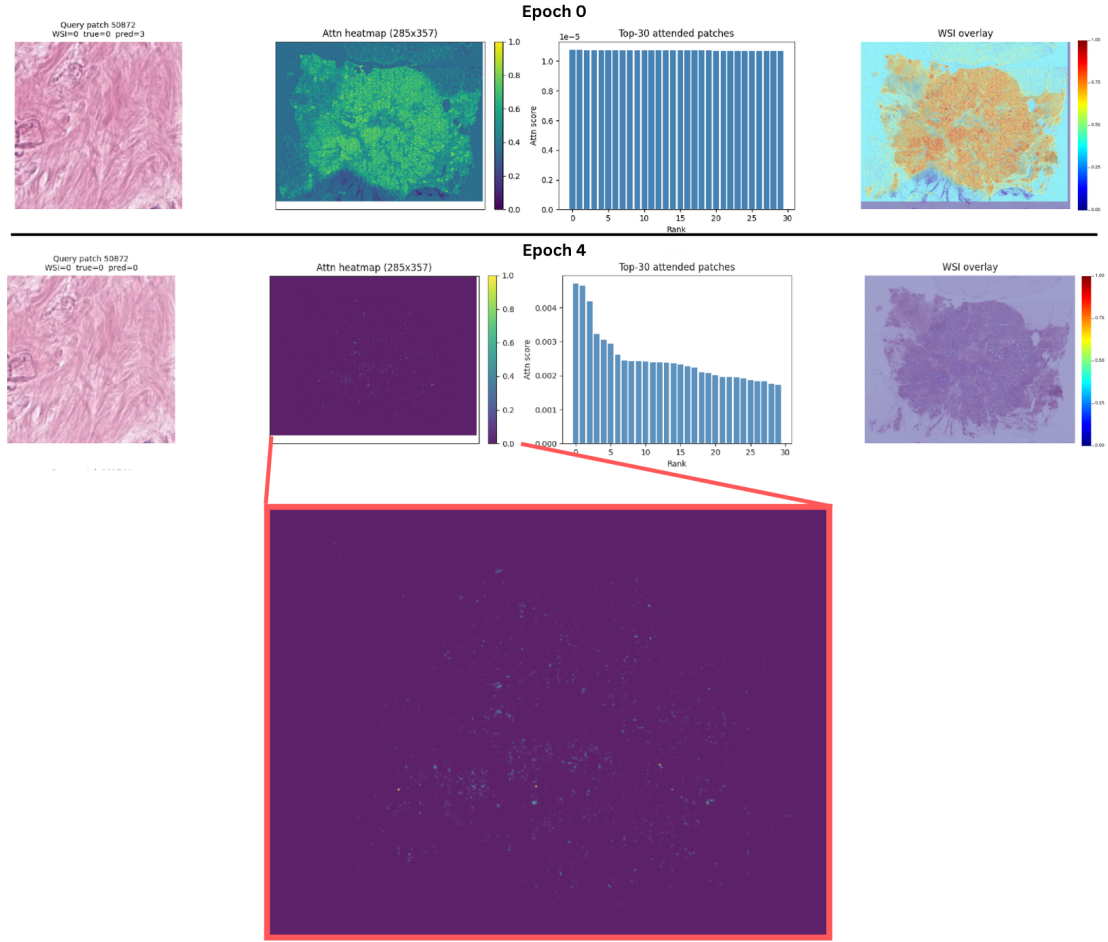


Figure 11.30: Visualization of heatmap made of attention weights from Graph RA-HIPT on BRACS dataset WSI

setting, was able to learn relations between whole classes. This graph is made directly from the relation branch attention weights, which are produced by the cross-attention between the query patch and the whole-slide embeddings. This combination of embeddings provided direct integration of local and global relations. These attention weights are influenced during training by our graph regularization structure, so the results of this graph represent the impact of our two proposed methods.

Looking at the medical perspective of these classes and their relations, the

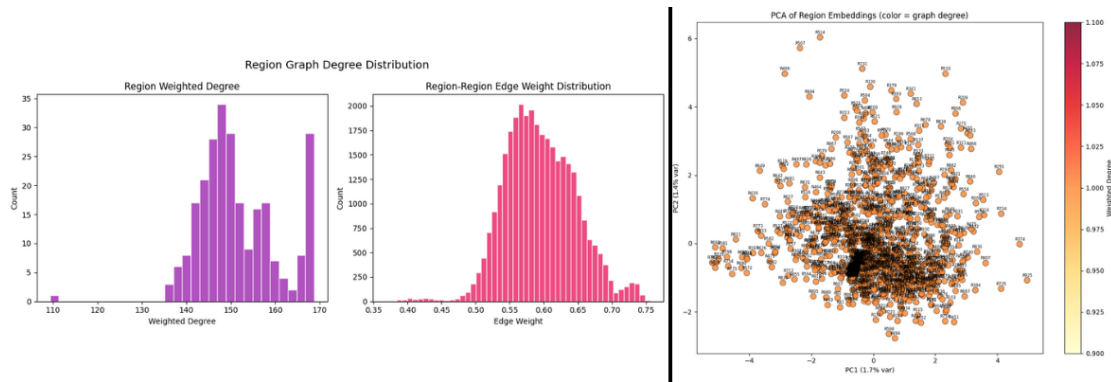


Figure 11.31: Visualization of PCA of region embeddings and Region graph degree distribution

following can be stated based on medical literature [49, 29, 66]:

1. **Isolation of blood vessels class** comes from a real biological concept, because blood vessel regions in histological slides are structurally different and architecturally separated from the epithelial-inflammatory microenvironment. Another reason is that they carry no compositional relationship with glandular, lymphocytic, or metaplastic tissue. The model has learned, without being told, that blood vessel patches are not relevant to other classes.
2. **Glandular-Metaplasia connection** shows the most dominant relation. This is also biologically correct, because glandular cells transform over time into metaplasia. The direct progression is chronic inflammation, then glandular atrophy, then metaplasia, then dysplasia, and the final stage is cancer. The model is able to show the highest-weight relation between these classes, which validates that the model is able to capture histologically meaningful and diagnostically relevant structure.
3. **Lymphocytic class** connects both glandular and metaplasia. This is another relevant correlation because these lymphocytic infiltrations co-occur with both classes in histological tissue. These lymphocytes produce cytokines

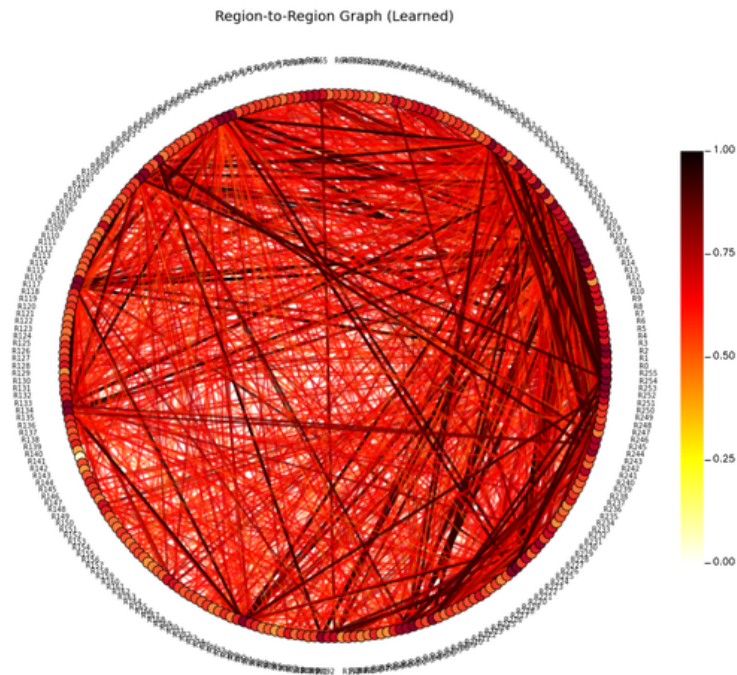


Figure 11.32: Visualization region to region relations derived from the regularization graph structure

and reactive species that drive the transition from glandular to metaplastic epithelium. This is another strong sign that the model is able to capture immune-epithelial crosstalk dynamics, which are another biologically relevant relation.

In the overall summary of the experiments, we can state a few results. The model is able to capture local and global relations, the regularization graph structure contains valid information between regions, a few selected regions have higher relations with each other, and the model is able to find relevant biological structures and concepts. The final missing part is direct validation with a domain expert, which will be done as a following step in a future work.

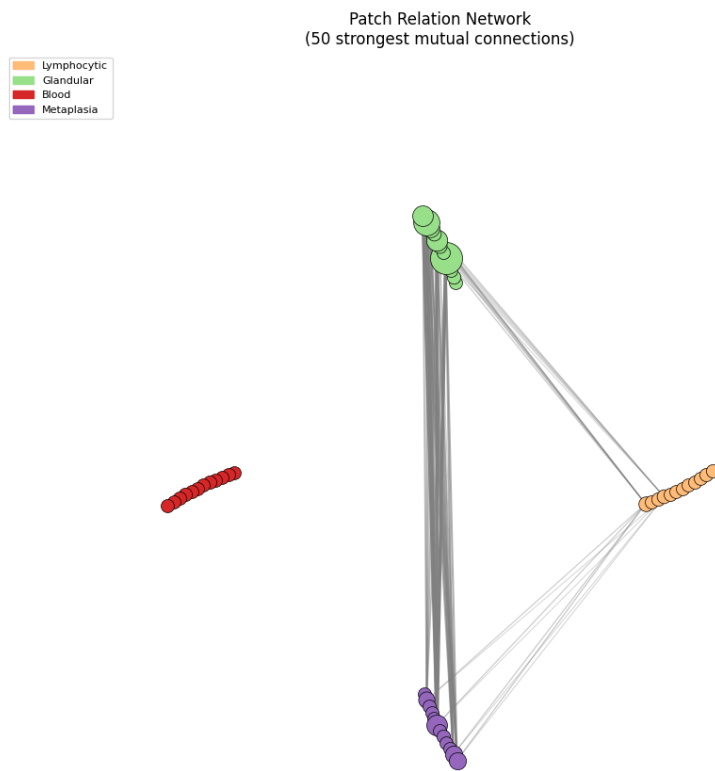


Figure 11.33: Visualization of class relations trained by the Graph RA-HIPT architecture

Chapter 12

Viewer tool for qualitative analysis (HistoView)

For our experiments with domain experts and participants we needed to create validation tool. This part is important, because all methods and results, which we proposed are closely bounded with the quality of the results. In a terms of controlled synthetic generation, we needed to validate if the generated stuctures are distinguishable between each other. On the other hand we need to also validate heatmaps generated by RA-HIPT architecture. This architecture is used to detect potential biomarkers, but are nearly imposible to validate without the pathologist. To our knowledge we did not find any tool, which can specifically suit our requirements. Because of this we proposed our custom tool suitable for two use cases: qualitative analysis of synthetic structures with laik users and qualitative analysis of the predicted heatmap biomarkers. Repository of this tool is available online [54] at <https://github.com/0Kozlik0/Heatmap.Viewer>.

12.1 Technology Stack

From the technology point of view the backend of the application is implemented with the use of FastAPI and Python. The frontend part of the application is implemented as a native HTML with CSS and JavaScript. As a high resolution image viewer was used OpenSeaDragon library [76], because of complexity of this feature we did not implement it by ourselves. Through the all experiments we run application locally without need of deployment at this stage of research.

12.2 Functionalities of the Application

The application contain only the core functionalities, which is crucial for specific use cases of our research.

Features:

- Whole-slide viewing with pan/zoom using Deep Zoom tiling
- Slide upload and loading from UI
- Support for pathology formats: .svs, .tif, .tiff, .ndpi, .mrxs, .vms, .scn, .bif and png.
- Heatmap overlay upload and display on top of the slide
- Optional patch image upload linked to a heatmap
- Patch preview panel with metadata
- Annotation tools
- Annotation management (show/hide, rename, delete, focus)
- Save and Export annotations

12.3 Use Cases

Our custom tool is designed to support fast and intuitive preview of histopathology images. This one tool combines whole-slide navigation, heatmap overlays, patch preview, and annotation features, which makes it useful for human-in-the-loop evaluation tasks. Fig. 12.1 shows the interface used during the annotation process, including the slide viewport, analysis controls, and annotation panel.

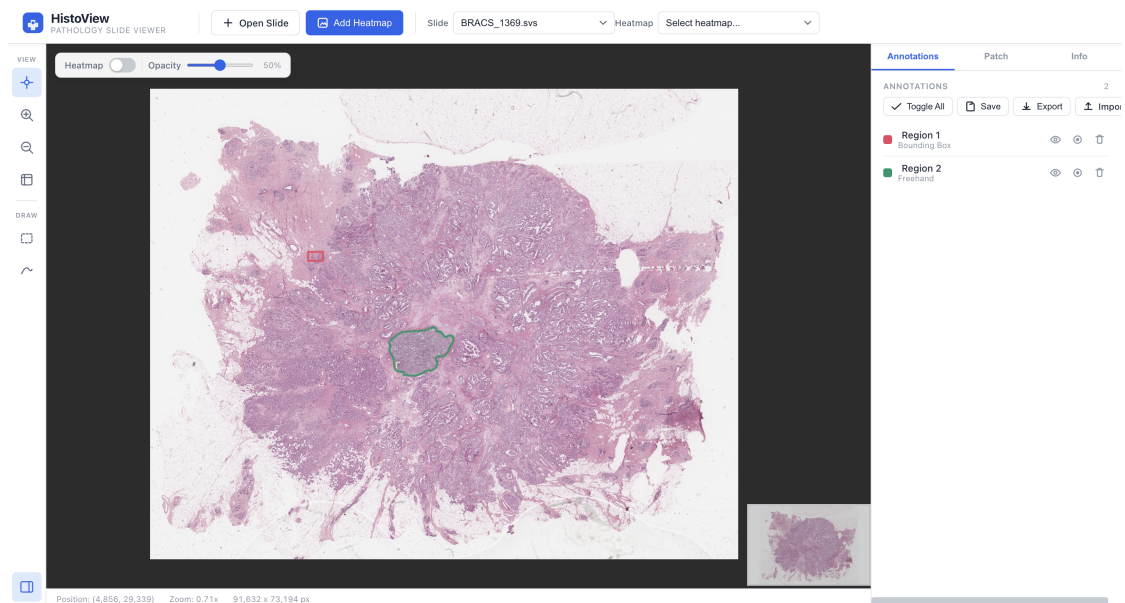


Figure 12.1: Preview of HistoView tool

12.3.1 Qualitative Analysis of Synthetic Structures with Laik Users

In this use case we used our tool in a case that laik users was assigned to assess synthetic histology-like structures based on reference images. The image of this case can be seen in Fig. 12.2. Users can zoom into generated regions, compare structures across multiple scales, and annotate areas similar to reference image. The annotation workflow helps collect structured feedback about the visible differ-

ences of synthetic structures, which is important metric for controlled validation of AI model.

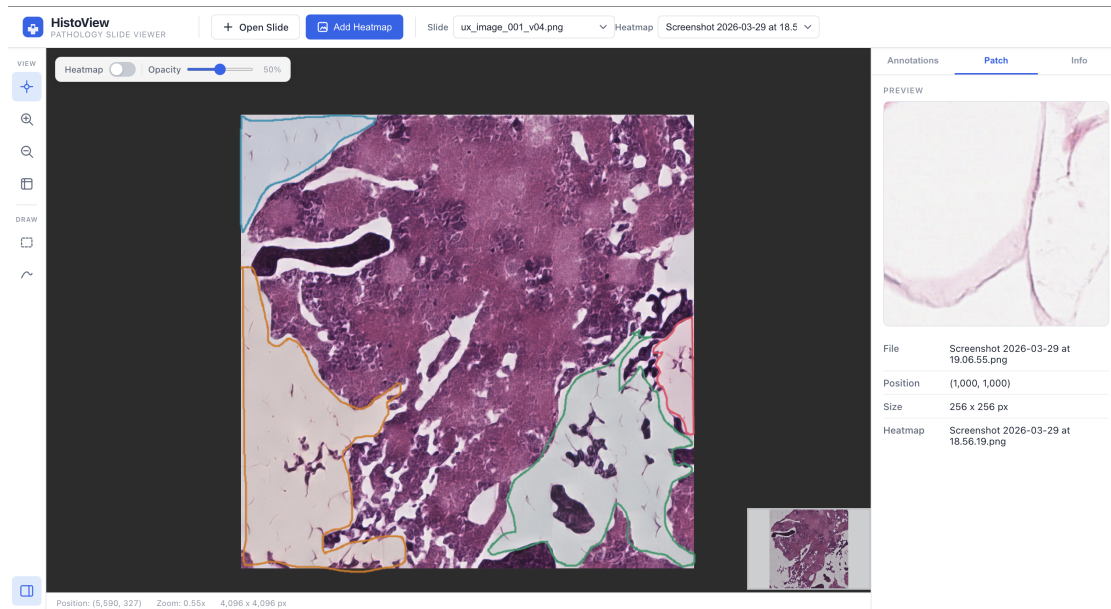


Figure 12.2: Preview for synthetic structure analysis use case

12.3.2 Qualitative Analysis of the Predicted Heatmap Biomarkers

In this second use case, the tool focus on validation of predicted biomarker heatmaps with our RA-HIPT architecture. These heatmaps are overlaid over the original slide. Histopathologist can controll overlay visibility and opacity to examine whether regions align with morphological meaningful tissue structures. Optional patch localization and manual annotations make it easier to provide useful qualitative evidence for model behavior and potential failure cases. The interface of this use case can be seen in Fig. 12.3.

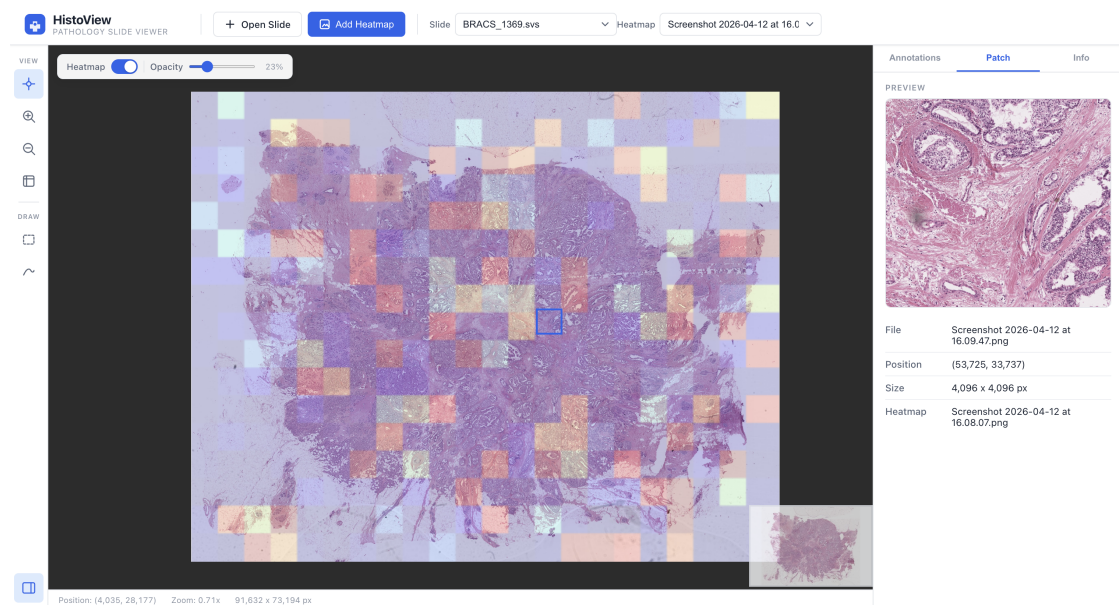


Figure 12.3: Preview for biomarker detection with heatmap analysis use case

Chapter 13

Conclusion

This thesis focused on improving AI-driven biomarker detection in histopathology by addressing three problems: missing integration of local and global relations, missing explainability of model predictions, and small number of annotated images, what makes model validation more complex. Because of these limitations, we proposed a complete pipeline composed of a global-local relation-aware prediction model and a controlled synthetic validation dataset.

13.1 Summary

From the architectural point of view, we designed the Relation-Aware Hierarchical Image Pyramid Transformer, which use combination of local features through patch embeddings with global features from whole slide embeddings. These embeddings are combined through the custom cross-attention mechanism. In addition, we proposed graph-structured regularization and custom loss components that support relational learning, attention head diversity, and overall stability. The model was evaluated in different training configurations, such as self-supervised

and weak-supervised. Another validation of the model was with both real and synthetic datasets, made with our second controlled synthetic validation dataset generator.

As a second part of our work, we designed and implemented mentioned controlled synthetic image generation pipeline based on VQ-VAE autoencoder, denoising diffusion models, semantic conditioning and inpainting-based stitching. This pipeline provides generation of large synthetic regions and whole-slide images with pre-defined semantic structure. For evaluation part we performed quantitative evaluation with metrics such as FID, IS, PSNR, SSIM, and sampling time. Complementary part to these metrics were ablation studies for seam visibility and qualitative evaluation with the 15 participants. These experiments confirmed that the generated data are suitable for downstream validation tasks.

Experimental analysis of RA-HIPT showed that the model is able to capture local-global relations and produce interpretable attention heatmaps. In weak-supervised experiments with controlled synthetic data and BRACS data, we performed visualizations of region to region relation and class-relation. Results of these visualizations provide insights that the model is able to detect real biological structures and capture long distance relations. On the other hand we also calculated classification accuracy and visualized confusion matrix, which confirms model ability to recognize differences between tissue classes. These findings support the hypothesis of this work: explicit integration of local and global features improves relational modeling in whole-slide images and provides a stronger basis for interpretable biomarker analysis.

Looking at the hypotheses formulated in the motivation, the outputs can be summarized as follows:

- **(RQ1)**: the proposed RA-HIPT architecture with cross-attention and relation-

aware regularization improved representation of local-global relations, and also confirmed stronger relational consistency and meaningful region-level interactions.

- **(RQ2)**: the controlled synthetic pipeline produced diagnostically structured data with competitive quantitative quality, enabling reproducible, stress-testing and validation when real annotated data are limited.
- **(RQ3)**: the explainability output and visualizations from our experiments such as attention maps, relation graphs, and class-level relation visualizations, which were directly derived from local-global relation integrations, provided biologically relevant interpretations that improves interpretability of model decisions.

Overall, we believe that our solution brings promising results with the implementation of local-global relations into the model architecture, with the focus on the interpretability and great foundations for controlled synthetic validation dataset framework. These parts creates a great foundation for the potencial biomarker detection and validation of the new AI model architectures, which are hard to validate because of their complexity.

13.2 Future Work

Future work can focus on three main directions. First, the current relation visualizations and heatmaps should be validated directly by domain experts to confirm relevance of highlighted regions and discovered class relations in a histology images. This step is important for final confirmation of model behavior.

Second, relevant direction is to extend the synthetic pipeline toward larger datasets with better tissue variability and quality, preserving the details and mod-

ifying the lossy compression parts such as inpainting and up/down sampling. Another steps includes further optimization of inpainting for very large resolutions and perform evaluation across additional downstream tasks other than the current biomarker detection model. The global goal of this part is to create framework for easy and fast controlled validation of models, potentially controlled with the suitable interface.

Third, the RA-HIPT framework can be further optimized. It can be focused on the model architecture and training process. Improvements can include additional work on regularization terms and stronger histology data focused supervision strategies. The long-term goal is to build an interpretable relation-aware system that can be transferred from controlled validation to real diagnostic support scenarios able to find potencial new biomarkers and structure relations.

Resumé

Úvod

Analýza histologických snímok je základom pri diagnostike a výskume biomarkerov. Digitálne snímky (WSI) kombinujú extrémne rozlíšenie s informáciou na viacerých škálach od bunkových detailov až po usporiadanie tkaniva na úrovni celej snímky. Manuálna expertíza patológom zostáva časovo náročná pri čom automatizované metódy môžu skrátiť prezeranie. Problém je že tieto metódy často pracujú len s malými výrezmi alebo neposkytujú dostatočnú vysvetliteľnosť. V klinickom kontexte je preto dôležité uvažovať nielen o presnosti na jednotlivých výrezoch, ale aj o vzťahoch medzi vzdialenými oblasťami snímky a o tom, či sa ich model učí alebo ich berie do úvahy.

Táto práca spája dva hlavné koncepty. Prvým je architektúra *Relation-Aware Hierarchical Image Pyramid Transformer* (RA-HIPT), ktorá kombinuje globálne embeddingy z hierarchického pyramidového transformera (HIPT) s lokálnymi embeddingami patchov 256×256 cez dvoj vetvový modul pozornosti (vetva tkaniva a vetva vzťahov) a voliteľnou grafovou regularizáciou medzi regiónmi. Druhým konceptom je *kontrolovaný syntetický validačný dataset*: VQ-VAE na získanie reprezentácií v latentnom priestore, latentná difúzia s podmienením sémantickou maskou, inpaintingové skladanie regiónov a farebná normalizácia. Takto vzniknuté syntetické

dáta umožňujú overiť hypotézy o integrácii lokálno globálnych vzťahov v prostredí, kde je plne známe cez sémantickú masku využitú pri generovaní snímok.

Pri reálnych klinických dátach sú často dostupné iba slabé anotácie na úrovni menších oblastí alebo celej snímky. Celková pixelová anotácia je časovo náročná pre patológa, preto si model musí pomôcť so samo alebo slabo učiacimi technikami pri trénovaní AI modelu. Vysvetliteľnosť výstupov, napríklad vizualizácia súvislosti medzi oblasťami, je dôležitou podmienkou dôveryhodnej asistencie patológovi v klinickom prostredí.

Prehľad súvisiacich oblastí

Digitálna histopatológia využíva klasifikáciu, segmentáciu, detekciu a generovanie syntetických snímok. Pri snímkoch je typické rozdelovať snímku na výrezy a následne agregovať informáciu počas tréovania. Hlboké siete začínajú od konvolučných architektúr až po vision transformery s mechanizmom pozornosti umožňujú učenie hierarchických a globálnych závislostí medzi regiónmi. Pre histologické snímky bol navrhnutý hierarchický prístup typu pyramidového transformera (HIPT), ktorý explicitne pracuje s viacerými úrovňami rozlíšenia [22]. Na tejto architektúre je postavená spracovanie globálnej informácie v tejto práci.

Generatívne modely pre histológiu zahŕňajú kvantizačné autoenkóbery (VQ-VAE) na vytváranie diskretného latentného priestoru vhodného na difúziu v menšej dimenzii. Tieto difúzne modely a ich smplovacie techniky (DDPM, DDIM, PLMS) iteratívne odšumujú latentný priestor. Inpaintingové mechanizmy na doplnenie neznámych oblastí medzi susednými výrezmi tak, aby spoje nemali viditeľné hranice. Podmienenie sémantickou maskou umožňuje riadiť rozloženie tried v snímke. Pri smplovaní syntetickej snímky sa často používa classifier-free guidance na zosilnenie vplyvu masky oproti bezpodmienečnej predikcii šumu.

Redukcia dimenzie a projekcie (napr. PCA) sa používajú na vizualizáciu rozloženia embeddingov pre jednotlivé regióny a na odhalenie kolapsu alebo zhlukovania reprezentácií regiónov pri počas tréningu tréningu. Pri interpretovateľnosti sa rozlišujú post-hoc metódy (napr. vizualizácie významnosti založené na gradientoch alebo propagácii váh) a architektúry, ktoré priamo vytvárajú mapy pozornosti alebo graf väzieb medzi triedami. V medicíne musí byť interpretácia konzistentná a nesmie skrývať neistotu modelu.

Pre úplnosť, čisto patchový prístup bez globálneho kontextu zostáva často výpočtovo najjednoduchší, avšak pri diagnostike založenej na vzťahoch medzi oblasťami môže nerozpoznať práve tie vzťahy, ktoré sú najdôležitejšie a na ktoré sa chceme v tejto práci zamerať. Preto má zmysel kombinovať hierarchické spracovanie histologických snímok s modulom pozornosti s využitím embeddingov a grafom medzi regiónmi.

Súvisiace smery výskumu

V digitálnej patológii má dlhodobu miesto viacero inštančné a slabo vedené učenie. Anotácia je často dostupná len na úrovni celej snímky alebo väčšej oblasti, zatiaľ čo modely pracujú so skupinami výrezov, taktiež aj nazývané (*bags*). Novšie prístupy skúmajú učenie črt alebo prototypov nad takýmito skupinami výrezov a ich prepojenie na slabo vedené učenie pri klasifikácii [7, 94]. S tým súvisí aj snaha zlepšiť štruktúru latentného priestoru tak, aby boli zhľuky črt lepšie využiteľné pri klasifikácii alebo interpretácii [36].

Na strane generovania sa latentné difúzne modely presadzujú aj pri vyšších rozlíšeníach obrazu. Existujú príbuzné práce, ktoré rozširujú difúziu do latentného priestoru naprieč škálami a umožňujú kvalitnejšie generovanie veľkých štruktúr [118]. Syntetické alebo čiastočne riadené generovanie sa tiež používa ako nástroj

na doplnenie dátových sád alebo na kontrolované experimenty, kde je potrebná vysvetliteľná štruktúra výstupu [79]. Naša práca sa radí na priesečník týchto trendov: namiesto čisto modelu čiernej skrinky a klasifikátora vedeného slabým učeníím navrhuje architektúru, ktorá explicitne modeluje vzťahy medzi regiónmi (cross-attention, graf) a zároveň používa syntetické snímky z kontrolovaného syntetického datasetu na overenie, či sa tieto väzby správajú na základe biologických morfológií.

Z perspektívy výskumného problému je dôležité, že literatúra potvrdzuje, že latentný priestor zostáva hlavným miestom, kde sa stretávajú koncepty čo model rozlišuje a generatívne ciele (ako vyzerajú konzistentné vzorky). Pre digitálnu histológiu to znamená, že kombinácia latentného priestoru (VQ-VAE), podmienenej difúzie a následnej validácie modelom typu RA-HIPT je širší koncept hľadania kontroly nad štruktúrou generovaného obrazu a aj nad vnútornými vzťahmi medzi časťami histologických snímok.

Ciele práce a výskumné otázky

Cieľom je navrhnúť a vyhodnotiť riešenie, ktoré (i) explicitne modeluje lokálno-globálne vzťahy na histologických snímkach, (ii) produkuje interpretovateľné mapy a grafy pre diskusiu s odborníkmi a (iii) umožňuje kontrolovanú validáciu cez syntetické WSI s pred definovateľnou štruktúrou. Práca formuluje tri otázky: **RQ1** - či architektúra s modulom pozornosti a grafom dokáže zachytiť zmysluplné väzby medzi regiónmi, **RQ2** - do akej miery kontrolovane vytvárané syntetické WSI podľa známej rozlohy štruktúr umožňujú overiť hypotézy AI modelov, **RQ3** - či sú vizualizácie natoľko zrozumiteľné, aby slúžili ako podklad pre expertnú validáciu.

Praktická časť je členená nasledovne: popis a štatistiky použitých dát, definícia

metriek pre generovanie a pre analýzu reprezentácií, rozsiahla časť o stavbe syntetického datasetu a meraní kvality (vq-vae, difúzia, podmienené generovanie, spájanie výrezov, abláčné štúdie, používateľské testovanie) a samostatná časť o RA-HIPT v režimoch učenia bez učiteľa a slabšie podmieneného tréningu na BCSS, syntetických WSI a BRACS. K tréningom s najlepšimi výstupmi sú sprostredkované presné konfigurácie s ich reportmi v elektronických prílohách.

Použité datasety a etické hľadisko

Experimenty využívajú verejné datasety BRACS a BCSS a taktiež vlastný kontrolovaný syntetický dataset tvorený z histologických snímok. BRACS [15] obsahuje 547 snímok od 189 pacientov, 4539 anotovaných oblastí a sedem slabo anotovaných tried karcinómu prsníka s exemplárnymi počtami ROI v jednotlivých triedach (N 484, PB 836, UDH 517, FEA 756, ADH 507, DCIS 790, IC 649). Problém tohto tried je nevyváženosť. BCSS [9] obsahuje 150 histologických snímok a približne 20 000 anotácií oblastí záujmu. Pre podmienenú difúziu bol použitý rozdelený súbor 41 731 tréningových a 100 validačných vzoriek s ROI maskami. Kontrolovaný syntetický dataset má 38 syntetických snímok s rozlíšením približne $13\,000 \times 13\,000$ pixelov, odvodených z 38 viactriednych sémantických másk (všetky z datasetu Histo-Seg [2]). Ide o inú tkanivovú doménu než rakovina prsníka, čo treba zohľadniť pri interpretácii štruktúr. Údaje slúžia čisto na výskumnú validáciu, nie na tréning. Výstupy práce nenahrádzajú klinické rozhodnutie a vyžadujú samostatné posúdenie patológmi pri akejkoľvek aplikácii mimo výskumu.

Metodika meraní

Pre generovanie sa používa Fréchetova inception vzdialenosť (FID) medzi reálnymi a syntetickými výrezmi. Inception Score (IS) ako hrubý ukazovateľ diverzity a klasifikovateľnosti vzoriek, ďalej PSNR a SSIM pre porovnanie generovaných patchov s referenciou pri podmienenom generovaní. Pri spojoch sa použili metódy na validáciu kvality spojov: zhoda hraníc segmentácie (Segment Anything) alebo superpixelov s časťou spoja, prípadne LPIPS v okrajových častiach pri porovnaní tvaru masiek spojov. Pri RA-HIPT sa okrem presnosti pomocnej klasifikácie sledovali krivky stratových funkcií, vizualizácie pozornosti jednotlivých hláv pozornosti, PCA embeddingov regiónov, stupňové rozdelenie uzlov v grafe vzťahov medzi regiónmi a diagnostiky kolapsu embeddingov.

FID a IS sú štandardné pre porovnanie AI generátorov obrázkov, no pre histológiu treba brať do úvahy, že Inception bol trénovaný na prírodných fotografiách mimo domény histológie. Ideálne hodnoty sa preto interpretujú relatívne medzi konfiguráciami v rámci práce, nie ako absolútna kvalita tkaniva. SSIM a PSNR pri podmienenom generovaní zachytávajú lokálnu zhodu s referenčným patchom počas experimentov. Vysoká kvalita generovaných snímok nie je v tejto práci hlavným cieľom, no presné generovanie pomocou sémantickej masky.

Pri RA-HIPT dopĺňajú tieto metriky primárne vizuálne a geometrické pohľady: PCA z embeddingov regiónov odhaľuje, či sa zhľuky rozumným spôsobom oddelujú alebo naopak kolabujú, graf vzťahov medzi regiónmi poskytuje vysvetliteľnú štatistiku stupňov váh medzi uzlami a ich súvislosti. Pomocná klasifikácia nad embeddingmi nie je samotná, ale slúži ako ďalší signál, či reprezentácia stále nesie informáciu napriek slabým anotáciám.

Navrhované riešenie: RA-HIPT a celkový prístup

Bežná analýza biomarkerov so slabými anotáciami a modelom bez globálneho kontextu je rozšírená o model zameriavajúci sa na vzťahy a druhým konceptom, ktorým sú kontrolované syntetické dáta. RA-HIPT vychádza z extrakcie embeddingov s pomocou HIPT na regiónoch 4096×4096 (prípadne jemnejších rozlíšeníach 256×256 v rámci HIPT) a následnej agregácii na úrovni celého WSI. Na vstupe lokálnej vetvy je vision transformer s blokom dvoj vetvového modulu pozornosti. Vetva *tissue* používa self-attention nad embeddingmi výrezov, pričom vetva *relation* používa cross-attention, kde jednotlivé lokálne črty pochádzajú z výrezu a kľúče/hodnoty z globálnych embeddingov, čo produkuje mapy pozornosti zamerané na vzťahy. Celková stratová funkcia agreguje členy pre tkanivo, vzťahy, diverzitu hláv pozornosti a ďalšie regularizácie s plánovaním váh počas tréningu. Grafová RA-HIPT rozširuje model o graf vzťahov medzi regiónmi a nové súvisiace stratové funkcie (diverzita, riedkosť, hladkosť, koherentnosť). Vzťahová vetva využíva okrem iného stratové funkcie typu focal a curiosity, no taktiež Mahalanobisom motivovanú kontrolu rozdielov medzi jednotlivými hlavami modulu pozornosti. Plánovanie hodnôt stratových funkcií je nastavené tak, aby sa predišlo kolapsu pozornosti alebo príliš jednotnému rozdeleniu váh medzi regiónmi. Časť stratovej funkcie využitá v tissue vetve je viazaná na pseudo-masky zachytávajúce tkanivo vs. pozadie. Tým sa má zlepšiť explicitné učenie dlhodobých závislostí oproti baseline bez grafu.

Postup kontrolovaného generovania syntetických dát

Postupnosť začína tréningom VQ-VAE na histologických výrezoch z BRACS (95 000 tréningových a 5 000 validačných výrezov 256×256 , normalizácia do rozsahu $(-1, 1)$), architektúra s kódovou knihou rozsahu radovo 10^4 embeddingov, perceptuálna

strata a diskriminátor v celej stratovej funkcii). Pri implementácii boli riešené artefakty v rohoch (bilineárna interpolácia namiesto najbližšieho suseda, a využitie reflexných okrajov). Druhým problémom boli nerovnomerné odtiene a zafarbenia medzi jednotlivých výrezov, čo je kritické pre následné spájanie týchto výrezov. Najlepší trénovací beh VQ-VAE je využitý pre zmrazený enkóder/dekóder pri difúzii. Nasleduje tréning latentného denoiseru (UNET) na nepodmienennej difúzii a následne semanticky podmienenej difúzii: vstupom je jednorázovo zakódovaná maska tried z BCSS (v hlavnej konfigurácii 21 tried), classifier-free guidance pri vzorkovaní, zmrazený dekóder VQ-VAE na obnovenie výrezov 256×256 . Na vyššej úrovni sa výrezy ukladajú na mriežku. Jednotlivé spoje sa riešia inpaintingom v latentnom priestore, používajú sa rovné alebo vlnité masky spojov a Reinhardova normalizácia farieb v Lab priestore medzi regiónmi. Pre veľké oblasti 4096×4096 sa používa postup zmenšenie–inpainting–zväčšenie kvôli obmedzenej pamäti grafickej karty. Celé WSI sa skladajú z blokov 4096×4096 s globálnou úpravou farieb. Kontrolované syntetické generovanie snímok bolo vykonané na 38 semantických maskách, čo vytvorilo 38 veľkých syntetických WSI s predpripravenými výrezmi 4096×4096 a 256×256 pre validáciu RA-HIPT.

Výsledky experimentov: syntetický pipeline

VQ-VAE a nepodmienená latentná difúzia

Na vyhodnotenie kvality rekonstrukcie a generovania z latentného priestoru bolo na 10 000 vzorkách namerané pre VQ-VAE FID **13,14**, priemerný Inception Score **1,33** (odchýlka IS v zápise experimentu **0,02**) a čas vyhodnotenia približne **5,14** min. Tréning nepodmienenej difúzie používal výrezy z BRACS (v rozdelení približne 9500 tréningových a 5000 validačných výrezov o veľkosti 256×256 , škálovanie intenzít jednotlivých pixelov bolo do rozsahu $(-1, 1)$), učiacu rýchlosť 0,0001, op-

timalizér Adam, dávku výzev 64, 15 epoch a MSE stratu v latentnom priestore. VQ-VAE enkóder a dekóder pri generovaní boli zmrazené.

Porovnanie vzorkovačov na 10 000 syntetických latentných obrázkoch je zhrnuté v tabuľke 13.1; pre rýchle experimenty je výhodný DDIM s 50 krokmi (najlepší kompromis medzi FID a časom), zatiaľ čo dlhé behy DDIM alebo PLMS v tejto konfigurácii neprinášajú lepšie FID než krátky DDIM.

Vzorkovač	T	FID ↓	IS ↑	Čas (min) ↓
DDPM	1000	35.54	1.36 ± 0.02	92.09
DDIM	50	33.37	1.28 ± 0.02	7.07
DDIM	200	36.11	1.23 ± 0.03	20.52
DDIM	500	38.11	1.26 ± 0.03	47.45
PLMS	50	38.65	1.30 ± 0.01	7.17
PLMS	200	39.37	1.25 ± 0.03	20.65
PLMS	500	39.75	1.27 ± 0.02	47.57

Table 13.1: Nepodmienená latentná difúzia: porovnanie vzorkovačov

Podmienená generácia na BCSS

Podmienený denoiser bol trénovaný na veľkom súbore regiónov z BCSS so zodpovedajúcimi maskami (21 tried v hlavnej konfigurácii, 1000 časových krokov difúzie, mierka classifier-free guidance 5,0 pri vzorkovaní). Architektúra UNET v najlepšom behu používa 160 základných kanálov, dva reziduálne bloky a pozornosť pri rozlíšeníach 16 a 8 v latentnom priestore. Masky tried je jednorázovo kódovaná a zmenšená na latentové rozlíšenie. Vygenerované podmienené výrezy priestorovo zodpovedajú maske tried. Artefakty textúr sú akceptovateľné vzhľadom na zameralenie práce na interpretovateľné rozloženie štruktúr, nie na fotorealistickú kvalitu snímok.

Porovnanie vzorkovačov na 10 000 vygenerovaných výrezoch je v tabuľke 13.2. DDIM s 50 krokmi zodpovedá preferovanému vzorkovaniu. Najnižšie FID v tejto

sade dosahuje PLMS s 500 krokmi.

Vzorkovač	T	FID ↓	PSNR (dB) ↑	SSIM ↑	Čas (min) ↓
DDPM	1000	161.47	10.10 ± 2.46	0.10 ± 0.05	152.54
DDIM	50	122.84	10.29 ± 3.14	0.13 ± 0.12	9.79
DDIM	200	122.29	10.46 ± 3.04	0.14 ± 0.12	32.41
DDIM	500	121.94	10.51 ± 3.04	0.13 ± 0.12	77.74
PLMS	50	121.65	10.32 ± 2.97	0.13 ± 0.11	9.96
PLMS	200	121.42	10.28 ± 3.04	0.13 ± 0.12	32.59
PLMS	500	120.41	10.53 ± 3.09	0.14 ± 0.12	77.88

Table 13.2: Podmienená generácia na BCSS: porovnanie vzorkovačov (metriky na 10 000 výrezoch).

Absolútne FID sú vyššie než pri nepodmienenej difúzii, čo zodpovedá priorite dodržania rozloženia tried podľa masky namiesto dokonalosti textúr. Na základe výsledkov validácie rozmiestnenia štruktúr sú výsledky v práci označené za dostatočné a to aj napriek artefaktom.

Skladanie regiónov, WSI a ablačná štúdia spojov

Región 1024×1024 bol zostavený zo štyroch podmienených výrezov 256×256 s inpaintingom spojov. Oproti naivnému spájaniu sú spoje vizuálne menej viditeľné. Pre 4096×4096 bol použitý postup zmenšenie–inpainting–zväčšenie, aby bolo možné spracovať veľkú mriežku v dostupnej pamäti grafickej karty. Globálne rozloženie tkaniva zostalo konzistentné s maskou. Pri celej WSI sa bloková štruktúra 4096×4096 kombinovala s globálnou farebnou normalizáciou. Viditeľnosť hraníc medzi veľkými blokmi môže zostať prítomná a je v práci otvoreným obmedzením pre ďalšie zlepšenie. V ablačných štúdiách dosiahla najlepšie výsledky kombinácia inpaintingu, **vlnitých** masiek spojov a farebnej normalizácie. Segment Anything a Felzenszwalbove superpixels slúžili na kvantifikáciu presnosti a recallu hraníc v pásme spoja: oproti naivnému dlaždeniu sa znížila zhoda hraníc s mriežkou.

Pri priamom porovnaní rovných a vlnitých spojov pomocou LPIPS v širších okrajových pásmach dosiahli vlnité masky lepšie výsledky ako rovné masky spojov.

Prieskum s laikmi

Šestnásť účastníkov (jeden pilot a pätnásť v hlavnom kole) plnilo úlohu hľadania v syntetickej snímke štruktúry podľa referenčného výrezu v aplikácii HistoView. Väčšina účastníkov úlohu dokončila v priebehu niekoľkých minút. Referenčný výrez bol podľa spätnej väzby dostatočný na orientáciu. Ako najľahšie nájditeľná sa javili oblasti tuku (vysoký kontrast voči okoliu), náročnejšie boli stromálne oblasti s postupným prechodom textúry. Tumory boli dobre hľadateľné pri silno viditeľných susedných štruktúrach s iným tkanivom, no pri nízkom zväčšení boli susedné oblasti občas vizuálne príbuzné. Skúmané boli predovšetkým kategórie stroma, lymfocytárna infiltrácia, tumor, cievy a tuk. Krátka tréningová podúloha pred hlavným testom podľa spätnej väzby znížila neistotu pri ovládaní anotačného nástroja a zjednodušila umiestnenie anotácie. Hlavné ťažkosti účastníci pripisovali skôr zo strany nástroja (či je anotácia zapísaná alebo citlivosť zoomu) než samotnej interpretácii tkaniva.

V ablačných štúdiách skladania sa porovnávali konfigurácie označené v práci ako B1-B8: od naivného dlaždenia cez postupné pridávanie inpaintingu, vlnitých spojov, zdieľanej inicializácie šumu a globálnej farebnej normalizácie. Čiarové grafy pre metriky spojov ukazujú konzistentné zlepšenie voči dlaždeniu B1. Segment Anything sa zameriava na presnosť a recall hraníc vo vzťahu k spoju a super-pixels na analogické trendy. Najnižšiu viditeľnosť prepojenia hraníc segmentácie s mriežkou dosiahla najsilnejšia kombinácia (inpainting + vlnité spoje + konzistencia farieb). Tým je experimentálne podložené, že zníženie spojov nie je len vizuálny dojem, ale merateľné aj cez segmentačné podúlohy.

Výsledky experimentov: RA-HIPT

Učenie bez učiteľa pri RA-HIPT (BCSS, regióny 4096×4096)

Konfigurácia pre základný model: optimalizér AdamW (učiacu rýchlosť 10^{-5} , váhový decay 10^{-4}), kosínusový rozvrh učenia ($T_{\max} = 25$, minimálna sadzba 10^{-6}), veľkosť dávky výrezov 16, 25 epoch. Cross-attention modul mal hĺbku 12, 12 hláv a rozmer embeddingu 192. Strata v režime učenia bez učiteľa kombinovala diverzitu, konzistenciu, kontrast, rekonstrukciu, curiosity a Mahalanobisovský motivovaný člen. Váha vzťahovej vetvy sa po úvodnom tréningu zvyšovala. Krivky tréningu boli spočiatku silne nestabilné, neskôr konvergovali. Vizualizácie pozornosti po jednotlivých hlavách ukázali, že po zavedení diversity regularizácie hlavy nekolabujú na identické mapy, ale špecializujú sa na rôzne časti tkaniva alebo pozadia.

Graph RA-HIPT, učenie bez učiteľa (BCSS, 4096×4096)

Pridanie regularizačnej grafovej štruktúry: AdamW, dávka výrezov 16, **15 epoch** (kratší beh), cross-attention hĺbka **6**, počet hláv 6, rozmer embeddingu **384**. Aktívne boli grafové stratové funkcie vrátane diverzity, riedkosti, hladkosti, predikcie, koherencie a obmedzení na grafe zachytávajúci vzťahy medzi regiónmi s postupným zapínaním jednotlivých stratových komponentov počas tréningu. Tepelné mapy agregovanej pozornosti ukázali zameranie na oblasti s tkanivom. Graf zachytávajúci vzťahy medzi regiónmi bol v porovnaní s neskorším behom s využitím slabých anotácií na syntetických dátach riedši, čo znamená menej výrazných hrán medzi regiónmi, toto v práci súvisí s chýbajúcou silnou metodikou tréningu s celkovými anotáciami a zložitou reálnych snímok. PCA embeddingov a diagnostiky kolapsu slúžili na sledovanie variability reprezentácie počas učenia. V niektorých fázach tréningu existovalo riziko kolapsu embeddingov, ktoré sa zmierňovalo

úpravami hyperparametrov a regularizačnými členmi obsahujúcich v celkovej strátovovej funkcii.

Z kvalitatívneho hľadiska nie je pri čistom SSL na reálnych WSI vždy možné bez doménového experta potvrdiť, či každá výrazná štruktúra na tepelnej mape zodpovedá konkrétnej biologickej morfológii. Presne táto medzera motivuje experiment na syntetických dátach so známou maskou pre celú snímku.

Graph RA-HIPT, s využitím slabých anotácií (kontrolovaný syntetický dataset, patch 256×256)

Konfigurácia: AdamW ($5 \cdot 10^{-5}$, decay 10^{-4}), dávka výrezov 16, 15 epoch. Cross-attention hĺbka **2**, **8** hláv, rozmer **512**. Využitie slabých anotácií a grafová regularizačná štruktúra boli súčasne zapnuté. **Známa sémantická maska** generátora umožnila porovnávať tepelné mapy s očakávaným rozmiestnením tried. Po tréningu pozornosť častejšie korešpondovala so štruktúrami v maske. Graf zachytávajúci vzťahy medzi regiónmi obsahoval **viac** významných spojení než v čistom SSL grafe na BCSS, čo experimentálne podporuje hypotézu, že syntetické dáta so známim rozložením štruktúr pomáha učiť model explicitnejšie vzťahy. Pomocná klasifikácia z embeddingov odhalila najväčšie problémy medzi morfologicky podobnými triedami (napr. nekrotické oblasti a odpad, krv, metaplázia), čo korešponduje s nevyváženosťou tried, no táto úloha nebola cieľom našej práce, preto sme sa na ňu ďalej nezameriavali.

Graph RA-HIPT, s využitím slabých anotácií (BRACS)

Hyperparametre zodpovedajú tréningy s využitím slabých anotácií (AdamW $5 \cdot 10^{-5}$, dávka výrezov 16, 15 epoch, hĺbka 2, 8 hláv, 512 rozmer, grafová štruktúra zapnutá). Tepelné mapy na **reálnych** WSI sú vizuálne menej homogénne než na

syntetických dátach, avšak po niekoľkých epochách model zvýrazňuje oblasti s biologicky súvisiacim tkanivom. PCA a rozdelenie stupňov uzlov v grafe naznačujú **menší** kolaps reprezentácie a stabilnejšie váhy hrán oproti čisto SSL grafovému experimentu na BCSS. Graf zachytávajúci vzťahy medzi regiónmi obnovuje susedstvá a väzby medzi regiónmi aj pri reálnych dátach. V texte práce sú vybrané výstupy diskutované aj vo vzťahu k medicínskej literatúre (napr. cievy, žľaza-metaplázia, lymfocytárna infiltrácia) ako nezávislá kontrola zmysluplnosti mapy, nie ako definitívna klinická interpretácia.

Zhodnotenie výskumných otázok

RQ1: na syntetických WSI s kontrolovanou maskou možno vizuálne skontrolovať, či tepelné mapy a graf vzťahov rešpektujú očakávané susedstvá tried. Hustejší graf zachytávajúci vzťahy medzi regiónmi pri slabo anotovanom tréningu Graph RA-HIPT na syntetických dátach podporuje hypotézu, že známa štruktúra pomáha učiť explicitnejšie väzby než čistý SSL graph na BCSS. Na BRACS s regularizačnou grafovou štruktúrou pri slabšie anotovanom tréningu model poskytuje konzistentnejšie regionálne vzťahy a dosahuje menej známkolapsu v PCA než čisto SSL experiment s grafovou štruktúrou na BCSS. Samotná medicínska korektnosť významu mapy však vyžaduje kontrolu patológom.

RQ2: tabuľky FID/IS pri nepodmienennej difúzii ukazujú, že DDIM 50 je výrazne rýchlejší než DDPM 1000 pri porovnateľnom alebo lepšom FID než dlhé behy DDIM/PLMS v tejto konfigurácii. Pri podmienení na BCSS najlepšie (najnižšie) FID v tabuľke dosahuje PLMS 500, no za cenu dlhého vzorkovania. DDIM 50 je praktický kompromis medzi FID približne 122–123, PSNR okolo 10,3 dB a časom pod 10 minút na 10 000 patchov. Metriky spojov a ablačné štúdie B1-B8 dokazujú, že metodika kontrolovaného generovania skutočne zlepšuje skladanie

výrezov oproti naivnému dlaždeniu, čo je predpoklad pre použitie syntetických WSI ako validačného datasetu.

RQ3: tepelné masky pre každú hlavu modulu pozornosti, PCA embeddingov a grafy zachytávajúce vzťahy medzi regiónmi poskytujú podklad na diskusiu s odborníkmi. Práca ich pri experimente s BRACS dátami dopĺňa o zmienky medicínskej literatúry ako nezávislú kontrolu výstupov. Avšak definitívna klinická validácia ostáva na strane patológov. Práca zámerne nepretvára klasifikáciu na jediný dôkaz úspechu, ale používa ju ako sekundárny signál pri nevyvážených triedach.

Implementácia a nástrojová časť

Experimenty sú implementované v jazyku Python s knižnicami pre hlboké učenie. Priebeh učenia, metriky a generované sumáre sú ukladané tak, aby bolo možné výsledky z príloh a logov spätne overiť. Veľké WSI sa spracúvajú po regiónoch a dávkach výrezov podľa pamäte GPU. Konfigurácie pre VQ-VAE, difúziu a RA-HIPT sú sprístupnené v technickej prílohe, aby boli opakovateľné v rovnakom softvérovom rámci.

Pre kvalitatívne úlohy bol pripravený prototyp anotačného a zobrazovacieho nástroja (HistoView) na prezeranie syntetických WSI a tepelných máp s ovládaním zoomu, panorámy a anotácií využitými v prieskume. Repozitár je verejne dostupný [54].

Obmedzenia a perspektívy

Medzi obmedzenia patrí doménový posun pri maskách odvodených z inej entity než histológia prsníka (vrátane odlišnej štruktúry tkanív oproti BRACS/BCSS), nerovnováha tried v BRACS ovplyvňujúca pomocnú klasifikáciu, artefakty pri veľmi veľkom rozlíšení generovaných snímok, pri latentnej kompresii VQ-VAE

(možná strata jemných textúr), časová a pamäťová náročnosť veľkých inpaintingových másk a možná viditeľnosť hraníc medzi blokmi 4096 pri celej WSI. Teoreticky limituje aj závislosť FID/IS na ImageNet-kompatibilnom extraktore znakov.

Závislosť významu tepelných máp a grafov na ďalšom potvrdení patológmi zostáva otvorená: práca poskytuje vizuálne a číselné podklady, nie definitívnu klinickú interpretáciu. Perspektívou je rozšírenie variability syntetických másk a farebnej normalizácií, doladenie inpaintingu na hraniciach blokov a pri extrémnych rozlíšeniach, špecializovanejšie trénovania a stratégie učenia bez učiteľa pre histológiu, rozšírenie počtu používateľských testov a formalizované hodnotenie s expertmi v slučke.

Súvisiace smery z literatúry (učenia so slabými anotáciami nad skupinami výrezov, štruktúrovaný latentný priestor, viacnásobná latentná difúzia) naznačujú, že ďalšia práca môže preskúmať menej známe prepojenie medzi generovanými morfológickými pravidlami a učením vzťahov v transformerovej architektúre, prípadne spoločné tréningové ciele, ktoré ešte výslovnejšie prepoja generátor a detektor biomarkerov na úrovni regiónov.

Záver

Práca integruje RA-HIPT s kontrolovaným syntetickým generovaním snímok a dodáva rozsiahlu experimentálnu a validačnú časť: od VQ-VAE s FID 13,14 cez systematické porovnanie DDPM/DDIM/PLMS na nepodmienennej latentnej difúzii (vrátane všetkých skúmaných počtov krokov a časov vzorkovania) až po podmienenú generáciu na BCSS s kompletnou tabuľkou FID/PSNR/SSIM a časov pre DDPM, DDIM i PLMS. Dokumentuje skladanie regiónov 1024×1024 a 4096×4096 , generovanie celej snímky, ablačné štúdie spojov s kvantitatívnymi metrikami a kvalitatívnym prieskum s laikmi pri vyhľadávaní štruktúr.

Na strane RA-HIPT práca podáva základy pre učenie bez učiteľa na BCSS datasete, rozšírenie regularizačnej grafovej štruktúry s diagnostikou kolapsu a PCA, slabo anotovanými grafovými experimentmi na kontrolovaných syntetických dátach (porovnanie tepelných máp so známou maskou a klasifikačnou maticou) a slabo anotovanými grafovými experimentmi na BRACS s porovnaním stability embeddingov voči čisto SSL grafovému nastaveniu. Výsledkom je metodicky zdokumentovaný rámec pre snímky s vysokým rozlíšením a validáciu vzťahov medzi regiónmi. Výstupy nenahrádzajú klinickú diagnózu a vyžadujú ďalšiu spoluprácu s patológmi pri prenose do praxe.

Bibliography

- [1] Mohamed A Abdou. “Literature review: Efficient deep neural networks techniques for medical image analysis”. In: *Neural Computing and Applications* 34.8 (2022), pp. 5791–5812.
- [2] Anum Abdul Salam et al. *Histo-Seg: H&E Whole Slide Image Segmentation Dataset*. Version 2. Dataset. Aug. 2025. DOI: 10 . 17632 / vccj8mp2cg . 2. URL: <https://data.mendeley.com/datasets/vccj8mp2cg/2>.
- [3] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. “On the surprising behavior of distance metrics in high dimensional space”. In: *International conference on database theory*. Springer. 2001, pp. 420–434.
- [4] Bulbul Agrawal, Shradha Dubey, and Manish Dixit. “Standard statistical feature analysis of image features for facial images using principal component analysis and its comparative study with independent component analysis”. In: *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019*. Springer. 2019, pp. 1027–1048.
- [5] Alhassan Ali Ahmed, Mohamed Abouzid, and Elżbieta Kaczmarek. “Deep learning approaches in histopathology”. In: *Cancers* 14.21 (2022), p. 5264.

- [6] Shahd A Alajaji et al. “Generative adversarial networks in digital histopathology: current applications, limitations, ethical considerations, and future directions”. In: *Modern Pathology* 37.1 (2024), p. 100369.
- [7] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. “BoQ: A Place is Worth a Bag of Learnable Queries”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 17794–17803.
- [8] Esther Ugo Alum. “AI-driven biomarker discovery: enhancing precision in cancer diagnosis and prognosis”. In: *Discover Oncology* 16.1 (2025), pp. 1–12.
- [9] Mohamed Amgad et al. “Structured crowdsourcing enables convolutional segmentation of histology images”. In: *Bioinformatics* 35.18 (2019), pp. 3461–3467.
- [10] Andrea Asperti and Valerio Tonelli. *Comparing the latent space of generative models*. 2022. arXiv: 2207.06812 [cs.LG]. URL: <https://arxiv.org/abs/2207.06812>.
- [11] Adrien Bardes, Jean Ponce, and Yann LeCun. “Vicreg: Variance-invariance-covariance regularization for self-supervised learning”. In: *arXiv preprint arXiv:2105.04906* (2021).
- [12] Shane Barratt and Rishi Sharma. “A note on the inception score”. In: *arXiv preprint arXiv:1801.01973* (2018). arXiv: 1801.01973. URL: <https://arxiv.org/abs/1801.01973>.
- [13] Kamal Berahmand et al. “Autoencoders and their applications in machine learning: a survey”. In: *Artificial Intelligence Review* 57.2 (2024), p. 28.
- [14] Samuel Bohumel. “Synthesis of Annotated Histology Data by Neural Networks”. In: (2025). Unpublished diploma thesis.

- [15] Nadia Brancati et al. “Bracs: A dataset for breast carcinoma subtyping in h&e histology images”. In: *Database 2022* (2022), baac093.
- [16] Rasmus Bro and Age K Smilde. “Principal component analysis”. In: *Analytical methods* 6.9 (2014), pp. 2812–2831.
- [17] Gabriele Campanella et al. “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images”. In: *Nature medicine* 25.8 (2019), pp. 1301–1309.
- [18] Sergen Cansiz. “Multivariate Outlier Detection in Python”. In: *Medium* (2021). Accessed: 28.05.2025. URL: <https://medium.com/data-science/multivariate-outlier-detection-in-python-e946cfc843b3>.
- [19] Marc-André Carbonneau et al. “Measuring disentanglement: A review of metrics”. In: *IEEE transactions on neural networks and learning systems* 35.7 (2022), pp. 8747–8761.
- [20] Ahmad Chaddad et al. “Generalizable and explainable deep learning for medical image computing: An overview”. In: *Current Opinion in Biomedical Engineering* (2024), p. 100567.
- [21] Chi-Long Chen et al. “An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning”. In: *Nature communications* 12.1 (2021), p. 1193.
- [22] Richard J. Chen et al. *Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning*. 2022. arXiv: 2206.02647 [cs.CV]. URL: <https://arxiv.org/abs/2206.02647>.
- [23] S Chen and W Guo. *Auto-Encoders in Deep Learning—A Review with New Perspectives. Mathematics 2023, 11, 1777*. 2023.
- [24] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PmLR. 2020, pp. 1597–1607.

- [25] Hoor college. *Signal Processing*. <https://www.astro.rug.nl/~vdhulst/SignalProcessing/Hoorcolleges/college03.pdf>. Accessed: 28.05.2025.
- [26] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. “The mahalanobis distance”. In: *Chemometrics and intelligent laboratory systems* 50.1 (2000), pp. 1–18.
- [27] Shujian Deng et al. “Deep learning in digital pathology image analysis: a survey”. In: *Frontiers of medicine* 14 (2020), pp. 470–487.
- [28] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.
- [29] Lin Ding, Mohamad El Zaatari, and Juanita L Merchant. “Recapitulating human gastric cancer pathogenesis: experimental models of gastric cancer”. In: *Stem Cells, Pre-neoplasia, and Early Cancer of the Upper Gastrointestinal Tract* (2016), pp. 441–478.
- [30] Carl Doersch. “Tutorial on variational autoencoders”. In: *arXiv preprint arXiv:1606.05908* (2016).
- [31] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).
- [32] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [33] Cian Eastwood et al. “Dci-es: An extended disentanglement framework with connections to identifiability, 2022”. In: *URL <https://arxiv.org/abs/2210.00364>* ().
- [34] Andre Esteva et al. “A guide to deep learning in healthcare”. In: *Nature medicine* 25.1 (2019), pp. 24–29.

- [35] Theodore Evans et al. “The explainability paradox: Challenges for xAI in digital pathology”. In: *Future Generation Computer Systems* 133 (2022), pp. 281–296.
- [36] Daniel Geissler et al. “Enhancing Interpretability Through Loss-Defined Classification Objective in Structured Latent Spaces”. In: *arXiv preprint arXiv:2412.08515* (2024).
- [37] “Generative Learning Trilemma”. In: (2023). URL: <https://www.researchgate.net/publication/376640959>.
- [38] Amirehsan Ghasemi et al. “Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review”. In: *Cancer Innovation* 3.5 (2024), e136.
- [39] Fatemeh Ghezloo et al. “Robust roi detection in whole slide images guided by pathologists’ viewing patterns”. In: *Journal of Imaging Informatics in Medicine* 38.1 (2025), pp. 439–454.
- [40] Benyamin Ghogh et al. “Uniform manifold approximation and projection (UMAP) and its variants: tutorial and survey”. In: *arXiv preprint arXiv:2109.02508* (2021).
- [41] Laurent Girin et al. “Dynamical variational autoencoders: A comprehensive review”. In: *arXiv preprint arXiv:2008.12595* (2020).
- [42] Miriam Hägele et al. “Resolving challenges in deep learning-based analyses of histopathological images using explanation methods”. In: *Scientific reports* 10.1 (2020), p. 6423.
- [43] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.
- [44] Andreas Holzinger et al. “What do we need to build explainable AI systems for the medical domain?” In: *arXiv preprint arXiv:1712.09923* (2017).

- [45] Artur Hore and Djemel Ziou. “Image quality metrics: PSNR vs. SSIM”. In: *20th International Conference on Pattern Recognition*. IEEE. 2010, pp. 2366–2369.
- [46] Yuantian Huang, Satoshi Iizuka, and Kazuhiro Fukui. “Diffusion-Based Semantic Image Synthesis from Sparse Layouts”. In: (2023), pp. 441–454.
- [47] Ekta Jain et al. “Whole slide imaging technology and its applications: Current and emerging perspectives”. In: *International Journal of Surgical Pathology* 32.3 (2024), pp. 433–448.
- [48] Laya Jose et al. “Generative adversarial networks in digital pathology and histopathological image processing: a review”. In: *Journal of Pathology Informatics* 12.1 (2021), p. 43.
- [49] Kenhub. *Stomach histology*. <https://www.kenhub.com/en/library/anatomy/stomach-histology>. Accessed: 2026-04-30. 2019.
- [50] Mahendra Khened et al. “A generalized deep learning framework for whole-slide image segmentation and analysis”. In: *Scientific reports* 11.1 (2021), p. 11579.
- [51] Minyoung Kim et al. “Relevance factor vae: Learning and identifying disentangled factors”. In: *arXiv preprint arXiv:1902.01568* (2019).
- [52] Sora Kim, Sungho Suh, and Minsik Lee. “Rad: Region-aware diffusion models for image inpainting”. In: (2025), pp. 2439–2448.
- [53] Frederick Klauschen et al. “Toward explainable artificial intelligence for precision pathology”. In: *Annual Review of Pathology: Mechanisms of Disease* 19.1 (2024), pp. 541–570.
- [54] Patrik Kozlík. *Heatmap.Viewer*. GitHub repository. 2026. URL: <https://github.com/0Kozlik0/Heatmap.Viewer> (visited on 04/14/2026).
- [55] Moez Krichen. “Convolutional neural networks: A survey”. In: *Computers* 12.8 (2023), p. 151.

- [56] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. “Variational inference of disentangled latent concepts from unlabeled observations”. In: *arXiv preprint arXiv:1711.00848* (2017).
- [57] Nhat Minh Le et al. “Learning biologically relevant features in a pathology foundation model using sparse autoencoders”. In: *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*.
- [58] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical Image Analysis* 42 (Dec. 2017), 60–88. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.07.005. URL: <http://dx.doi.org/10.1016/j.media.2017.07.005>.
- [59] Luping Liu et al. “Pseudo Numerical Methods for Diffusion Models on Manifolds”. In: (2022). URL: <https://openreview.net/forum?id=PlKWVd2yBkY>.
- [60] Xiao Liu et al. “Learning disentangled representations in the imaging domain”. In: *Medical Image Analysis* 80 (2022), p. 102516.
- [61] Francesco Locatello et al. “A sober look at the unsupervised learning of disentangled representations and their evaluation”. In: *Journal of Machine Learning Research* 21.209 (2020), pp. 1–62.
- [62] Francesco Locatello et al. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.
- [63] Haijing Luan et al. “Review of deep learning-based pathological image classification: From task-specific models to foundation models”. In: *Future Generation Computer Systems* 164 (2025), p. 107578.
- [64] Andreas Lugmayr et al. “Repaint: Inpainting using denoising diffusion probabilistic models”. In: (2022), pp. 11461–11471.

- [65] Andreas Madsen, Siva Reddy, and Sarath Chandar. “Post-hoc interpretability for neural nlp: A survey”. In: *ACM Computing Surveys* 55.8 (2022), pp. 1–42.
- [66] Johanna M Mäkinen et al. “Epithelial cell proliferation and glandular atrophy in lymphocytic gastritis: effect of H pylori treatment”. In: *World Journal of Gastroenterology: WJG* 9.12 (2003), p. 2706.
- [67] Pawan Kumar Mall et al. “A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities”. In: *Healthcare Analytics* 4 (2023), p. 100216.
- [68] Christine McCaffrey et al. “Artificial intelligence in digital histopathology for predicting patient prognosis and treatment efficacy in breast cancer”. In: *Expert Review of Molecular Diagnostics* 24.5 (2024), pp. 363–377.
- [69] Leland McInnes, John Healy, and James Melville. “UMAP: uniform manifold approximation and projection for dimension reduction. arXiv”. In: *arXiv preprint arXiv:1802.03426* 10 (2018).
- [70] Goeffrey J McLachlan. “Mahalanobis distance”. In: *Resonance* 4.6 (1999), pp. 20–26.
- [71] Grégoire Montavon et al. “Layer-wise relevance propagation: an overview”. In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 193–209.
- [72] Tianhang Nan et al. “Deep learning quantifies pathologists’ visual patterns for whole slide image diagnosis”. In: *Nature Communications* 16.1 (2025), p. 5493.
- [73] Mohammad Sadegh Nasr et al. “Clinically relevant latent space embedding of cancer histopathology slides through variational autoencoder based image compression”. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–5.

- [74] Sergey Nikolenko. “Generative AI II: Discrete Latent Spaces”. In: *Synthesis AI* (2023). Accessed: 28.05.2025. URL: <https://synthesis.ai/2023/03/21/generative-ai-ii-discrete-latent-spaces/>.
- [75] Ahmad Obeid et al. “Advancing Histopathology with Deep Learning Under Data Scarcity: A Decade in Review”. In: *arXiv preprint arXiv:2410.19820* (2024).
- [76] *OpenSeadragon*. Open-source web-based viewer for high-resolution zoomable images. 2026. URL: <https://openseadragon.github.io/> (visited on 04/14/2026).
- [77] Dominika Petříková and Ivan Cimrák. “Survey of recent deep neural networks with strong annotated supervision in histopathology”. In: *Computation* 11.4 (2023), p. 81.
- [78] Slobodan Petrovic. “A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters”. In: *Proceedings of the 11th Nordic workshop of secure IT systems*. Vol. 2006. Citeseer. 2006, pp. 53–64.
- [79] Matteo Pozzi et al. “Generating and evaluating synthetic data in digital pathology through diffusion models”. In: *Scientific Reports* 14.1 (2024), p. 28435.
- [80] Arsela Prelaj et al. “Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review”. In: *Annals of Oncology* 35.1 (2024), pp. 29–65.
- [81] Gokuldas (Vedant) Sarvesh Raikar, Amisha Sarvesh Raikar, and Sandesh Narayan Somnache. “Advancements in artificial intelligence and machine learning in revolutionising biomarker discovery”. In: *Brazilian Journal of Pharmaceutical Sciences* 59 (2023), e23146.
- [82] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. “Deep learning for medical image processing: Overview, challenges and the future”. In:

- Classification in BioApps: Automation of decision making* (2018), pp. 323–350.
- [83] Erik Reinhard et al. “Color Transfer between Images”. In: *IEEE Computer Graphics and Applications* 21.5 (2001), pp. 34–41.
 - [84] Carl O Retzlaff et al. “Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists”. In: *Cognitive Systems Research* 86 (2024), p. 101243.
 - [85] Oren Rippel et al. “Metric learning with adaptive density discrimination”. In: *arXiv preprint arXiv:1511.05939* (2015).
 - [86] Lucia Roa-Peña, Francisco Gómez, and Eduardo Romero. “An experimental study of pathologist’s navigation patterns in virtual microscopy”. In: *Diagnostic pathology* 5.1 (2010), p. 71.
 - [87] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.
 - [88] Pouria Rouzrokh et al. “A Current Review of Generative AI in Medicine: Core Concepts, Applications, and Current Limitations”. In: *Current Reviews in Musculoskeletal Medicine* (2025), pp. 1–21.
 - [89] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. “Dynamic routing between capsules”. In: *Advances in neural information processing systems* 30 (2017).
 - [90] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models”. In: *arXiv preprint arXiv:1708.08296* (2017).
 - [91] Julian Schön, Raghavendra Selvan, and Jens Petersen. “Interpreting latent spaces of generative models for medical images using unsupervised methods”. In: *MICCAI Workshop on Deep Generative Models*. Springer. 2022, pp. 24–33.

- [92] Ramprasaath R Selvaraju et al. “Grad-CAM: Why did you say that?” In: *arXiv preprint arXiv:1611.07450* (2016).
- [93] Ghazal Shamsipour et al. “Improve the efficiency of handcrafted features in image retrieval by adding selected feature generating layers of deep convolutional neural networks”. In: *Signal, image and video processing* 18.3 (2024), pp. 2607–2620.
- [94] Dawei Shen, Yaozhong Zhang, and Seiya Imoto. “Weakly Supervised Vector Quantization for Whole Slide Images Classification”. In: *bioRxiv* (2024), pp. 2024–08.
- [95] Rushikesh Shende. “Autoencoders, Variational Autoencoders (VAE), and -VAE”. In: *Medium* (2023). Accessed: 28.05.2025. URL: <https://medium.com/@rushikesh.shende/autoencoders-variational-autoencoders-vae-and-%CE%B2-vae-ceba9998773d>.
- [96] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising diffusion implicit models”. In: *arXiv preprint arXiv:2010.02502* (2020).
- [97] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. “A survey of dimensionality reduction techniques”. In: *arXiv preprint arXiv:1403.2877* (2014).
- [98] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. “Deep neural network models for computational histopathology: A survey”. In: *Medical image analysis* 67 (2021), p. 101813.
- [99] Jan Stühmer, Richard Turner, and Sebastian Nowozin. “Independent subspace analysis for unsupervised learning of disentangled representations”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1200–1210.

- [100] Sergey Suchkov et al. “Personalized and Precision Medicine (Ppm) Through the View of Biomarker-Driven Applications to Support Their Global Impact in the Upgraded Healthcare-Related Services”. In: (2024).
- [101] Mohammad Mustafa Taye. “Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions”. In: *Computation* 11.3 (2023), p. 52.
- [102] David Tellez et al. “Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology”. In: *Medical image analysis* 58 (2019), p. 101544.
- [103] Erico Tjoa and Cuntai Guan. “A survey on explainable artificial intelligence (xai): Toward medical xai”. In: *IEEE transactions on neural networks and learning systems* 32.11 (2020), pp. 4793–4813.
- [104] Carlo Tomasi. *Computer Vision: Image Convolution, Filtering, and Gradient*. <https://courses.cs.duke.edu/fall113/compsci527/notes/convolution.pdf>. COMPSCI 527 Supplementary Lecture Notes, Duke University, Last Modified on August 6, 2013. Aug. 2013.
- [105] Maximilian E Tschuchnig, Gertie J Oostingh, and Michael Gadermayr. “Generative adversarial networks in digital pathology: a survey on trends and future potential”. In: *Patterns* 1.6 (2020).
- [106] James ED Tweel et al. “Automated whole slide imaging for label-free histology using photon absorption remote sensing microscopy”. In: *IEEE Transactions on Biomedical Engineering* 71.6 (2024), pp. 1901–1912.
- [107] Aaron Van Den Oord, Oriol Vinyals, et al. “Neural discrete representation learning”. In: *Advances in neural information processing systems* 30 (2017).
- [108] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

- [109] Wei Wang et al. “Diffusion model for medical image denoising, reconstruction and translation”. In: *Computerized Medical Imaging and Graphics* (2025), p. 102593.
- [110] Weilun Wang et al. “Semantic image synthesis via diffusion models”. In: *arXiv preprint arXiv:2207.00050* (2022).
- [111] Zhou Wang et al. “Image quality assessment: From error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.
- [112] Jiawen Wei, Hugues Turbé, and Gianmarco Mengaldo. “Revisiting the robustness of post-hoc interpretability methods”. In: *arXiv preprint arXiv:2407.19683* (2024).
- [113] Yuxiang Wei et al. “Orthogonal jacobian regularization for unsupervised disentanglement in image generation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6721–6730.
- [114] Lilian Weng. “From Autoencoder to Beta-VAE”. In: *Lil’Log* (2018). Accessed: 28.05.2025. URL: <https://lilianweng.github.io/posts/2018-08-12-vae/>.
- [115] Xiaoliang Xie et al. “Evaluating cancer-related biomarkers based on pathological images: a systematic review”. In: *Frontiers in oncology* 11 (2021), p. 763527.
- [116] Yifan Xu et al. “Transformers in computational visual media: A survey”. In: *Computational Visual Media* 8 (2022), pp. 33–62.
- [117] Cheng Yang, Lijing Liang, and Zhixun Su. “Real-world denoising via diffusion model”. In: *arXiv preprint arXiv:2305.04457* (2023).
- [118] Srikar Yellapragada et al. *ZoomLDM: Latent Diffusion Model for multi-scale image generation*. 2025. arXiv: 2411.16969 [cs.CV]. URL: <https://arxiv.org/abs/2411.16969>.

- [119] Yu Yu, Weibin Zhang, and Yun Deng. “Frechet inception distance (fid) for evaluating gans”. In: *arXiv preprint arXiv:2105.04906* (2021). arXiv: 2105.04906. URL: <https://arxiv.org/abs/2105.04906>.
- [120] Chong Zhang et al. “AI in spotting high-risk characteristics of medical imaging and molecular pathology”. In: *Precision Clinical Medicine* 4.4 (2021), pp. 271–286.
- [121] Xuanru Zhou et al. “Generative Artificial Intelligence in Medical Imaging: Foundations, Progress, and Clinical Translation”. In: *arXiv preprint arXiv:2508.09177* (2025).
- [122] Carmine Zoccali et al. “Biomarkers in clinical epidemiology studies”. In: *Clinical Kidney Journal* 17.6 (2024), sfae130.

Declaration of Use of Artificial Intelligence

In accordance with the Rector's Measure of the Slovak University of Technology in Bratislava of 28 November 2024 on declaring the use of artificial intelligence (AI) tools in final theses, I state below the parts of this work and the activities in its preparation where I used AI tools, how they were used, and which tools were involved.

Translation between languages.

The Slovak **Resumé** was prepared from the English version with the help of generative AI translation assistance. (ChatGPT, GPT-5.3)

Text editing, reformulation, and grammar.

Minor reformulations and consistency edits to the English thesis text, together with AI-assisted grammar and style checking. (Perplexity AI)

Routine python scripts for visualizations

Generation of selected routine visualization scripts for creation of plots, and auxiliary visualizations in the Experiments chapter (with subsequent manual review, labeling, and final approval). (Cursor AI)

Bibliography

Software, documentation, and code.

Completing and validation of source-code documentation. Refactoring and cleaning of self-written code with AI assistance. (Cursor AI)

I confirm that all methodological decisions, experiments, interpretation of results and final responsibility for the content of the thesis remain with the author. The AI tools were not used as authors or co-authors of the thesis.

In Bratislava, 4.5.2026

Bc. Patrik Kozlík

