



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Efficient Multiscale Deep Learning for Artifact Detection in Digital Pathology

TESI DI LAUREA MAGISTRALE IN
COMPUTER SCIENCE AND ENGINEERING - INGEGNERIA IN-
FORMATICA

Author: **Andrea Faccioli**

Student ID: 271316

Advisor: Prof. Giacomo Boracchi

Co-advisors: Emin Zerman, Matti Pellikka

Academic Year: 2025-26

Abstract

Digital pathology has enabled the integration of computational methods into clinical workflows through the analysis of Whole Slide Images. However, the presence of preparation and acquisition artifacts, such as tissue folds or external objects, can negatively affect both human diagnosis and the performance of computational pathology systems. This research aims to develop a robust, computationally efficient quality control pipeline by adapting Vision Transformers for automated artifact detection. Furthermore, it seeks to formally model and structurally integrate the detection of scanner-specific stitching artifacts into a unified deep learning architecture. A novel multiscale Vision Transformer architecture is proposed to effectively capture both local and global features, leveraging a two-branch framework that processes concentrically cropped image patches at different magnifications. To improve training procedure that is tainted by the scarcity of annotated data, a knowledge distillation strategy was adopted using artifact masks generated by a Convolutional Neural Network teacher model. Results demonstrate that the proposed model significantly outperformed both the baseline teacher model and a single-scale Vision Transformer architecture in terms of segmentation accuracy. Additionally, the structural integration of the stitching artifact detection module successfully identified scanner-specific anomalies without compromising the performance of the overall quality control pipeline and with minimal computational overhead. Automated, early artifact detection streamlines the diagnostic process, reducing the time and resources spent on manual review and re-scanning, ultimately improving patient outcomes.

Keywords: Computational Pathology, Vision Transformers, Artifact Detection, Whole Slide Imaging, Knowledge Distillation.

Abstract in lingua italiana

La patologia digitale ha permesso l'integrazione di metodi computazionali nei flussi di lavoro clinici attraverso l'analisi di Whole Slide Images (immagini a intero vetrino). Tuttavia, la presenza di artefatti di preparazione e acquisizione, come pieghe del tessuto o oggetti esterni, può influire negativamente sia sulla diagnosi umana che sulle prestazioni dei sistemi di computational pathology. Questa ricerca mira a sviluppare una pipeline di controllo qualità robusta ed efficiente dal punto di vista computazionale, adattando un'architettura basata su Vision Transformer per il rilevamento automatizzato degli artefatti. Inoltre, lo studio cerca di modellare formalmente e integrare il rilevamento di artefatti di stitching specifici dello scanner all'interno di un'architettura unificata. Viene proposta una nuova architettura Vision Transformer multiscala per catturare efficacemente sia le feature locali che quelle globali, sfruttando un modello a due rami che elabora riquadri di immagini ritagliate concentricamente a diversi ingrandimenti. Per migliorare la procedura di training, penalizzata dalla scarsità di dati annotati, è stata adottata una strategia di knowledge distillation utilizzando maschere di artefatti generate da un modello teacher basato su una Convolutional Neural Network. I risultati dimostrano che il modello proposto supera significativamente sia il modello teacher di baseline, sia un'architettura Vision Transformer a singola scala in termini di accuratezza della segmentazione. Inoltre, l'integrazione del modulo di rilevamento degli artefatti di stitching ha identificato con successo anomalie specifiche dello scanner senza compromettere le prestazioni dell'intera pipeline di controllo qualità e con un overhead computazionale minimo. Il rilevamento automatizzato e precoce degli artefatti ottimizza il processo diagnostico, riducendo il tempo e le risorse impiegate per la revisione manuale e il re-scanning, migliorando in ultima analisi gli esiti per i pazienti.

Keywords: Computational Pathology, Vision Transformers, Artifact Detection, Whole Slide Imaging, Knowledge Distillation.

Contents

Abstract	i
Abstract in lingua italiana	iii
Contents	v
1 Introduction	1
1.1 Introduction to digital pathology	1
1.2 Problem statement	2
1.2.1 Research questions	3
1.3 Scope	3
1.4 Outline	4
2 Theory	5
2.1 Digital microscopy	5
2.2 Artifacts in digital pathology: classification and causes	6
2.3 Artifact detection in digital pathology: from statistical priors to deep learning methods	8
2.4 Vision Transformer	12
2.5 SegFormer	14
3 Methodology	17
3.1 Design Research Methodology	17
3.1.1 DRM application in the context of this study	18
3.2 Training dataset	19
3.2.1 Macro artifacts	19
3.2.2 Stitching artifact	21
3.3 Testing dataset	21
3.3.1 Macro artifacts	22

3.3.2	Stitching artifact	23
3.4	Evaluation methodology	25
3.4.1	Macro artifacts	25
3.4.2	Stitching artifacts	26
3.4.3	Inference speed	27
4	Implementation	29
4.1	Experimental setup	29
4.2	Macro artifacts dataset pre-processing	30
4.3	Tissue segmentation network	34
4.4	Artifact segmentation network	34
4.4.1	Standard SegFormer	35
4.4.2	Multiscale SegFormer	35
4.4.3	Multitask SegFormer	37
4.5	Overview of the proposed architecture	40
5	Results	47
5.1	Evaluation of macro artifacts	47
5.2	Evaluation of stitching artifact	63
5.3	Evaluation of the inference speed	72
6	Conclusions	75
6.1	Discussion	75
6.1.1	Addressing RQ1: Artifact Detection & Knowledge Distillation . . .	75
6.1.2	Addressing RQ2: Stitching Artifact Detection & Architecture Inte- gration	76
6.2	Limitations and future work	77
6.3	Ethical considerations	79
	Bibliography	81
	List of Tables	87
	List of Tables	87
	List of Symbols	89

1 | Introduction

Cancer remains a leading cause of mortality across the globe, with an incidence of diagnosed cases of 22 million and an estimated number of deaths of 9.74 million in 2022. These numbers are predicted to rise within the next decades, projected to 35.3 million and 18.5 million respectively [5]. Such data supports the importance of the research in the field. The diagnosis of cancer is a complex process that relies on the expertise of pathologists, who examine tissue samples under a microscope to identify abnormalities and determine the presence of malignancy. While traditionally performed using light microscopes, recent advancements in technology have led to the digitization of pathology slides, enabling the use of Whole Slide Images (WSIs) for diagnostic purposes. WSIs are high-resolution digital representations of tissue samples that can be viewed and analyzed on a computer screen, building the foundation for the field of digital pathology. The preparation of these WSIs is a complex laboratory routine that includes biopsy extraction, chemical fixation, dehydration, embedding, sectioning into thin slices, and staining (most commonly with Hematoxylin and Eosin) to create diagnostic contrast among cellular structures [24].

1.1. Introduction to digital pathology

Digital pathology is a rapidly evolving field that has transformed the way pathologists analyze and interpret tissue samples. By digitizing traditional glass slides into WSIs, digital pathology allows for remote access, easier storage, and the application of advanced computational techniques to assist in diagnosis [21]. Second-opinion consultations, quick access to prior cases, and medical education are some others widespread applications of digital pathology [19]. Moreover, the digitization of pathology slides has paved the way for the integration of machine learning algorithms into the diagnostic workflow, shaping the emerging field of Computational Pathology.

CPath is a sub-branch of digital pathology that leverages algorithms and machine learning techniques to analyse medical images. It offers a collection of instruments that can help pathologists in different phases of the clinical workflows. Quality control tools can be used to identify and flag slides with technical flaws, such as artifacts, before they

reach the diagnostic desk. Image analysis algorithms can assist in the identification and quantification of specific features within the tissue, such as size, shape, and distribution of cells, which previously required time-consuming manual visual estimation by pathologists. Finally, deep learning models can be trained to perform specific diagnostic tasks, such as classifying tissue samples as benign or malignant, grading tumors, and predict time-to-event outcomes (e.g. overall survival) [8].

1.2. Problem statement

Although digital pathology enables advanced computational analysis of WSIs, the reliability of both human diagnosis and CPath algorithms depends heavily on the quality of the images entering the workflow. To understand where quality issues arise and how they affect downstream analysis, it is useful to consider the clinical digital pathology workflow. This workflow can be divided into three main phases, each of which contributes to the final diagnostic outcome [16]:

- **Pre-analytical phase:** Encompasses all steps from tissue acquisition to slide preparation and digitization. Because it involves numerous physical and hardware-dependent steps, it is highly susceptible to introducing visual flaws.
- **Analytical phase:** Involves the actual examination and interpretation of the histological images to formulate a diagnosis. In modern digital workflows, this consists of a pathologist evaluating the WSIs on a diagnostic monitor, increasingly supported by Computer-Aided Diagnosis (CAD) tools, deep learning models, and image analysis algorithms.
- **Post-analytical phase:** Covers the processes following the diagnostic interpretation. This includes the generation and delivery of the final pathology report, communication with the treating clinical team, and the long-term digital archiving of the WSIs.

Among these phases, the pre-analytical stage is particularly critical because errors introduced at this point directly affect all subsequent steps of the workflow. Approximately 66% of reported workflow errors originate during the pre-analytical phase [1]. One of the most important challenges within pre-analytical quality control is the detection of image artifacts, which is the focus of this thesis.

Artifacts are unwanted structures or inconsistencies that obstruct the clear visualization of tissue structures, originating from diverse sources including flaws in slide preparation and improper settings during slide digitalization [22, 27]. Detecting these artifacts in the

early stages of a pathology workflow has multiple beneficial outcomes. First, it streamlines the work of human pathologists by flagging slides that need to be re-cut, re-stained, or re-scanned before they ever reach a diagnostic desk, saving hours of professional time. Second, and crucially, it ensures that Computational Pathology (CPath) algorithms are not misled by technical flaws or surgical debris. If left unchecked, these artifacts can mask essential clinical features or trigger false positives and false negatives. Wang et al. [46] demonstrated that the introduction of common artifacts—such as altered staining, tissue folds, and air bubbles—causes measurable performance degradation in CPath models, manifesting both as a decrease in accuracy and a significant increase in prediction variance. Therefore, robust, automated artifact detection is not merely a quality control convenience; it is a fundamental requirement for the safe and reliable deployment of CPath tools in clinical pathology.

While the clinical necessity of pre-analytical quality control is clear, deploying such systems in routine practice presents a significant computational challenge. Whole Slide Images (WSIs) are gigapixel-sized, and artifacts manifest at vastly different scales, ranging from massive tissue folds to small debris. Moreover, the lack of comprehensive, publicly accessible datasets presents a major barrier to research. Driven by strict patient privacy frameworks and competitive industry interests, organizations frequently withhold annotated medical images. This data fragmentation makes it exceptionally difficult to train, validate, and generalize deep learning architectures across different clinical environments.

1.2.1. Research questions

Driven by the objective to build a robust, clinically viable quality control pipeline, this thesis seeks to answer the following research questions:

RQ1: How effectively and efficiently can Vision Transformers be adapted through knowledge distillation for automated artifact detection in Whole Slide Images (WSIs)?

RQ2: How can stitching artifact detection be formally modeled, structurally integrated into a Vision Transformer architecture, and robustly evaluated within a digital pathology pipeline?

1.3. Scope

This study strictly focuses on the development and evaluation of a deep learning-based artifact detection system for Whole Slide Images (WSIs) acquired with Grundium Ocus scanners in digital pathology. The scope is limited to a subset of common artifacts,

including tissue folds, foreign artifacts, pen markings, and out-of-focus regions, as well as scanner-specific stitching artifacts. The research will primarily utilize publicly available datasets, alongside manually annotated Grundium Ocus WSIs acquisition. The staining type will be limited to Hematoxylin and Eosin (H&E), which is the most widely used in clinical practice. The algorithms developed will be tested on datasets captured with Grundium Ocus devices, and the RQ2 will specifically address stitching artifacts from this type of scanner (tile-based acquisition, see Section 2.1). Vision Transformers will be the primary deep learning architecture explored.

1.4. Outline

The remainder of this thesis is structured as follows: Chapter 2 provides the theoretical background necessary for the study, detailing the context of digital microscopy, the classification and causes of artifacts, existing quality control methods, and the foundational concepts of Vision Transformers and the SegFormer architecture. Chapter 3 outlines the research methodology and thoroughly describes the composition of the training and testing datasets alongside the formal evaluation methodology utilized. Chapter 4 details the technical implementation, covering the experimental setup, data pre-processing, and the iterative development of the artifact segmentation networks. Chapter 5 presents a comprehensive evaluation of the results, analyzing the models' quantitative and qualitative performance in detecting the different types of artifacts, as well as assessing their computational efficiency. Finally, Chapter 6 concludes the thesis by discussing the findings in relation to the research questions, acknowledging current limitations, suggesting directions for future work, and reflecting on the ethical considerations of the research.

2 | Theory

2.1. Digital microscopy

Traditional microscopy has been the cornerstone of pathology for centuries, allowing pathologists to examine tissue samples at high magnification. However, the advent of digital microscopy has revolutionized the field by enabling the digitization of glass slides into high-resolution images [7]. The process begins when a prepared glass slide is placed into the scanner. A highly precise, motorized mechanical stage automatically moves the slide along the X and Y axes, ensuring the entire tissue surface is systematically exposed to the objective lens. Light passes through the tissue and the objective lens (typically at 20x or 40x magnification) and hits a high-resolution digital camera equipped with Charge-Coupled Device (CCD) or Complementary Metal-Oxide Semiconductor (CMOS) sensors. These sensors convert the optical light signals into digital data. Two primary scanning methods are employed: tile-based (grid) scanning and line-based scanning [30]. Line-based scanning captures image data continuously in long, uninterrupted strips. In contrast, tile-based scanning captures discrete, overlapping tiles that are later stitched together to form the complete image.

To make viewing efficient, WSI data is typically structured in a multiresolution "pyramidal" format [2] paired with spatial tiling. While the pyramidal structure provides progressively down-sampled versions above the massive, high-resolution base layer, tiling subdivides each level into discrete, independently encoded rectangular regions. Instead of decoding full horizontal line strips, viewing software can fetch only the specific tiles that intersect with the user's viewport. This allows the system to instantly display a low-resolution overview and seamlessly load the heavy, high-resolution data strictly for the area the pathologist zooms into. Grundium Ocus M40 scanner, for example, can produce images with a resolution of up to 0.25 microns per pixel at 40x magnification, that are stored in pyramidal formats with multiple levels of resolution (i.e. 40x, 10x, 2.5x), as illustrated in Figure 2.1.

Because the uncompressed base layer of a Whole Slide Image (WSI) can easily span sev-

eral gigabytes, efficient data compression is a mandatory step in the digitization pipeline. Typically, the individual tiles or lines comprising the image are compressed using lossy algorithms, such as JPEG, before being packaged into the final pyramidal container (e.g., TIFF, SVS, or DICOM formats). JPEG compression works by transforming the image data into the frequency domain using the Discrete Cosine Transform (DCT) [3], quantizing the coefficients to reduce precision, and then encoding the data using entropy coding techniques. While JPEG compression drastically reduces storage requirements and network bandwidth for remote viewing, it introduces a necessary trade-off between file size and image fidelity. Excessive compression ratios can lead to the loss of fine morphological details or introduce digital compression artifacts, such as blockiness, ringing, or color bleeding. Consequently, compression levels must be carefully evaluated.

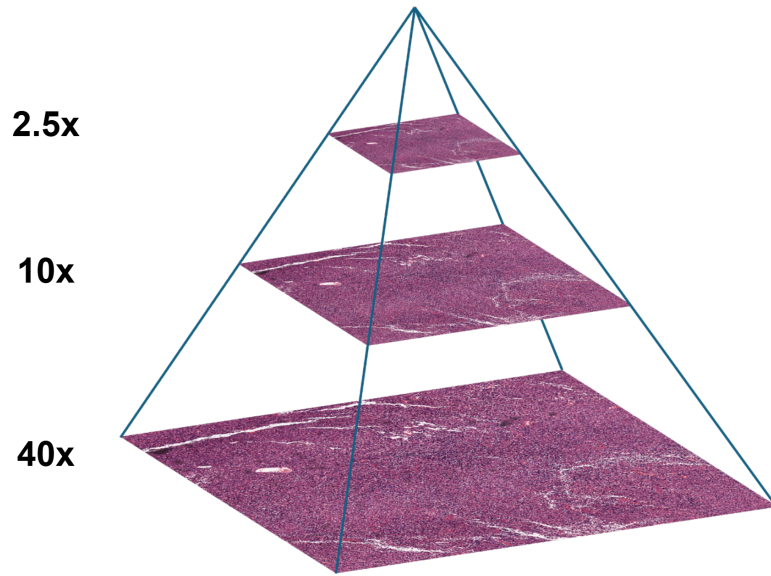


Figure 2.1: Pyramidal image structure in digital pathology. The base level contains the high-resolution image (40x), while higher levels contain down-sampled versions (10x, 2.5x) for efficient viewing and navigation.

2.2. Artifacts in digital pathology: classification and causes

Kanwal et al. [24] proposed a comprehensive classification system for artifacts in digital pathology, categorizing them based on their underlying causes within the WSI generation pipeline.

The acquisition of a histopathological glass slides typically begins with a *biopsy*, where a tissue sample is extracted from the patient. To preserve the tissue’s structural integrity

and prevent degradation, the sample undergoes *fixation* using specific chemical solutions. The tissue is then processed through a series of steps, including *dehydration*, *clearing*, and *embedding*, where it is infiltrated with a supportive medium (e.g., paraffin wax) to facilitate thin *sectioning*. The sections are then flattened and placed onto a glass slide during the *flotation* step. To highlight structural components and provide diagnostic contrast, the slide undergoes *staining*, where specific dyes are applied to the tissue in order to enhance the visibility of cellular and extracellular features. The tissue is then *mounted* beneath a coverslip and enters *storage*. Finally, the physical slide is *digitized* using a high-resolution microscopy scanner to produce the final WSI. The entire acquisition pipeline is illustrated in Figure 2.2.

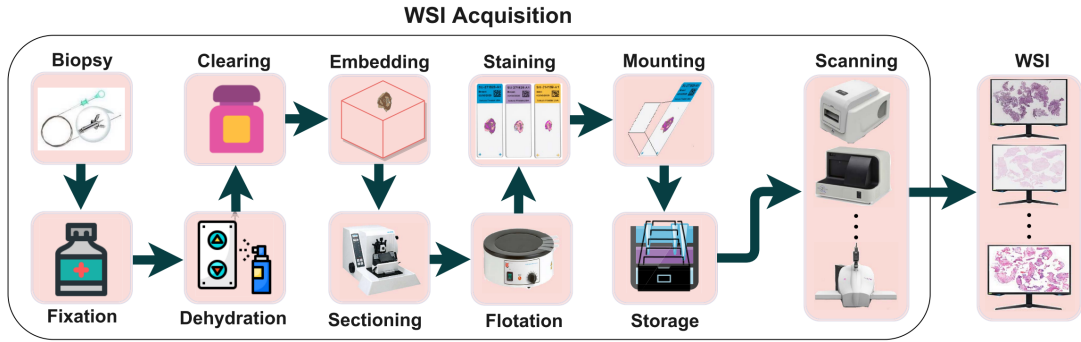


Figure 2.2: WSI acquisition pipeline illustrating the sequential steps from biopsy to digitization. Each step can introduce specific types of artifacts that may affect the quality of the final image [24].

Each of these steps can introduce specific types of artifacts, each with distinct rectification possibilities. Based on their origin, artifacts can be broadly classified into three categories:

- **Tissue-level artifacts** are produced during the initial acquisition and chemical processing of the sample, spanning from the biopsy to the staining steps. Because they physically alter the tissue itself, these artifacts are frequently impossible to rectify without obtaining an entirely new patient biopsy. Examples include tissue folding, tearing, or incomplete staining.
- **Slide-level artifacts** are associated with the final manual workflow steps, primarily mounting and storage. Unlike tissue-level errors, these can sometimes be resolved by cleaning or re-mounting the physical slide. Examples include dust, air bubbles, or scratches on the slide.
- **Scanner-level artifacts** are introduced exclusively during the digital acquisition phase and do not exist on the physical glass slide. If detected, these can be possibly solved simply by re-scanning the slide. Blur is the most common scanner-level

artifact, stitching errors can also occur based on the scanning method.

2.3. Artifact detection in digital pathology: from statistical priors to deep learning methods

Artifact detection in digital pathology has evolved significantly over the years, transitioning from traditional computer vision techniques to advanced deep learning models [37].

Early attempts relied heavily on image processing techniques, such as edge detection and thresholding of specific visual characteristics like color, saturation, and intensity. Palokangas et al. [34] utilized k-means clustering to successfully extract folded areas based on saturation and intensity. Kothari et al. [29] built on top of this and proposed to identify tissue folds by calculating the difference between color saturation and intensity, and then applying thresholding to extract the folded regions. Swiderska-Chadaj et al. [41] demonstrated the use of handcrafted color and morphological features in whole-slide image analysis, employing CMYK (Cyan, Magenta, Yellow, Key) color-space information and mathematical morphology to distinguish blood vessels from visually similar non-informative blood regions in hemorrhoid tissue slides. Traditional computer vision techniques, however, were limited in their ability to generalize across different tissue types and staining protocols, as well as in their capacity to handle sparse tissue regions.

As the field advanced, researchers began extracting “handcrafted” features—such as texture, shape, and contrast—and feeding them into classic machine learning classifiers like Support Vector Machines (SVMs) and Random Forests. Janowczyk et al. introduced *HistoQC* [20], an open-source quality control tool that extracts handcrafted features (color, brightness, contrast, and edges) and feeds them into ML models to identify both artifacts and artifact-free regions. It progressively refines a tissue/useful-region mask using thresholding, morphology, blur detection, color/texture metrics, and optional trained classifiers such as Gaussian Naive Bayes or Random Forests. Shakhawat et al. [38] detected air bubbles and tissue folds at low resolutions by extracting features such as contrast, correlation, homogeneity, energy, entropy, variance, sum average, sum variance, sum entropy, difference variance, and difference entropy, and then feeding them into a SVM classifier. Campanella et al. [6] provide a useful transition between these paradigms by directly comparing handcrafted sharpness descriptors combined with a Random Forest classifier against deep convolutional neural networks for detecting out-of-focus regions in WSIs. Their results show that feature-based machine learning can achieve strong performance

when informative sharpness metrics are carefully selected, while convolutional neural networks models can learn discriminative representations directly from image patches and achieve superior performance, motivating the shift from manual feature engineering toward data-driven deep learning approaches for artifact detection.

The most recent and high-performing methods rely on deep neural networks. These models bypass handcrafted features entirely, learning complex, non-linear representations of artifacts directly from large datasets. One of the first deep learning approaches was proposed by Senaras et al. [36] who introduced *DeepFocus*, a convolutional neural network (CNN) trained to classify image patches as in-focus or out-of-focus. Unlike earlier handcrafted methods, DeepFocus learns discriminative blur features directly from raw pixel data. The authors explicitly motivate deep learning by arguing that blur in histology is nonlinear and variable, making conventional handcrafted feature approaches less suitable. Subsequent studies on focus quality assessment focused on streamlining the annotation process by applying synthetic augmentation [28] [22] and improving the model’s efficiency by leveraging lightweight architectures [47]. Foucart et al. [12] extended artifact detection beyond single-artifact methods by training deep residual networks [14] on different artifact type annotations (blur, tissue folds, tears, ink stains), comparing tile-level classification and pixel-level segmentation strategies to identify artifact-free tissue regions suitable for downstream analysis. While covering multiple artifact types, their method simply distinguishes between artifact and artifact-free regions, without providing a more granular classification of specific artifact types. Smit et al. [39] extended artifact detection to multi-class semantic segmentation through an encoder-decoder architecture. A pixel-level segmentation model was trained to classify each pixel into one of seven classes: normal tissue, out-of-focus regions, tissue folds, ink, dust, marker, and air bubbles. This approach, combined with a decision tree-based quality control module, translated the predicted artifact masks into practical quality-control actions such as cleaning, rescanning, recutting, or taking no action. Similarly, Haghighat et al. [13] proposed *PathProfiler*, a ResNet-based multi-label quality assessment framework that predicts patch-level artifact heatmaps and slide-level usability, focus, and H&E staining scores, showing that deep features can better capture complex quality issues such as staining degradation and tissue folds than handcrafted feature-based classifiers.

Building on this direction, recent work has continued to refine multi-class artifact segmentation and/or classification by designing more comprehensive pipelines that combine improved deep learning architectures, finer-grained pixel-level predictions, and more efficient inference strategies to support practical WSI quality control. Weng et al. [49] proposed *GrandQC*, a two-stage deep-learning framework in which a tissue detection module

first separates tissue from background, and a subsequent artifact segmentation module performs pixel-level multi-class classification only within the detected tissue regions. This approach significantly reduces computational overhead by avoiding artifact classification in background areas. Yet, the artifact segmentation module relies on a standard CNN architecture, which may not fully capture the complex spatial relationships and contextual information necessary for accurate artifact classification. Kanwal et al. [25] further investigated the trade-off between detection performance and computational efficiency by proposing an artifact-processing pipeline based on a mixture-of-experts (MoE) strategy. Instead of training a single multi-class model, they trained independent binary experts for specific artifact morphologies, including damaged tissue, blur, tissue folds, air bubbles, and histologically irrelevant blood, and then fused their predictions to obtain a final artifact classification thanks to a gating mechanism. Importantly, their study also introduced Vision Transformer-based models into the artifact detection pipeline, comparing Vision Transformer (ViT-Tiny [42]) backbones with MobileNet-based CNNs [18] in both MoE and multi-class settings. While the MobileNet-based MoE achieved the best overall balance between performance and computational cost, the inclusion of Vision Transformers reflects the growing interest in attention-based architectures for capturing broader spatial context and improving robustness to out-of-distribution WSIs. *SlideInspect* [35] brings together many of these recent developments within a single quality-control framework, combining multi-stage tissue and artifact segmentation, Transformer-based contextual feature extraction, stain quality assessment, and slide-level action recommendation. Specifically, it uses low-magnification tissue segmentation to restrict subsequent analysis to relevant tissue regions, performs pixel-level multi-class artifact segmentation using an architecture with a Swin Transformer backbone [32], and converts the resulting artifact and stain-quality measurements, obtained from the stain intensity and contrast of the histological tissue, into actionable outputs such as no action, re-scan, re-mount, or re-cut. However, its artifact discrimination remains relatively coarse, as the segmentation output is limited to three artifact classes (out-of-focus regions, mounting artifacts, and tissue folds), thereby grouping several visually distinct artifacts, such as bubbles, dust, pen marks, and coverslip issues, under the broader mounting artifact category. Finally, Wang et al. [48] proposed *DiffusionQC*, which departs from fully supervised artifact classification and segmentation by reframing artifact detection as an outlier detection problem. Instead of learning explicit boundaries between predefined artifact classes, *DiffusionQC* trains a latent diffusion model based on a Variational Autoencoder (VAE) on artifact-free tissue patches and identifies artifacts through the residual error produced during denoising, with an enhanced variant using contrastive learning to further separate clean and artifact-containing samples in latent space. This formulation substantially reduces the

need for pixel-level artifact annotations and enables promising cross-stain generalization, but it also sacrifices semantic specificity, since detected regions are flagged as artifact-like deviations rather than being assigned to distinct artifact categories. Taken together, these recent methods show that deep learning-based WSI quality control is moving toward increasingly comprehensive and deployment-oriented pipelines, yet important trade-offs remain between annotation cost, artifact-class coverage, computational efficiency, and generalization to unseen stains, tissues, and artifact types.

Table 2.1 summarizes the key characteristics of the most relevant work in deep learning-based WSI quality control, highlighting their artifact class coverage, model architecture and training data.

Table 2.1: Summary of related work in Deep Learning-based WSI artifact detection

Work	Year	Output Artifact Classes	Architecture	# Training WSIs
Senaras et al. <i>DeepFocus</i> [36]	2018	Out-of-focus	CNN	16
Foucart et al. [12]	2018	Artifact vs Artifact-free	CNN	ND
Smit et al. [39]	2021	Out-of-focus, tissue folds, ink, dust, marker, air bubbles	CNN	100
Haghighat et al. <i>PathProfiler</i> [13]	2021	Staining issue, tissue folds, out-of-focus, other artifact	CNN	119
Weng et al. <i>GrandQC</i> [49]	2024	Tissue folds, pen markings, edge and air bubbles, out-of-focus, foreign objects	CNN	420
Kanwal et al. [25]	2024	Damaged tissue, out-of-focus, tissue folds, air bubbles, blood	MoE (MobileNet and ViT)	35
Scotto et al. <i>SlideInspect</i> [35]	2026	Out-of-focus, mounting artifacts, tissue folds	ViT	2160
Wang et al. <i>DiffusionQC</i> [48]	2026	Artifact vs Artifact-free	VAE + ViT	24

2.4. Vision Transformer

The Vision Transformer (ViT) is a deep learning architecture that applies the principles of the Transformer, originally developed for natural language processing, to computer vision tasks [10]. The core idea behind ViT is to treat an image as a sequence of patches, similar to how words are treated in text processing. The image is split into fixed-size patches (e.g., 16x16 pixels), which are then flattened into 1-dimensional vectors and passed through a learnable linear projection to create patch embeddings. These embeddings are then complemented with positional encodings to let the model know where a patch is located. The resulting sequence of patch embeddings is fed into multiple Transformer encoder blocks that consist of multi-head self-attention mechanisms, Multi-Layer Perceptrons (MLPs), and layer normalization. Finally, a classification head is applied to the output of the Transformer encoders to produce the final prediction. ViT has demonstrated competitive performance on various image classification benchmarks, often outperforming traditional convolutional neural networks (CNNs) when trained on large datasets. Figure 2.3 illustrates the classic ViT architecture.

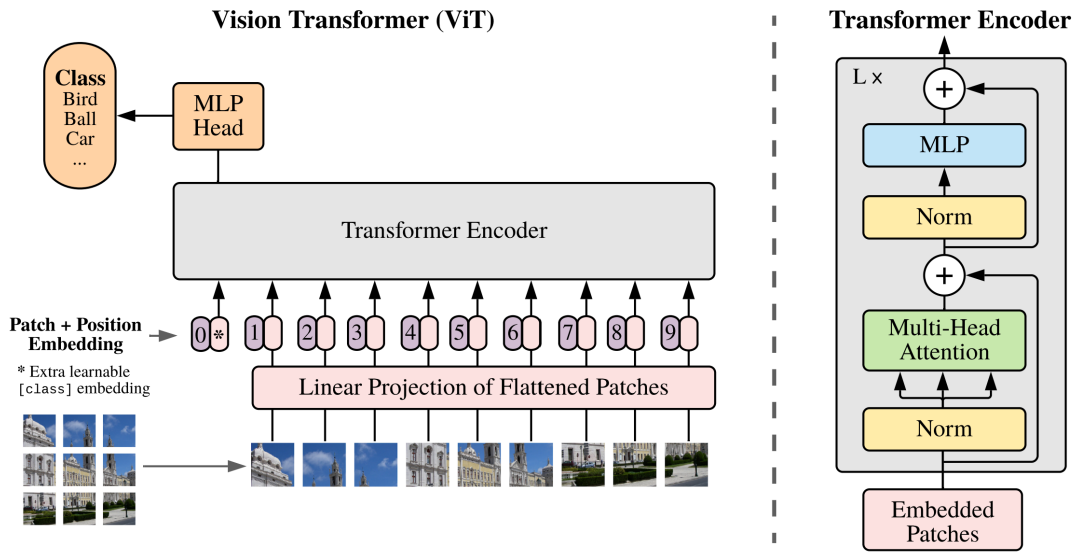


Figure 2.3: Vision Transformer (ViT) architecture. The input image is divided into fixed-size patches, which are then embedded and processed through multiple Transformer encoder blocks to produce the final output [10].

Unlike convolutions that only look at immediate neighboring pixels, **self-attention** [45] connects any pixel to any other pixel between different patches, regardless of how far apart they are. This helps the model understand the broader context of an entire image. Following the patch embedding stage of the ViT, the input image is represented as a

sequence of discrete tokens. Let this sequence be denoted as:

$$X = [x_1, x_2, \dots, x_n] \quad (2.1)$$

where n is the total number of patches, and each x_i represents the continuous vector embedding of the i -th patch.

The core of the architecture relies on the self-attention mechanism, which allows the model to dynamically weigh the importance of different patches relative to one another, capturing global spatial dependencies. To achieve this, each token x_i is linearly projected into three distinct learned representations using learnable weight matrices W_q , W_k , and W_v :

$$q_i = x_i W_q \quad (\text{Query}) \quad (2.2)$$

$$k_i = x_i W_k \quad (\text{Key}) \quad (2.3)$$

$$v_i = x_i W_v \quad (\text{Value}) \quad (2.4)$$

Intuitively, the **query** vector (q_i) represents the information the current patch i is seeking from the rest of the image. The **key** vector (k_j) represents the semantic content contained within patch j . The **value** vector (v_j) holds the actual feature representation of patch j that will be aggregated and passed forward.

To determine the amount of focus (or "attention") patch i should allocate to patch j , we compute an attention score using the dot product between the query of patch i and the key of patch j . To stabilize gradients during training, this dot product is scaled by the square root of the key dimension, d_k :

$$\text{score}(i, j) = \frac{q_i \cdot k_j}{\sqrt{d_k}} \quad (2.5)$$

These raw scores indicate the alignment between patches. To convert them into a valid probability distribution, a softmax function is applied across all patches m for the given patch i . This yields the attention weights, a_{ij} , ensuring that the weights for any patch i sum to one:

$$a_{ij} = \frac{\exp(\text{score}(i, j))}{\sum_{m=1}^n \exp(\text{score}(i, m))} \quad (2.6)$$

With the attention weights calculated, the self-attention context vector for patch i , denoted as z_i , is computed as the weighted sum of all value vectors v_j . This operation allows the model to pull relevant feature information from the entire image context into

the current token:

$$z_i = \sum_{j=1}^n a_{ij} v_j \quad (2.7)$$

Finally, to facilitate gradient flow and preserve the original feature information, this aggregated context is added back to the original token via a residual connection:

$$x'_i = x_i + z_i \quad (2.8)$$

where x'_i is the updated token representation that is subsequently passed to the next component of the Transformer block.

2.5. SegFormer

While the standard Vision Transformer excels at image classification, it faces notable limitations when applied to dense prediction tasks such as semantic segmentation. Specifically, ViT outputs single-scale, low-resolution feature maps and incurs exceptionally high computational costs when processing high-resolution images. To address these challenges, the SegFormer architecture was proposed by Xie et al. [51] as a more efficient and effective Transformer-based model for semantic segmentation. Figure 2.4 illustrates the SegFormer architecture, showing its main components.

The first major contribution is the hierarchical Transformer encoder. Unlike standard ViT-style models, which produce a single low-resolution feature map, SegFormer produces multi-scale features at different resolutions: roughly 1/4, 1/8, 1/16, and 1/32 of the input image. This is important for segmentation because the model needs both fine spatial detail and high-level semantic context. At the same time, the hierarchical design lowers the quadratic cost of standard self-attention by reducing the number of tokens to be processed at each stage. Moreover, the encoder uses overlapping patch embeddings instead of non-overlapping ViT patches. This helps preserve local continuity between neighboring patches, which is useful for dense prediction in digital pathology images. The second contribution is that SegFormer removes positional encodings. Many Vision Transformers use fixed positional embeddings, but these can hurt performance when the test image resolution differs from the training resolution. SegFormer instead uses a Mix-FFN block with a 3×3 convolution inside the feed-forward network, which implicitly provides positional information. This design choice allows for higher flexibility.

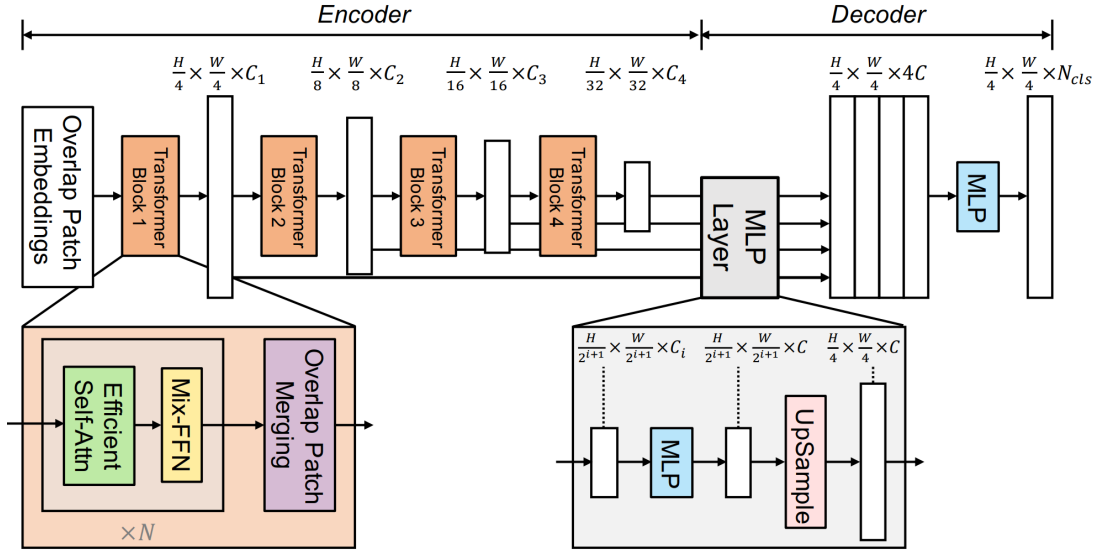


Figure 2.4: SegFormer architecture. SegFormer introduces a hierarchical Transformer encoder that produces multi-scale feature maps, and a lightweight MLP-based decoder that fuses these features for accurate segmentation, while maintaining computational efficiency [51].

3 | Methodology

This chapter outlines the methodological framework utilized to develop and evaluate the proposed artifact detection models. First, it introduces the research methodology that guides the overall structure of the research, detailing its four constituent stages and their specific application to this project. Following this, the chapter provides a comprehensive overview of the datasets constructed for both the training and testing phases. Finally, it establishes the formal evaluation methodology, defining the quantitative metrics used to rigorously assess the proposed models.

3.1. Design Research Methodology

This study employs the Design Research Methodology (DRM) proposed by Blessing and Chakrabarti [4]. This methodology outlines four primary stages to ensure the rigorous, evidence-based development of new artefacts:

- **Criteria Formulation:** Define the ultimate goals of the project utilizing *success criteria* and *measurable criteria*. Success criteria establish the long-term business or practical aims of the research, whereas measurable criteria serve as practical, quantifiable translations used to actually evaluate the project’s outcomes within its specific scope.
- **Descriptive Study I (DS-I):** This stage investigates current approaches to similar or affine problems through literature review and direct observation. The objective is to identify the factors influencing the measurable criteria, ultimately resulting in one or more “reference models”.
- **Prescriptive Study (PS):** Utilizing the insights gained from DS-I, practical interventions—such as new tools, methods, or guidelines—are developed. These interventions aim to transition the existing state into the desired, improved state.
- **Descriptive Study II (DS-II):** This final stage focuses on the formal evaluation of the newly developed artefact. It involves *application evaluation* to check if the tool actually affects the specific factors it was designed to target, and *success evaluation*

to determine if it ultimately improves the overall measurable success criteria.

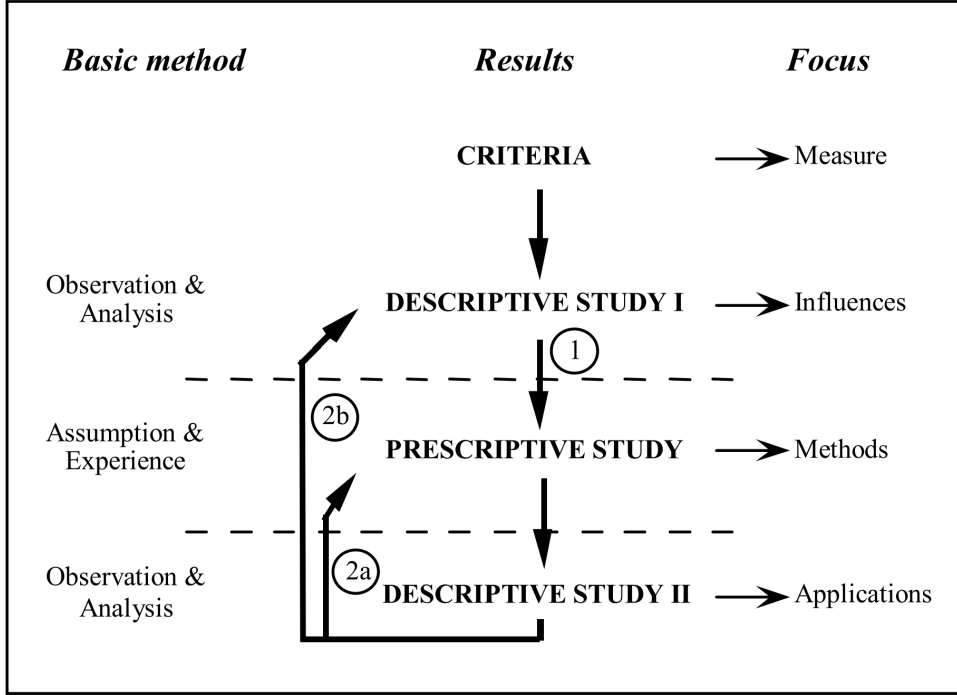


Figure 3.1: Design Research Methodology framework overview [4]

DRM is not meant to be a strictly sequential process. Instead, many iterations take place, and some stages might even run in parallel. Figure 3.1 reports an high-level overview of the DRM framework.

3.1.1. DRM application in the context of this study

The four stages of the Design Research Methodology correspond directly to the main phases of this project:

- **Criteria Formulation:** Project goals are defined in Section 1.2.1. Measurable criteria are defined through task-specific evaluation metrics, including Balanced Accuracy, mIoU (mean Intersection over Union), and Dice Score. The evaluation methodology is formalised in Section 3.4.
- **Descriptive Study I (DS-I):** This stage consists of the literature review presented in Chapter 2. Selected implementations were installed and executed on the test set to establish a benchmark for comparative analysis.
- **Prescriptive Study (PS):** This stage encompasses the formal modeling and experimentation of the candidate architectures. Implementation details are provided in Chapter 4.

- **Descriptive Study II (DS-II):** The proposed methodology is evaluated in Chapter 5. Ongoing evaluations informed further analysis of the prior work and guided refinements to the implementation.

3.2. Training dataset

The composition of the training dataset was designed to address the dual objectives outlined in Section 1.2.1. It comprises two distinct subsets: a first set containing common artifacts (such as tissue folds, out-of-focus regions, foreign objects, and pen markings), and a second set dedicated specifically to stitching artifacts. From now on, the term “macro artifacts” will be used to refer to the former subset. A summary of the sources of the resulting training dataset is provided in Table 3.1.

Table 3.1: Sources of the artifact segmentation training dataset.

Artifacts	Source	# WSIs	# WSIs used (cf. § 4.2)
Macro	GrandQC	13,543	75
Stitching	Grundium Ocus acquisitions	12	12

3.2.1. Macro artifacts

As described above, macro artifacts term refer to the common artifacts that appear in WSIs, such as tissue folds, out-of-focus regions, foreign objects, and pen markings.

To overcome data scarcity, a form of knowledge distillation has been adopted. Knowledge distillation, within the scope of this project, is the process of leveraging the outputs of a reference model (or "teacher") to generate additional training data for a new model (or "student") [15]. This approach allowed us to partially bypass the time consuming and costly process of manual annotation, while still providing a rich dataset for training the student model. In this case, the teacher model is the GrandQC tool [49]. GrandQC is a deep learning-based tool that has been trained on a large dataset of annotated histopathological images (420 WSIs) and it has demonstrated strong performance in identifying various types of artifacts, including blurriness, pen marks, and tissue folding. The authors of GrandQC have made the tool available as open-source software, together with the artifact masks generated on 13,543 WSIs from 32 cohorts from the TCGA repository. The TCGA (The Cancer Genome Atlas) repository is a comprehensive resource of genomic and clinical data from thousands of cancer patients, owned by the National Cancer Institute. The generated masks are published under a Non-Commercial license [26], which allows us to use them for research purposes. The WSIs contained in the TCGA repository, paired with

the corresponding masks generated by the GrandQC tool, compose the macro artifacts training set.

The available masks contain the following 7 classes of macro artifacts:

- Normal tissue
- Tissue Fold
- Dark Spot and Foreign Object
- Pen Marker
- Edge and Air Bubble
- Out of Focus
- Background

A mask is stored in an image in PNG format at 10x magnification, where each pixel value corresponds to the class of artifact present at that location. For example, a pixel value of 1 would correspond to "Normal tissue", while a pixel value of 2 would correspond to "Fold". In our implementation, the class "Edge and Air Bubble" has been ignored (i.e. set pixels to value 255), as the authors of GrandQC have merged two artifacts characterized by very different visual features into a single class, which could lead the student model to confusion. After color mapping is applied, the mask can be visualized as a color-coded image, where each color represents a different class of artifact. This allows for easy identification and analysis of the artifacts present in the WSIs. An example of the masks generated by GrandQC, paired with the corresponding image, is shown in Figure 3.2.

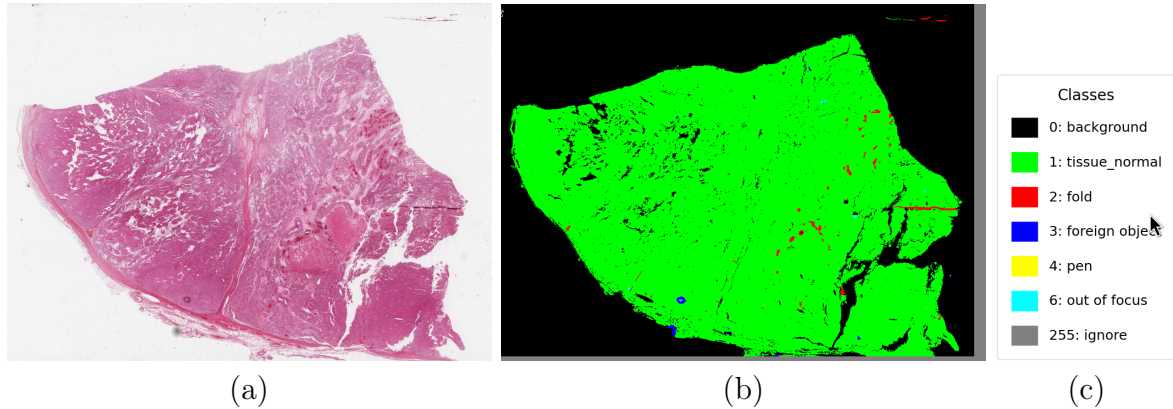


Figure 3.2: Example of artifact masks generated by GrandQC: (a) input image, (b) predicted artifact mask, and (c) class legend.

3.2.2. Stitching artifact

Stitching artifacts are a specific type of artifact that can occur during the process of creating a whole slide image (WSI) from multiple smaller images (camera field of views, FOVs). These artifacts can manifest as visible seams or misalignments between the tiles, which can affect the quality and accuracy of the WSI. GrandQC does not provide specific masks for stitching artifacts, as they are not explicitly included in the artifact classes. Moreover, due to the nature of this artifact, its morphology is highly connected to the specific scanning device and software used for the stitching process. Therefore, the macro artifacts dataset was complemented with a small set of manually annotated acquisitions containing stitching artifacts.

12 WSIs were scanned using a Grundium Ocus device with the stitching post-processing step and/or slide locking mechanism disabled. The resulting WSIs were then manually annotated using QuPath, an open-source software for digital pathology image analysis. The annotation style consists in "LineString" objects, that have been later rasterized into binary masks at the same resolution of the macro artifacts dataset (10x magnification). A visualization of a sample annotation for the stitching artifact is reported in Figure 3.3. 504 patches were extracted from the annotated WSIs. During this stage, dilation was applied to each stitching line, forcing the line width to be spanning 4 pixels.

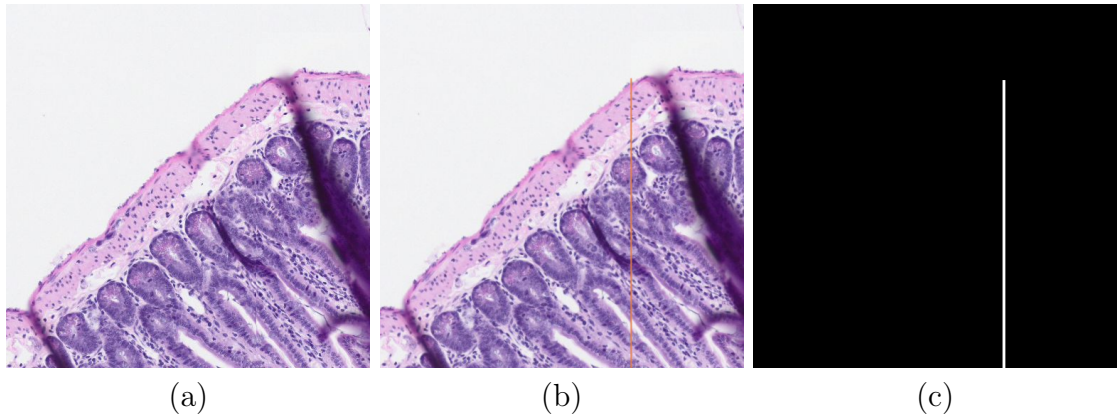


Figure 3.3: Example of stitching artifact annotation: (a) input image, (b) annotation in QuPath, and (c) corresponding dilated binary mask.

3.3. Testing dataset

Following the methodology established for the training data, the testing dataset is similarly partitioned into two subsets, ensuring a robust evaluation of the model's dual objectives. An overview of the sources of the testing dataset and is provided in Table 3.2.

Table 3.2: Sources of the artifact segmentation testing dataset.

Artifacts	Source	# WSIs
Macro	Grundium Ocus acquisitions	8
Stitching	Grundium Ocus acquisitions	5

3.3.1. Macro artifacts

The testing dataset for the macro artifacts was constructed by scanning 8 WSIs with a Grundium Ocus device, that were later manually annotated with QuPath. The preview images of the selected WSIs are shown in Figure 3.4.

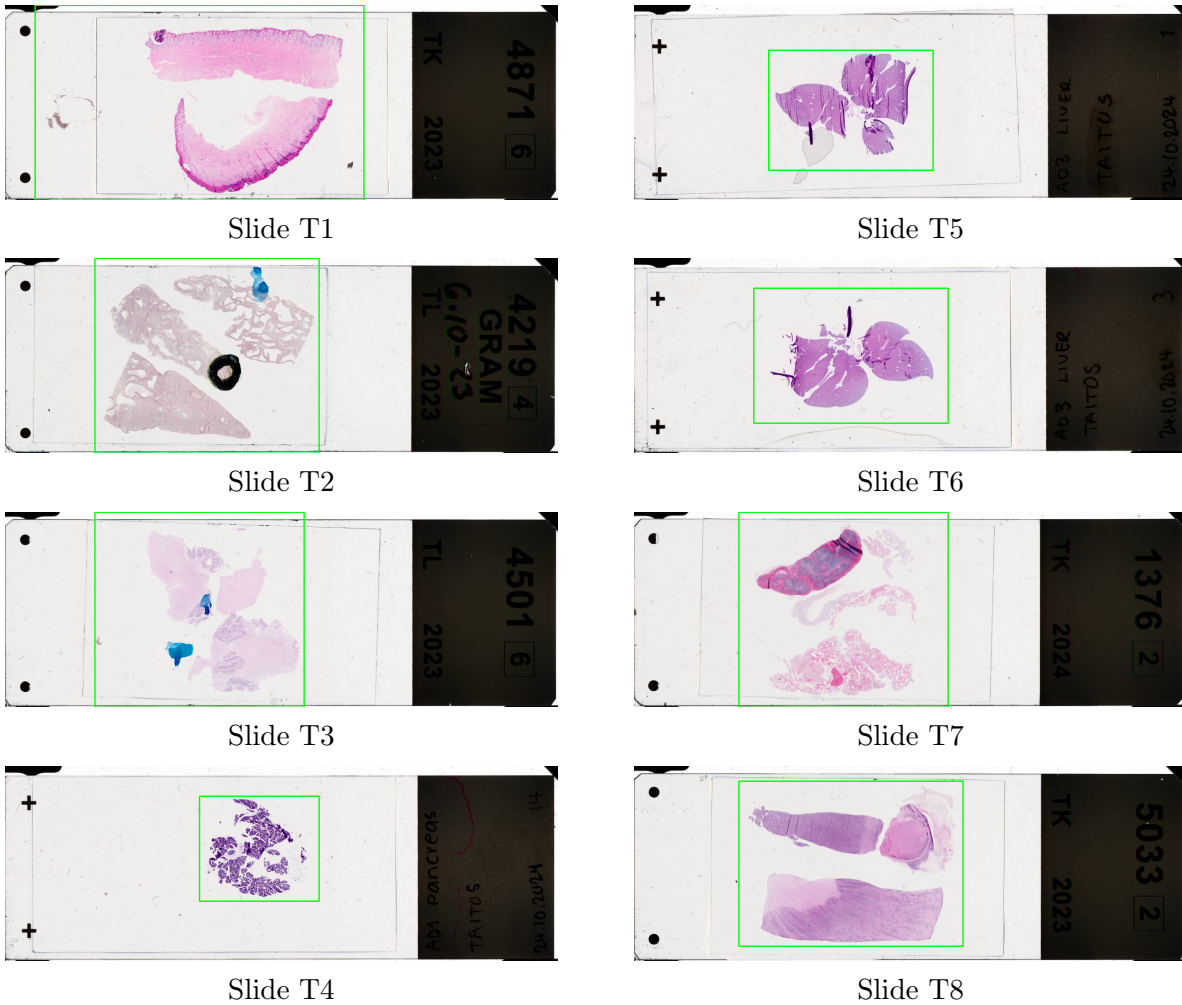


Figure 3.4: Macro artifacts testing examples

The annotation style consists in "Polygon" objects, that have been later rasterized into color-coded masks at the same resolution of the training dataset (10x magnification). The WSIs were selected to contain a variety of artifacts, including folds, pen marks, foreign objects, and out of focus regions. Only artifact classes were annotated, while

normal tissue and background were left unannotated (i.e. set to value 0 in the masks). This methodology, consistently with similar studies in the field [25, 48], was adopted to limit the annotation time, while still providing a comprehensive evaluation of the model's performance on the artifact classes. Two examples of testing samples with corresponding masks are shown in Figure 3.5.

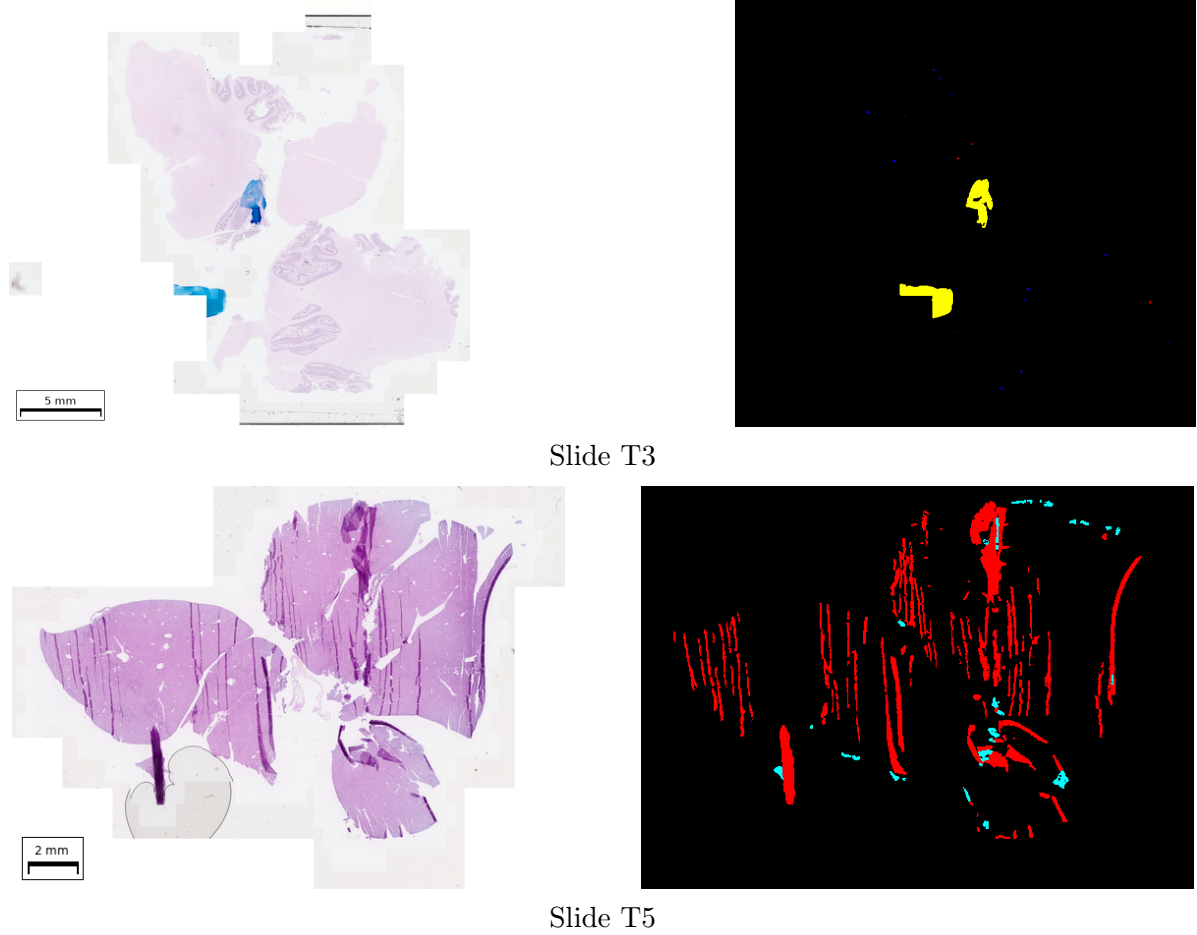


Figure 3.5: Examples of testing samples with corresponding ground truth masks.

3.3.2. Stitching artifact

The testing dataset for the stitching artifacts was generated from 5 WSIs scanned with the same Grundium Ocus device, that were later manually annotated with QuPath. The preview images of the selected WSIs are shown in Figure 3.6.

The annotation style consists in "LineString" objects, that have been later rasterized into binary masks at the same resolution of the training dataset (10x magnification), as described in Section 3.2.2. In order to streamline the annotation process and standardize the positive stitching samples selection, a semi-automated annotation approach was adopted.

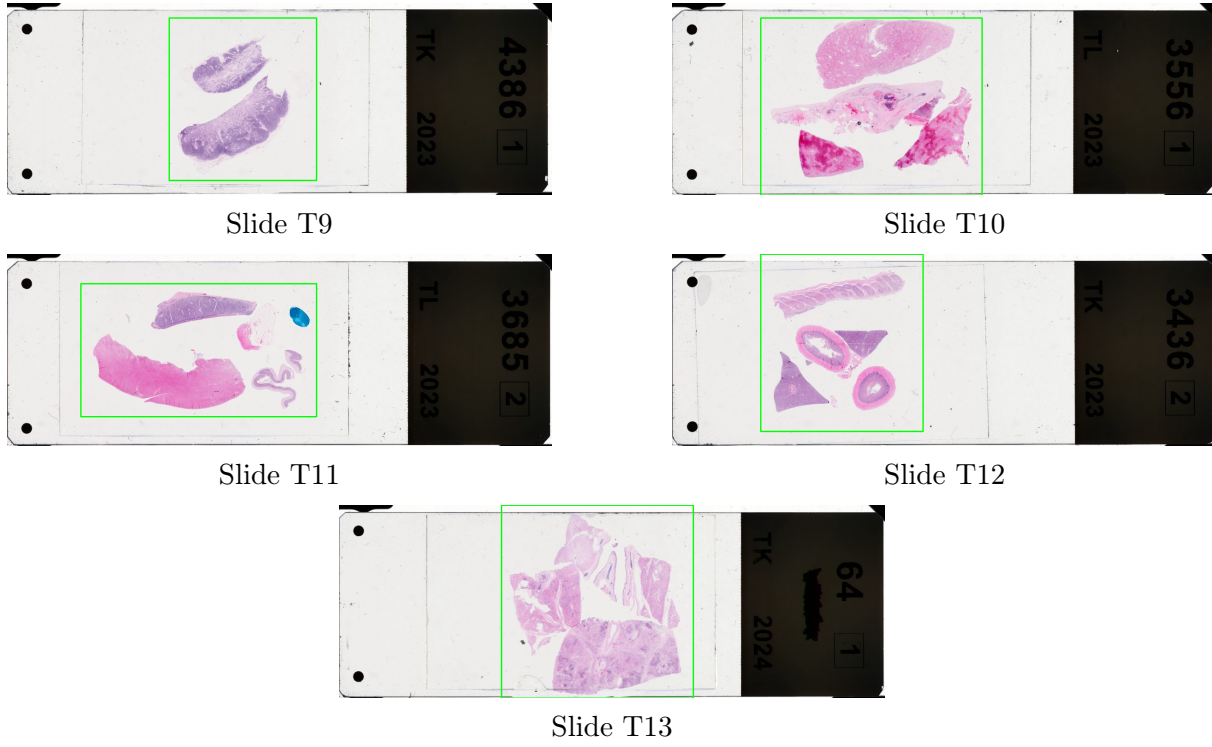


Figure 3.6: Stitching artifact testing examples.

A scanning pipeline generates three aligned acquisitions for each WSI, as suggested in the Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices guidance document of the U.S. Food and Drug Administration (FDA) [43]. The three acquisitions are as follows:

- Original scan
- OffsetX (scan region shifted by half FOV (Field of View) in x)
- OffsetY (scan region shifted by half FOV in y)

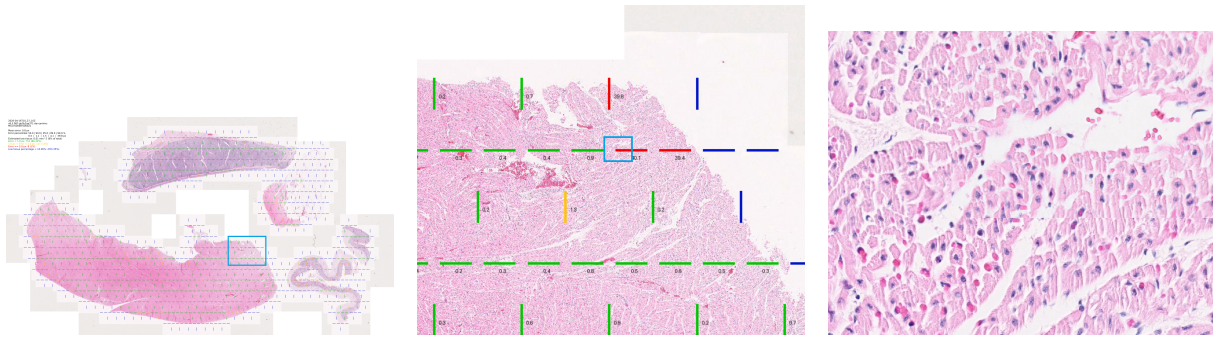


Figure 3.7: Hann-windowed phase correlation tool visualization.

Around each seam in the original scan, the pipeline extracts paired ROIs from the original

and corresponding offset scan (OffsetX or OffsetY). It then estimates local alignment shifts using normalized, Hann-windowed phase correlation. Phase correlation estimates how much one image tile is shifted relative to another by comparing them in the frequency domain. The "jump" magnitude is then used as a robust geometric error score because it reflects spatial misregistration, not simple color or texture variation. Seams with a jump magnitude above 10 pixels are flagged as positive stitching samples, establishing the ground truth for the annotation process in QuPath.

3.4. Evaluation methodology

The evaluation is performed both through quantitative metrics and qualitative analysis of the output of the networks. The quantitative evaluation has the goal of providing an objective measure of the performance of the tested models, through the use of *measurable criteria* defined in the Criteria Formulation stage of the DRM. The qualitative analysis, on the other hand, have been performed by comparing the predicted artifact masks with one or more reference models analysed in the Descriptive Study I stage (c.f. Chapter 2).

3.4.1. Macro artifacts

Quantitative performance for macro artifact detection was measured considering the following binary classification metrics:

- **Recall:** The proportion of actual artifact pixels correctly identified by the model. It is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

where TP is the number of true positive pixels and FN is the number of false negative pixels. A high recall indicates that the model is able to identify most of the artifact pixels. Recall alone, however, does not provide information about the number of false positives. A model that classifies almost all pixels as artifacts would have a high recall but would not be useful in practice.

- **Mean Intersection over Union (mIoU):** The mean IoU is a common metric for evaluating the performance of semantic segmentation models. It is calculated as the average IoU across all classes, where IoU for a single class is defined as:

$$IoU = \frac{TP}{TP + FP + FN}$$

where TP is the number of true positive pixels, FP is the number of false positive

pixels, and FN is the number of false negative pixels. In other words, IoU for a single class is defined as the area of overlap between the predicted mask and the ground truth, divided by the total area covered by both. A high mIoU indicates that the model is able to accurately segment the artifact pixels while minimizing false positives and false negatives.

- **Dice Score:** Also known as the F1 score for segmentation, this metric is highly prevalent in medical imaging due to its robustness against class imbalance. It is calculated as:

$$Dice = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Because the true positives are weighted by a factor of 2, the Dice Score is less penalizing toward minor boundary errors than IoU, making it advantageous when evaluating artifacts with highly irregular morphologies.

These metrics were calculated independently for each class and averaged across all classes to determine Balanced Accuracy, Average mIoU, and Average Dice Score.

3.4.2. Stitching artifacts

The grid-like nature of stitching artifacts, as well as their thin morphology, makes the evaluation of their detection performance particularly challenging. Predicted stitching lines are a few pixels wide due to the dilation applied to the training annotations, while ground truth annotations are rasterized as lines with a width of 1 pixel. This mismatch makes it infeasible the use of standard pixel-wise metrics such as IoU and Dice Score.

Therefore, an approach based on line coverage has been adopted. Let G represent the set of pixels comprising the 1-pixel wide ground truth annotations, and P represent the set of pixels constituting the predicted binary mask. Two metrics have been calculated for each test sample:

- **Recall:** This metric measures the proportion of the ground truth stitching line length successfully overlapped by the predicted binary mask. It is calculated as the ratio of the intersection between the prediction and the ground truth to the total area of the ground truth:

$$Recall = \frac{|P \cap G|}{|G|}$$

- **False discovery rate (FDR):** Predicted lines are isolated into distinct segments via connected component analysis. This metric calculates the proportion of pixels within the predicted mask that do not overlap with any ground truth line. It is

formulated as the relative complement of G in P over the total predicted area:

$$FDR = \frac{|P \setminus G|}{|P|}$$

where $|P \setminus G|$ denotes the number of predicted pixels that do not intersect with any ground truth line.

3.4.3. Inference speed

The inference speed of the tested models is a crucial aspect to consider, especially in the context of clinical applications where timely results are essential. The inference time is measured from the moment the model reads the input WSI until the artifact segmentation network produces the raw predictions. This includes file reading, decoding and any necessary preprocessing steps, such as downscaling and patch extraction, as well as the time taken for the model to process the input and generate the output logits. The inference time does not include the time taken for post-processing steps, such as mask overlaying or stitching artifact thresholding, as these steps are not present in all the tested models and could introduce variability that is not directly related to the model's performance. The inference speed of the tested models is assessed through the following metrics:

- **Average inference time per WSI (s):** The mean computational time required to process a complete WSI, measured in seconds, aggregated across the test set.
- **Average throughput in square millimeters per second (mm²/s):** The mean physical tissue area processed per second. This metric takes into account both the inference time and the size of the WSIs, providing a more comprehensive measure of the model's performance, taking into account the inherent variability in tissue area across different WSIs.

4 | Implementation

4.1. Experimental setup

The experiments were conducted on a laptop equipped with an Intel(R) Core(TM) Ultra 7 265H (16 CPUs), 64 GB of RAM, and an NVIDIA RTX PRO 1000 Blackwell Generation GPU with 8 GB of VRAM. The models were tested in a WSL (Windows Subsystem for Linux) environment running Ubuntu 24.04. The implementation was done using Python 3.12.

Artifact detection in computational pathology is generally formulated as either a classification or a segmentation problem. Within these categories, researchers can adopt patch-based or full-image approaches. Because feeding entire whole slide images (WSIs) into a deep learning model is computationally infeasible due to their large size, a patch-based approach is standard practice [23, 25, 35, 49]. The WSI is divided into smaller patches processed independently, which reduces computational load and enables efficient training and inference. Furthermore, a pixel-wise segmentation approach provides granular data regarding the spatial location and extent of artifacts. Inspired by prior literature [35, 49], a two-stage approach was adopted as architectural choice. In the first stage, a tissue segmentation network was trained at a lower magnification level (i.e., 1x) to identify tissue regions and discard the background, minimizing computational overhead. In the second stage, an artifact segmentation network was utilized to process only the tissue-containing patches at a higher magnification level (i.e., 10x). This pipeline ensures that computational resources are allocated exclusively to relevant WSI regions. Both these networks are based on the SegFormer architecture [51], which is a state-of-the-art model for semantic segmentation tasks. A visual representation of the proposed two-stage approach is shown in Figure 4.1. The exact architecture and training details of these networks are described in the following sections.

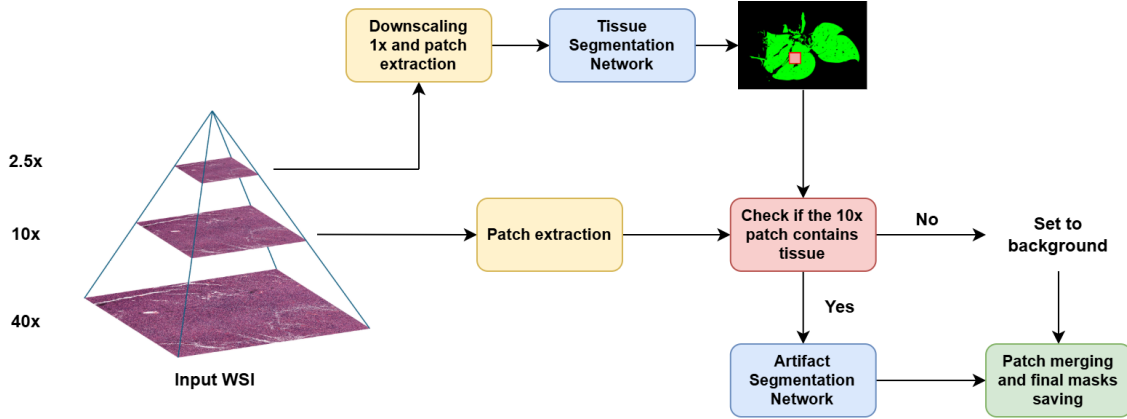


Figure 4.1: Two-stage pipeline adopted for tissue and artifact segmentation.

4.2. Macro artifacts dataset pre-processing

By visually inspecting the available masks, we identified the presence of bad quality segmentations, such as wrong detected artifacts and grid-like patterns. A manual revision of the masks was performed to remove the most evident errors and limit the error propagation from the teacher to the student model. This methodology, often referred to as Human-in-the-Loop (HITL), is a common practice in knowledge distillation to ensure the quality of the generated training data. HITL development consists in the integration of human intelligence, oversight, and judgment into the machine learning lifecycle. In the work of Wu et al. [50], this supervision can be categorized as Human-in-the-loop data processing. Examples of discarded masks are presented in Figure 4.2.

After the revision, a total of 75 WSIs were selected for the training phase. The TCGA source projects of the selected WSIs are listed in Table 4.1.

Analysis of the dataset’s class distribution revealed that the background class comprises over half of a typical slide image, which holds no relevance for artifact detection.

From the 75 selected WSIs, two distinct datasets were built for the training of the tissue segmentation network and the artifact segmentation network, respectively. The tissue segmentation network was trained on a dataset of 512 x 512 pixel patches extracted at 1x magnification using the OpenSlide library and paired with the corresponding artifact mask at the same magnification, while the artifact segmentation network was trained on a dataset of 512 x 512 pixel patches extracted at 10x magnification. For the tissue segmentation network, the binary masks were generated by collapsing the "Normal tissue" and all the artifact classes into a unified "Tissue" class, while the "Background" class was left unchanged. A total of 4,862 patches were generated for the tissue segmentation

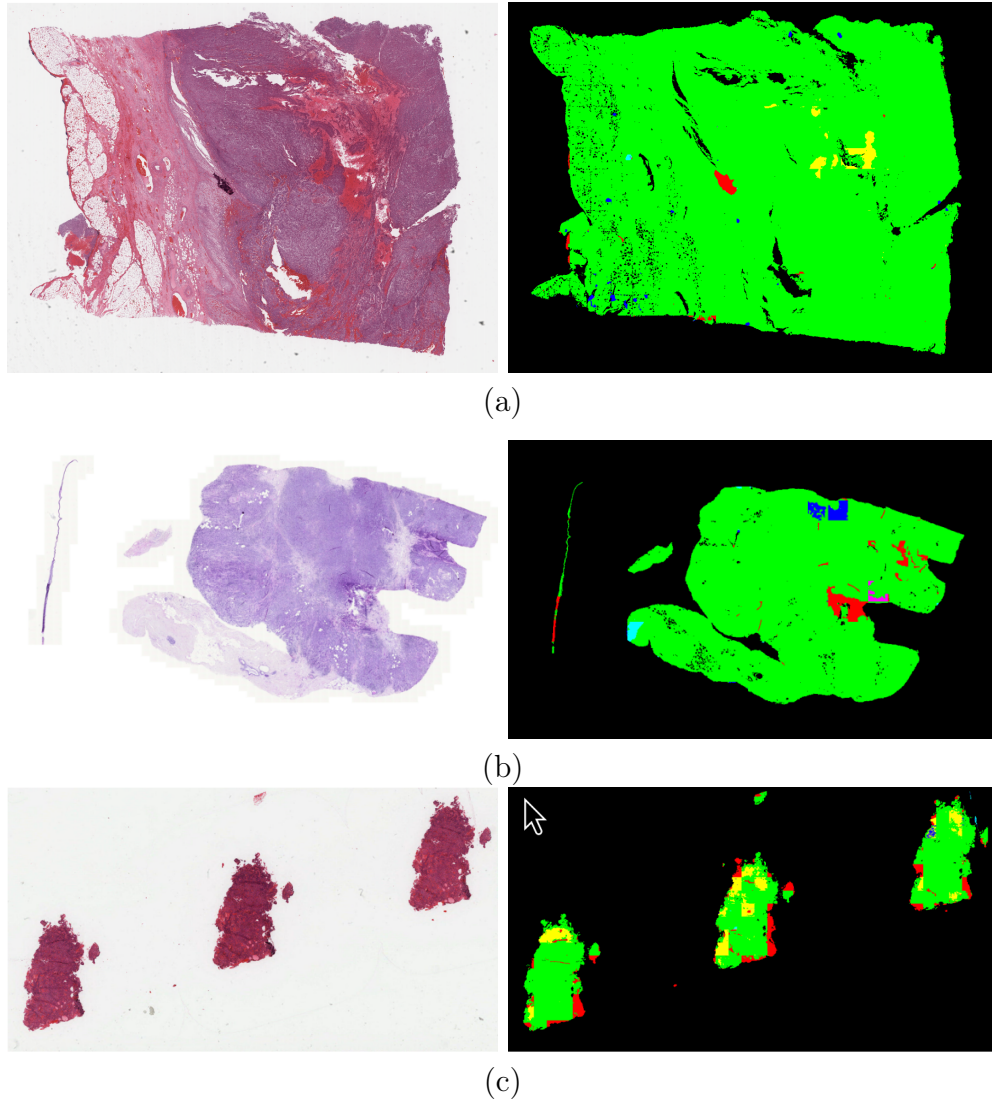


Figure 4.2: Examples of discarded masks: (a) mask with false positive pen marking (*yellow*), (b) mask with false positive foreign object (*blue*), and (c) mask with grid-like pattern of false positives artifacts (*yellow* and *red*).

Table 4.1: Source projects of the selected training WSIs.

Project	Masks
Adrenocortical Carcinoma (ACC)	14
Breast Invasive Carcinoma (BRCA)	8
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC)	11
Cholangiocarcinoma (CHOL)	5
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC)	2
Esophageal Carcinoma (ESCA)	4
Glioblastoma Multiforme (GBM)	4
Head and Neck Squamous Cell Carcinoma (HNSC)	9
Kidney Renal Clear Cell Carcinoma (KIRC)	4
Liver Hepatocellular Carcinoma (LIHC)	7
Pancreatic Adenocarcinoma (PAAD)	5
Pheochromocytoma and Paraganglioma (PCPG)	2
Total	75

network training dataset.

A total of 81,532 patches were generated for the artifact segmentation network training dataset. The class distribution of the artifact segmentation training dataset is reported in Table 4.2.

Table 4.2: Training class distribution measured on labeled pixels.

Class	Labeled pixels (%)
Background	52.14
Tissue normal	46.17
Fold	0.63
Foreign object	0.14
Pen marker	0.70
Out of focus	0.22

To address severe class imbalance issue, multiple mitigation strategies were implemented:

- **Syntetic out of focus augmentation:** Due to the low representation of the "Out of focus" class (0.22%), a synthetic augmentation method was implemented to increase the number of training samples for this class. Applied to 2.5% of strictly clean, artifact-free regions, the algorithm relies on a custom shape-generation function to avoid unrealistic geometric blurs. It constructs organic, stain-like topologies by aggregating multiple randomized anisotropic Gaussians, introducing high-frequency noise for edge irregularity, and refining the boundaries via morphological closing and opening operations. The resulting binary mask undergoes heavy Gaussian

smoothing to create a feathered alpha channel, allowing for a biologically plausible, soft-edged transition between the blurred injection and the sharp, native tissue. The corresponding ground-truth segmentation masks are simultaneously updated to incorporate these synthetic artifact classes. Figure 4.3 demonstrates this process.

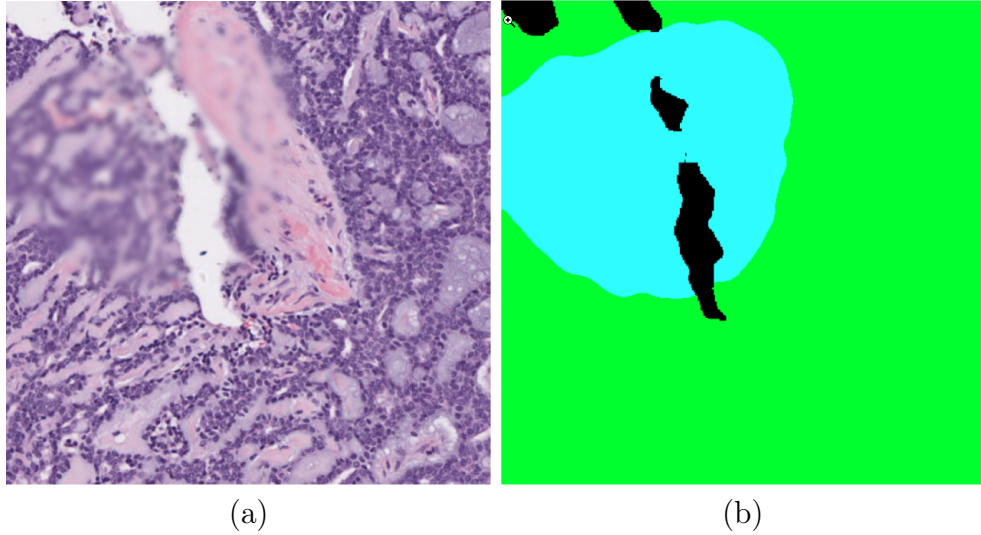


Figure 4.3: Example of synthetic out of focus augmentation: (a) original image with synthetic blur applied, and (b) corresponding updated artifact mask (*green*: normal tissue, *cyan*: out of focus).

- **Undersampling:** The training dataset was undersampled based on the following rules:
 - 95% of the patches containing more than 95% of background pixels were discarded.
 - 98% of the patches containing only background and normal tissue pixels were discarded.
- **Weighted loss:** A weighted loss function was utilized during training to prioritize minority classes. Weights were calculated based on the inverse frequency of each class in the training dataset. Further details about the loss function and the weight calculation are provided in Section 4.5.

After applying these balancing strategies, the final training dataset for macro artifacts comprised 17,928 patches. The adjusted class distribution is shown in Table 4.3.

These patches have been merged with the stitching artifacts training dataset described in 3.2.2, resulting in the final training dataset reported in Table 4.4.

Table 4.3: Training class distribution measured on labeled pixels before and after applying the balancing strategies.

Class	Labeled pixels (%) - before	Labeled pixels (%) - after
Background	52.14	20.89
Tissue normal	46.17	66.42
Fold	0.63	3.95
Foreign object	0.14	0.83
Pen marker	0.70	5.13
Out of focus	0.22	2.78

Table 4.4: Summary of the artifact segmentation training dataset.

Artifacts	Source	# WSIs	# Extracted patches
Macro	GrandQC	75	17,928
Stitching	Grundium Ocus acquisitions	12	504
Final Training Dataset	Merged	87	18,432

4.3. Tissue segmentation network

The tissue segmentation network has the goal of correctly detecting the portion of the image that contains actual tissue, in order to focus the artifact segmentation network only on the relevant areas of the image. The architecture of the tissue segmentation network is a SegFormer with a MiT-B0 encoder and a single decoder head. The network is trained using a binary cross-entropy loss function, where the target is a binary mask indicating the presence of tissue in the image. The output of the tissue segmentation network is a binary mask that highlights the regions of the image that contain tissue, which is then used as input for the pre-processing step of the artifact segmentation network. Once a good baseline performance has been achieved within the tissue segmentation stage, the focus of the project has shifted on the artifact segmentation network described in the next section, which is the main contribution of this work.

4.4. Artifact segmentation network

The artifact segmentation network is designed to detect and segment both macro-artifacts and stitching artifacts in the histopathological images. The development of this network followed a top-down approach, starting with a simpler architecture and gradually adding interventions to improve performance and support for more types of artifacts. Iterations of the artifact segmentation network were developed in line with the DRM methodology (cf. Section 3.1), with each iteration introducing new features and/or architectural

changes based on the performance of the previous version and advances in the literature. Incremental versions of the artifact segmentation network are tracked in the following Sections 4.4.1-4.4.3 and the final architecture is described in detail in Section 4.5.

4.4.1. Standard SegFormer

The first iteration of the artifact segmentation network was a standard SegFormer architecture with a MiT-B0 encoder and a single decoder head. The network was trained using a cross-entropy loss function, where the target is a multi-class mask indicating the presence of macro-artifacts inside the 512x512 image patches. The output of the network is a multi-class mask that highlights the regions of the image that contain macro-artifacts, normal tissue, and background. The predictions of the single patches are then stitched together to create a full-size mask for the whole WSI. The architecture of the standard SegFormer is shown in Figure 4.4. The results obtained with this architecture were used as a baseline for the subsequent iterations of the artifact segmentation network, as well as to measure the effectiveness of the distillation from the teacher model to the student model.

4.4.2. Multiscale SegFormer

The second iteration of the artifact segmentation network was a multiscale SegFormer architecture, which is an extension of the standard SegFormer architecture. This architecture takes as input not only the original 512x512 image patch at 10x magnification, but also a concentric 5x magnification patch centered on the same location. In this way, the network can exploit both the local information from the 10x patch and the broader context from the 5x patch. This idea is motivated by the different morphological characteristics of the different artifacts: within the same WSI, we could find both small artifacts, such as dust or small out of focus areas, and large artifacts, such as tissue folds. The small artifacts are better detected at higher magnification, while the large artifacts are better detected at lower magnification. An example of a tissue fold at both 10x and 5x magnification is shown in Figure 4.5, where the larger folding structure is more clearly visible in the 5x view.

The architectural variation introduced in the multiscale SegFormer is inspired by the multi-resolution approach applied to Convolutional Neural Networks for breast cancer analysis proposed by Van Rijthoven et al. [44], where a dual-branch parallel architecture is used to process the two magnification levels separately; the extracted features from the first branch are duplicated and concatenated to the second branch. In HookNet [44], two

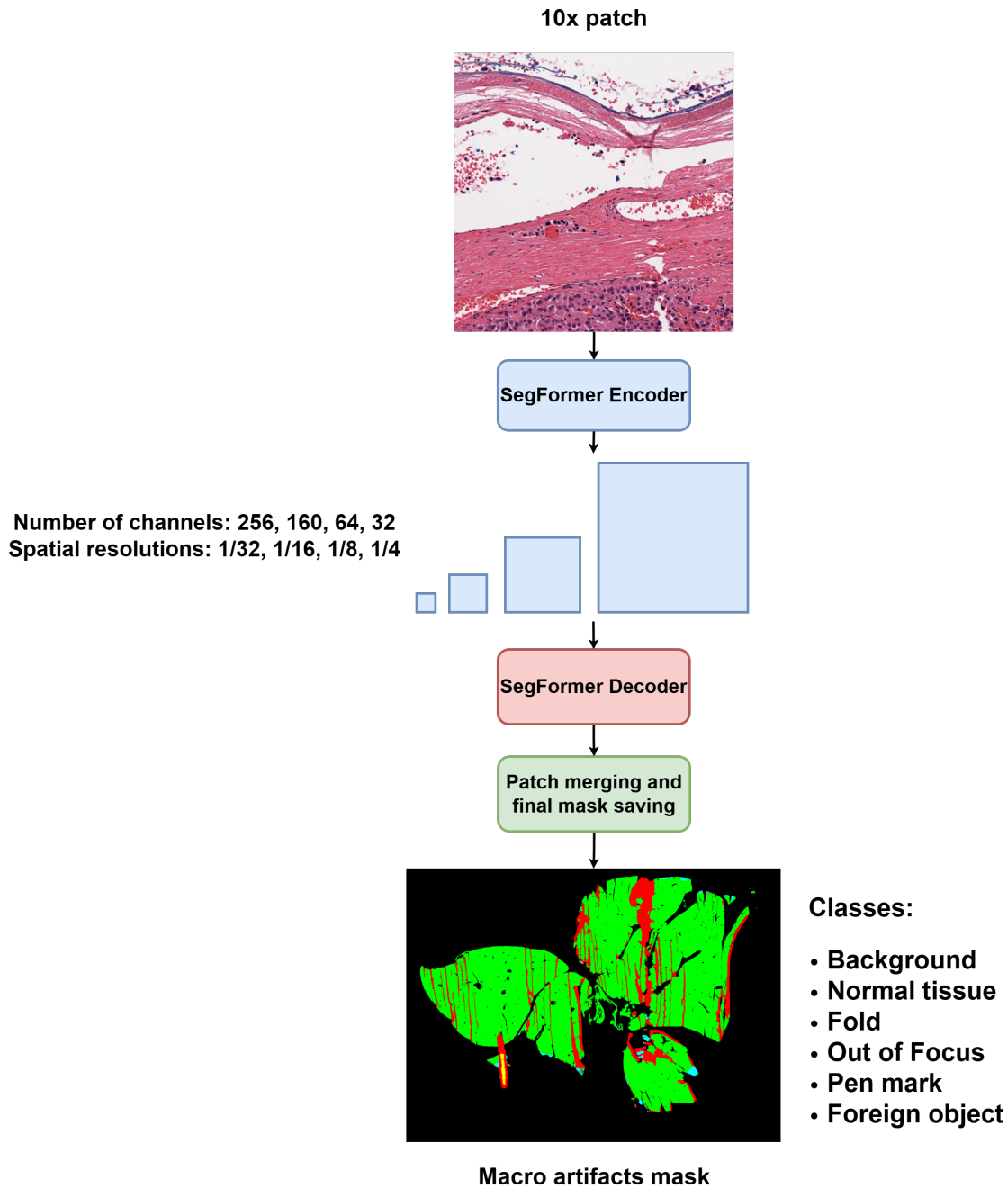
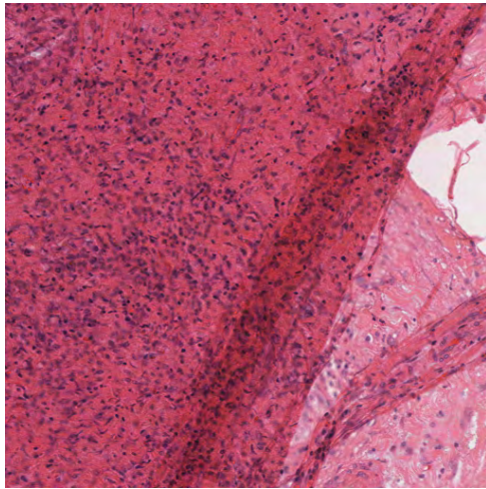
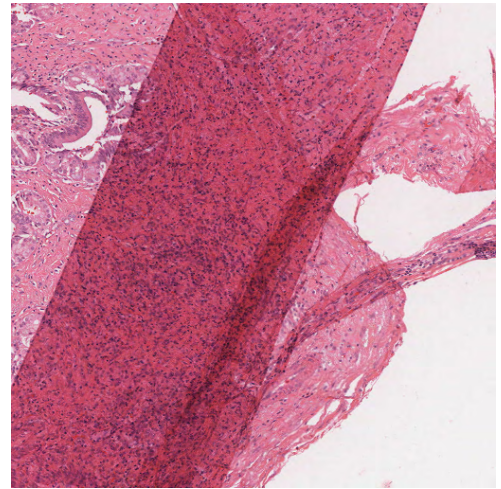


Figure 4.4: Overview of the Standard SegFormer architecture. The network consists of a hierarchical Mix Transformer (MiT) encoder that extracts multi-scale features from the input image using overlapping patch merging, dynamically encoding spatial information without the need for strict positional encodings. These multi-resolution feature maps (at 1/4, 1/8, 1/16, and 1/32 scales) are subsequently passed to a lightweight All-MLP decoder. The decoder unifies the channel dimensions, upsamples the features to a common resolution, and concatenates them to generate the final high-resolution semantic segmentation mask.



(a) 10x magnification crop



(b) 5x magnification crop

Figure 4.5: Example tissue fold at two magnifications. The 5x view helps reveal the larger folding structure more clearly, supporting the multiscale architecture motivation.

loss functions are applied to the two branches, one for each magnification level, and the final loss is a weighted sum of the two losses, that are jointly optimized during training.

The multiscale SegFormer architecture translates the parallel encoder design to Vision Transformers, but with a single shared decoder head. This solution ensures that only one macro-artifact segmentation mask is generated as output and allows to train the network with a single loss function, without the need to define a separate loss for each branch and to find the optimal weights for the two losses. The architecture of the multiscale SegFormer is shown in Figure 4.6. The network consists of two parallel branches, one for each magnification level, with two separate MiT-B0 encoders. The feature maps extracted from the 5x branch (green) are concatenated to the feature maps extracted from the 10x branch (blue). In order to match the region of interest of the 10x patch, the 5x feature maps are centerly cropped and then upsampled to the same spatial resolution of the 10x feature maps. The concatenated feature maps are then passed to a single shared decoder head, which generates the final multi-class segmentation mask.

4.4.3. Multitask SegFormer

The third iteration of the artifact segmentation network is a multitask SegFormer architecture, which is an extension of the multiscale SegFormer architecture. This architecture, reported in Figure 4.7, is designed to detect and segment both macro-artifacts and stitching artifacts in the histopathological images. The network complements the macro-artifact segmentation head with a second decoder head, that is in charge of detecting stitching

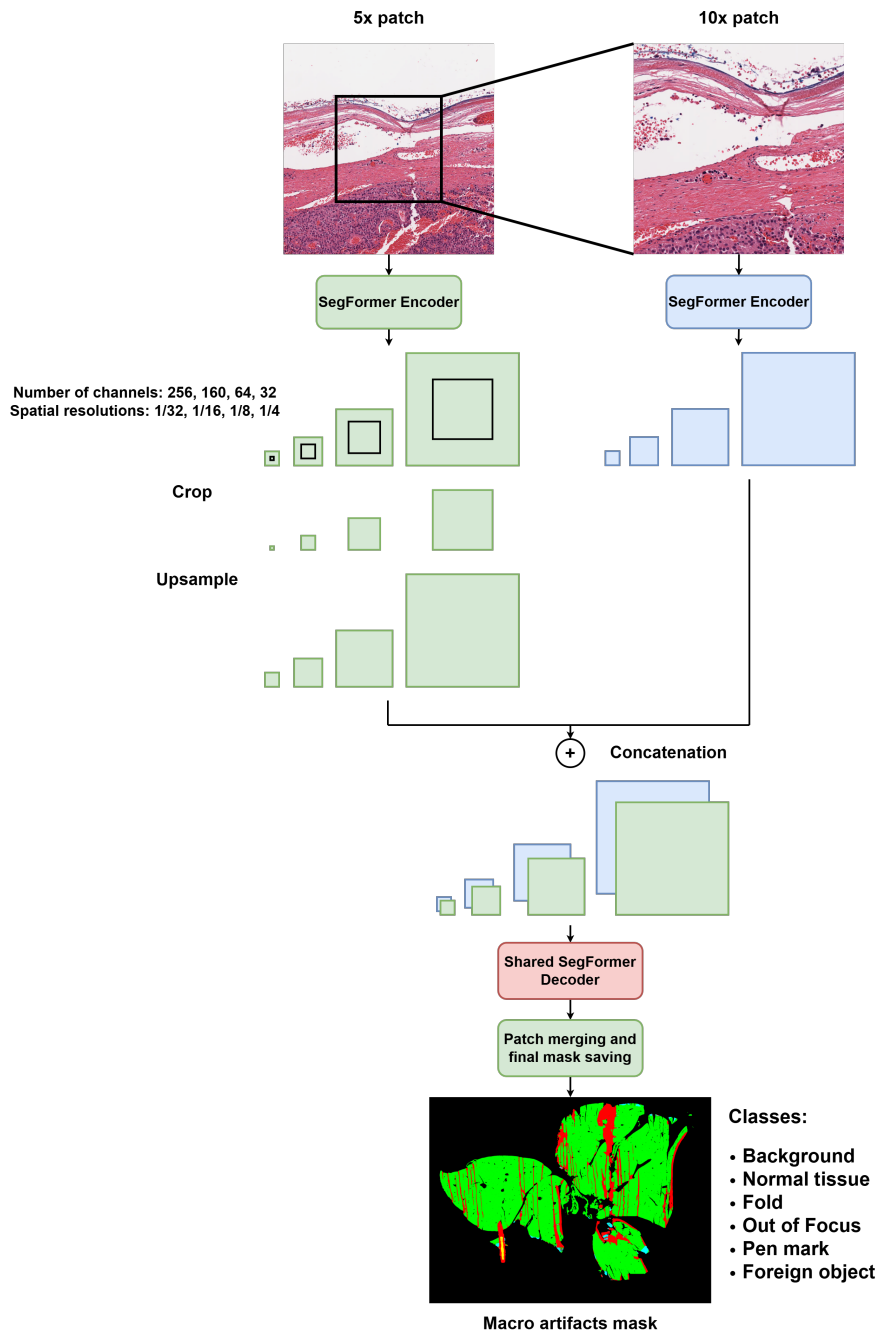


Figure 4.6: Overview of the Multiscale SegFormer architecture. The network consists of two parallel branches, one for each magnification level, with two separate MiT-B0 encoders. The features maps extracted from the 5x branch (green) are concatenated to the features maps extracted from the 10x branch (blue). In order to match the region of interest of the 10x patch, the 5x feature maps are centerly cropped and then upsampled to the same spatial resolution of the 10x feature maps. The concatenated feature maps are then passed to a single shared decoder head, which generates the final predictions.

artifacts. The stitching artifact segmentation head takes as input the higher resolution features extracted from the 10x branch ($1/4$ scale), as the stitching artifacts are generally small and require a higher magnification to be detected. It consists of a lightweight CNN decoder that passes the output of a Strip Pooling module through two sequential 3×3 convolution layers with padding=1. Each convolution is followed by batch normalization and a ReLU activation, forming a two-layer convolutional refinement block. Finally, a 1×1 convolution maps the refined 64-channel feature map to a single stitch logit channel. The logits are then passed to a sigmoid activation function to produce probabilities for the presence of stitching artifacts at each pixel location. Those probabilities are then thresholded and passed through a post-processing module performing morphological operations to generate the final binary mask for stitching artifacts. The architecture of the stitching head is shown in Figure 4.7.

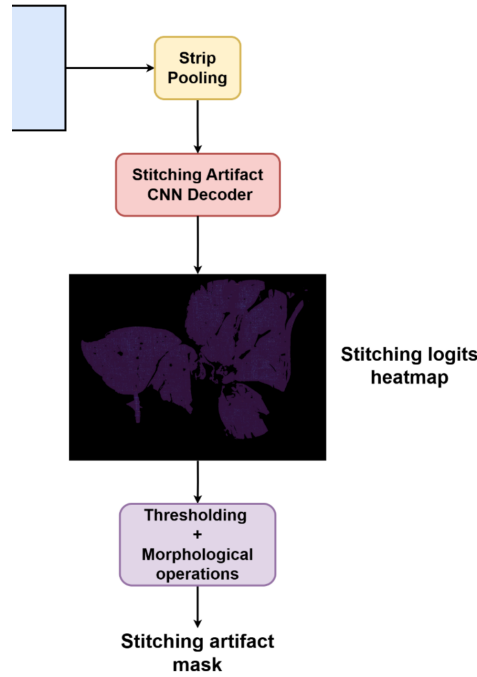


Figure 4.7: Stitching artifact head. It takes as input the higher resolution features extracted from the 10x branch ($1/4$ scale) and leverages a thresholding + morphological operations module to refine the final mask.

Particularly beneficial for the detection of stitching artifacts is the use of a tailored pooling strategy. The Strip Pooling module, introduced by Hou et al. [17], addresses the limitations of conventional spatial pooling methods that rely on square receptive fields, which often struggle to capture anisotropic or elongated object structures. To overcome this, the Strip Pooling module employs two parallel one-dimensional average pooling operations—

one horizontal ($1 \times W$) and one vertical ($H \times 1$)—to aggregate contextual information along distinct spatial dimensions. These strip-like pooled features are subsequently expanded back to their original 2D dimensions, fused together, and passed through a sigmoid activation to generate a refined spatial attention map. By multiplying this attention map element-wise with the original input features, the module effectively captures long-range dependencies and complex structural details, significantly improving the semantic segmentation of continuous objects like roads, rivers, and pillars. In the context of stitching artifact detection, the Strip Pooling module enhances the model’s ability to identify long lines spanning across big portions of the image. The architecture of the Strip Pooling module is shown in Figure 4.8.

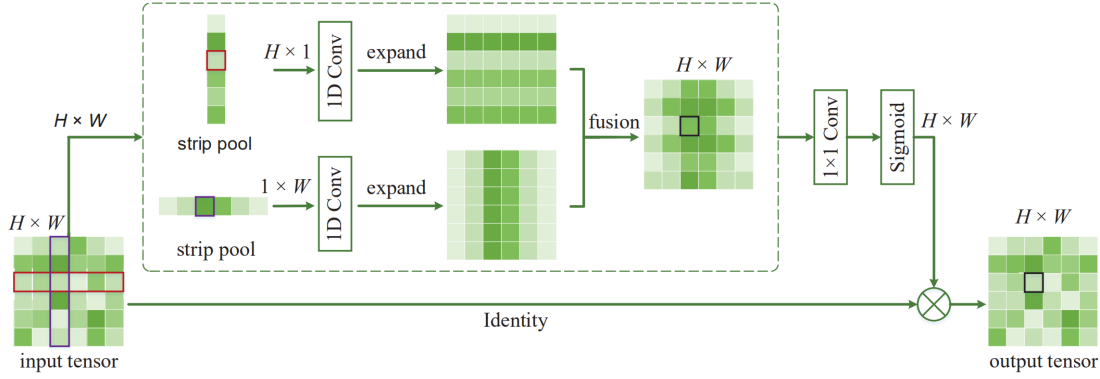


Figure 4.8: **Strip pooling module.** Schematic illustration of the strip pooling module used in the stitching artifact decoder [17].

4.5. Overview of the proposed architecture

The final architecture of the artifact segmentation network is a multitask multiscale SegFormer, which combines the features illustrated in the previous sections. The network consists of two parallel branches, one for each magnification level, with two separate MiT-B0 encoders. The features maps extracted from the 5x branch are concatenated to the features maps extracted from the 10x branch and passed to a shared decoder head for macro-artifact segmentation. The higher resolution features extracted from the 10x branch are also passed to a second decoder head for stitching artifact segmentation. The architecture of the proposed multitask multiscale SegFormer is shown in Figure 4.9.

Training the multitask multiscale architecture relies on a composite loss function that explicitly decouples general multi-class tissue segmentation from the specialized task of binary stitch artifact detection. The total loss, $\mathcal{L}_{total}$, is formulated as a linear combination

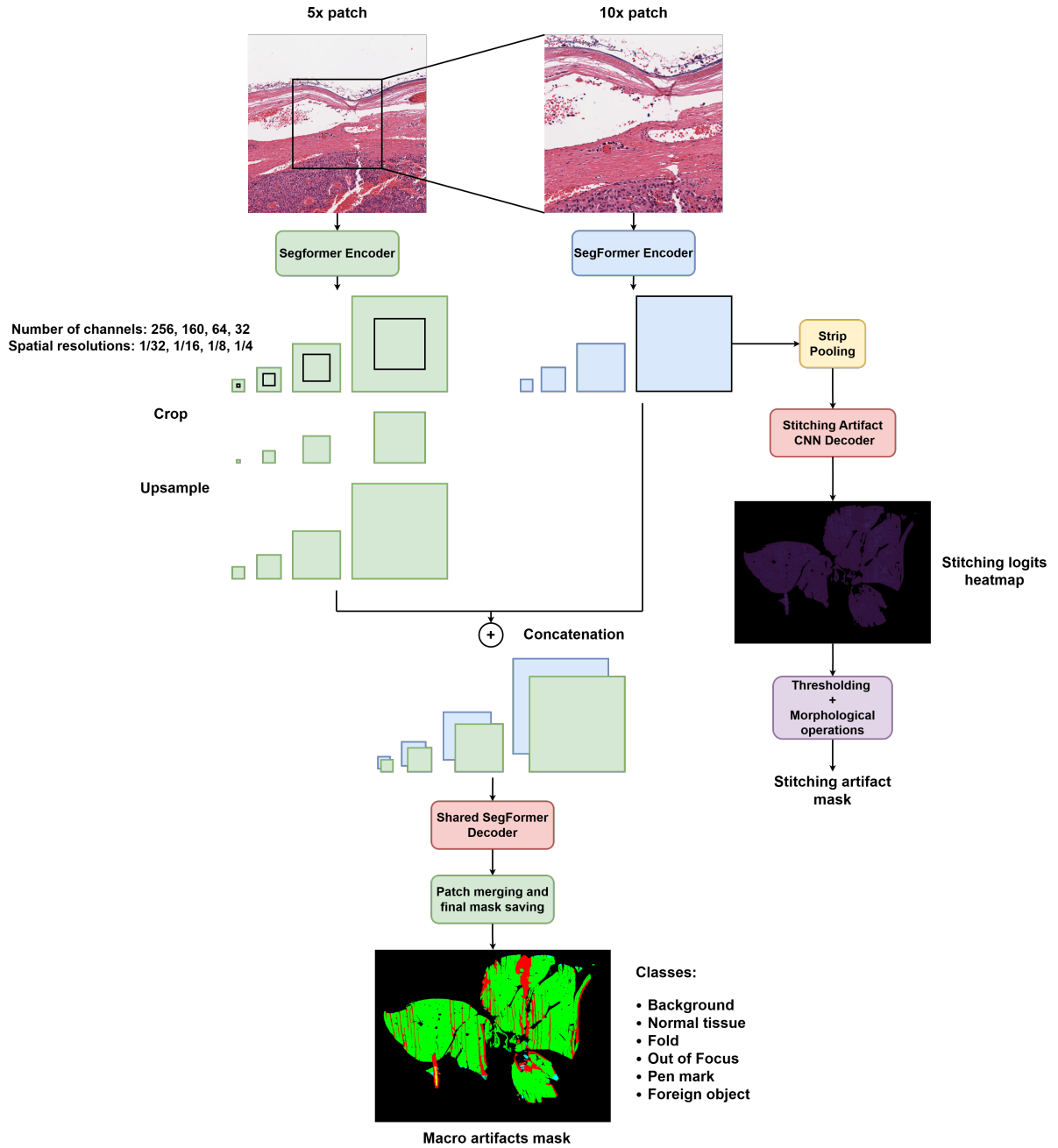


Figure 4.9: Overview of the proposed Multitask Multiscale SegFormer architecture.

of a macro-level and a stitch-level loss:

$$\mathcal{L}_{total} = \mathcal{L}_{macro} + \lambda_{stitch}\mathcal{L}_{stitch} \quad (4.1)$$

where λ_{stitch} is a weighting coefficient empirically set to 1.0.

To address the severe class imbalance typical of whole-slide images, the macro branch employs a class-weighted Cross-Entropy loss ($\mathcal{L}_{macro}$). To guarantee sufficient gradient signals for underrepresented classes, the per-class weights w_c are calculated using normalized inverse-square root pixel frequencies ($w_c \propto 1/\sqrt{N_c}$), where N_c is the total pixel count for class c in the training set, and are clamped to the interval $[0.0, 4.0]$ to maintain optimization stability. The resulting class weights are reported in Table 4.5.

Table 4.5: Weights assigned to each class in the macro-artifact segmentation branch, calculated based on the pixel distribution in the training set.

Class	Weight
Background	0.438
Tissue normal	0.244
Fold	1.015
Foreign object	2.236
Pen marker	0.872
Out of focus	1.196

Conversely, the stitch branch isolates localized stitching artifacts through a combined loss function designed to simultaneously address positive-negative pixel imbalance and spatial overlap. Specifically, the stitch objective is the unweighted sum of a Binary Focal Loss [31] and a Dice Loss [9, 40] ($\mathcal{L}_{stitch} = \mathcal{L}_{focal} + \mathcal{L}_{dice}$). The Focal Loss component prevents overwhelming background dominance by dynamically down-weighting well-classified pixels. It is defined as:

$$\mathcal{L}_{focal} = \alpha_t(1 - p_t)^\gamma \text{BCE}(p, y) \quad (4.2)$$

where p represents the predicted sigmoid probability, $y \in \{0, 1\}$ is the binary ground truth, and $p_t = p \cdot y + (1 - p)(1 - y)$. The focusing parameter is set to $\gamma = 2.0$, and the weighting factor α_t is derived using a balancing parameter $\alpha = 0.5$.

This is complemented by the Dice Loss, which directly optimizes region-based spatial agreement:

$$\mathcal{L}_{dice} = 1 - \frac{2 \sum (p \cdot y) + \epsilon}{\sum p + \sum y + \epsilon} \quad (4.3)$$

where a smoothing term $\epsilon = 10^{-6}$ is incorporated to preserve numerical stability.

Finally, to enforce distinct feature learning and prevent gradient interference between the network branches, a strict spatial masking strategy is applied during training. The macro loss $\mathcal{L}_{macro}$ is exclusively evaluated on image patches devoid of stitch pixels, whereas the stitch loss $\mathcal{L}_{stitch}$ is computed solely on patches that contain the targeted stitch artifacts.

The **morphological operations module** applied to the output of the stitching head (Figure 4.10) is designed to refine the raw predictions of the network and generate a clean binary mask for stitching artifacts. The module consists of a sequence of morphological operations, including closing, opening, and dilation, and a connected component analysis to filter out small false positive detections.

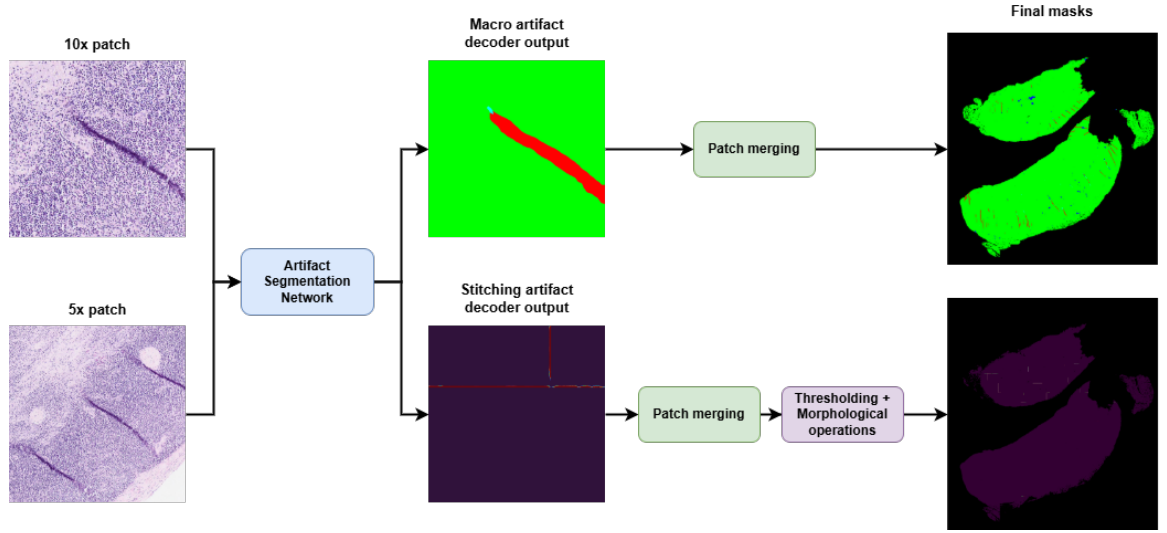


Figure 4.10: Proposed artifact segmentation network pipeline.

The stitch decoder produces a probability map that is first converted into a binary candidate mask by applying a fixed probability threshold: pixels with predicted stitch probability greater than or equal to the selected threshold (i.e. 0.90) are assigned to the stitch class, while all other pixels are suppressed. The resulting binary mask is then refined with a morphology-based pipeline designed for the expected geometry of stitch artifacts, which typically appear as thin, elongated horizontal or vertical line structures. First, a directional closing operation is applied separately along the vertical and horizontal axes. Long oriented closing kernels are used to bridge small gaps between fragmented stitch predictions. The vertical and horizontal results are then merged, preserving line continuity while avoiding isotropic expansion of unrelated regions. Second, a directional opening step removes small or weakly oriented noise by retaining structures that survive long vertical or horizontal opening kernels. Finally, connected-component filtering removes blob-like detections: components are kept only when they have an elongated or sparsely occupied bounding-box geometry, consistent with linear stitch artifacts. Overall,

this post-processing converts the raw decoder probabilities into a cleaner binary stitch mask by improving continuity of fragmented linear predictions, and suppressing compact false-positive regions. Figure 4.11 illustrates the morphological operations pipeline.

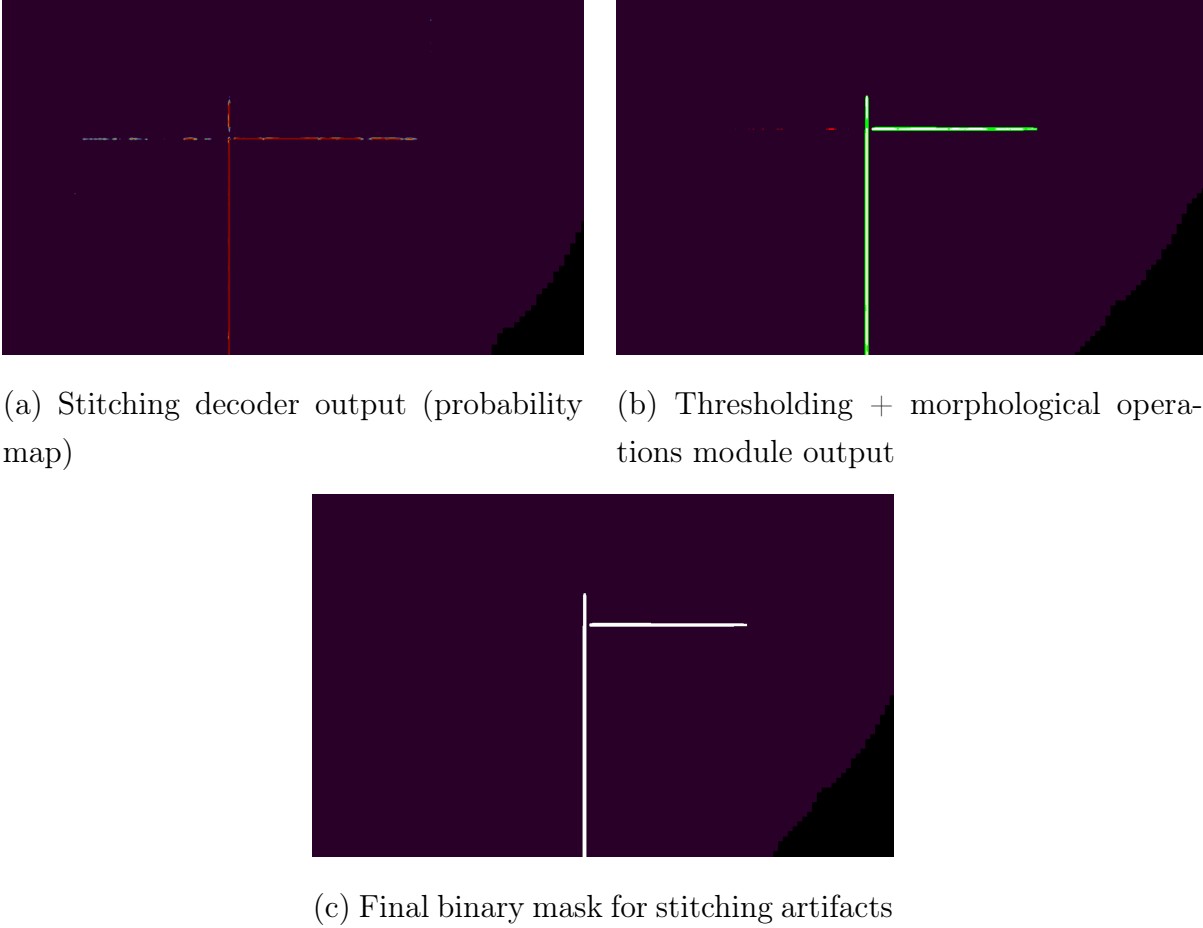


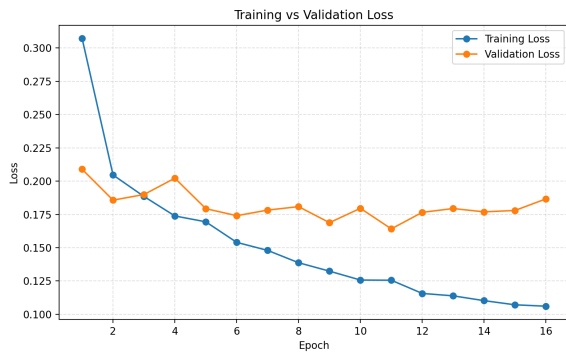
Figure 4.11: Morphological post-processing pipeline for stitching artifact detection. The raw output of the stitching decoder (a) is a probability map that is first thresholded to generate a binary candidate mask. This mask is then refined with a sequence of morphological operations designed to enhance the continuity of linear stitch predictions and suppress compact false positives. (b) shows the intermediate step after thresholding and morphological operations: *white* pixels represent the thresholded candidate stitch predictions, *green* pixels represent the result of the directional closing operation, and *red* pixels represent the result of the directional opening operation and connected-component filtering. (c) shows the final binary mask for stitching artifacts after the post-processing pipeline.

For reproducibility, the training hyperparameters used for all the iterations of the artifact segmentation network are reported in Table 4.6. The training and validation loss curves

for all the iterations of the artifact segmentation network are shown in Figure 4.12.

Table 4.6: Default hyperparameters utilized for the network training phase.

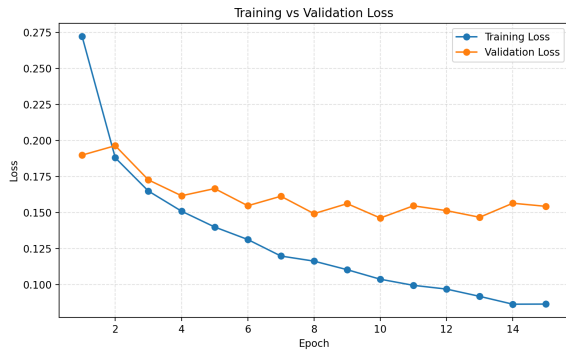
Parameter	Value
Epochs	20
Early stopping patience	5
Patch size	512x512
Batch size	8
N. patches	18567
Optimizer	AdamW
Learning rate	6e-5



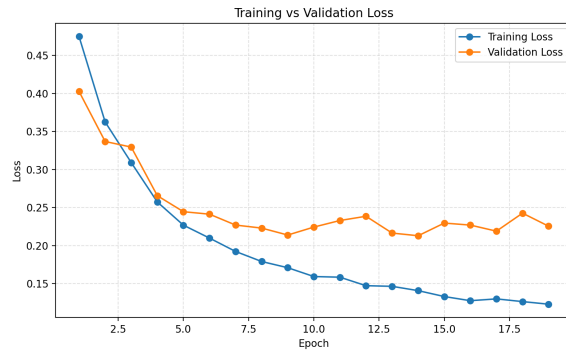
(a) Standard SegFormer



(b) Multiscale SegFormer (10x - 2.5x)



(c) Multiscale SegFormer (10x - 5x)



(d) Multitask Multiscale SegFormer (10x - 5x)

Figure 4.12: Training vs Validation Loss Curves

5 | Results

This chapter evaluates the performance of the models tested throughout the development phase. We present the results separately for the macro artifacts test set (T1-T8) and the stitching artifact test set (T9-T13). The results are collected both per-slide and aggregated across the test sets. Finally, we analyze the inference speed of the proposed architecture and compare it with GrandQC and the different iterations of our models. Model efficacy is assessed quantitatively using the *measurable criteria* defined in Chapter 3 and qualitatively through visual inspection of the predicted segmentation masks.

5.1. Evaluation of macro artifacts

In this section, we present the results of the proposed method for macro artifact segmentation on the macro artifacts test set (T1-T8). The evaluation is performed separately for each slide and then aggregated across the test set. It includes a comparison of the proposed method with GrandQC and a standard SegFormer model, as well as an analysis of the impact of multiscale information and the integration of the stitch detection branch on the performance of macro artifact segmentation. The output masks of the models are visually compared with the ground truth and two benchmark methods (GrandQC and HistoQC) for each slide in the test set.

Table 5.1 summarizes the aggregated performance of the models on the macro artifacts test set (T1-T8). Average artifact balanced accuracy, average mean Intersection over Union (mIoU), and average Dice are computed as the mean of artifact balanced accuracy, mIoU and Dice across the 8 slides. Within a slide, artifact balanced accuracy, mIoU and Dice are computed as the mean recall, IoU and Dice across the artifact classes present in the ground truth.

The first two rows of the table offer a comparison between GrandQC and a standard SegFormer model, both working at 10x magnification. The standard SegFormer model outperforms GrandQC across all the metrics, with a notable improvement in average mIoU (0.429 vs 0.358) and average Dice (0.548 vs 0.431). Rows 3 and 4 present the results of

Table 5.1: Model aggregated performance results - Macro artifacts test set (T1-T8)

Model	Input magnification	Stitch Detection	Avg Artifact Balanced Accuracy	Avg mIoU	Avg Dice
GrandQC	10x	N	0.744	0.358	0.431
Standard SegFormer	10x	N	0.762	0.429	0.548
Multiscale SegFormer	10x - 2.5x	N	0.761	0.532	0.666
Multiscale SegFormer	10x - 5x	N	0.890	0.483	0.608
Multiscale SegFormer	10x - 5x	Y	0.849	0.512	0.636

the multiscale SegFormer model at different magnification combinations. Both multiscale configurations outperform the single resolution SegFormer, supporting the hypothesis that multiscale information is beneficial for artifact segmentation. The 10x-5x configuration achieves the highest average artifact balanced accuracy (0.890), while the 10x-2.5x configuration achieves the highest average mIoU (0.532) and average Dice (0.666). Finally, the last row of the table shows the results of the proposed multitask multiscale SegFormer model with the integration of the stitch detection branch. The model maintains comparable performance to the multiscale SegFormer architectures without stitch detection, demonstrating that the addition of the stitch detection branch does not significantly compromise the model’s ability to segment macro artifacts, while providing the added functionality of stitch detection.

Figures 5.1, 5.2 and 5.3 present the per-class recall, IoU and Dice for the macro artifacts test set (T1-T8). The heatmaps show the performance of each model across the different artifact classes, allowing for a more detailed analysis of the strengths and weaknesses of each model in segmenting specific artifact types. Some observations from the heatmaps include:

- GrandQC shows reduced capability in accurately segmenting out-of-focus areas and foreign objects, as evidenced by the lower IoU and Dice values for these classes compared to the other models.
- Pen marker artifacts are generally well segmented across all models, with little variation in performance, suggesting that this artifact type may be easier to detect and segment, due to its color and size characteristics.
- The multiscale SegFormer models register a significant improvement in the recall of tissue folds and foreign objects compared to the single scale SegFormer, indicating that the multiscale approach may be particularly beneficial for segmenting this

artifact type, which can vary greatly in size and appearance.

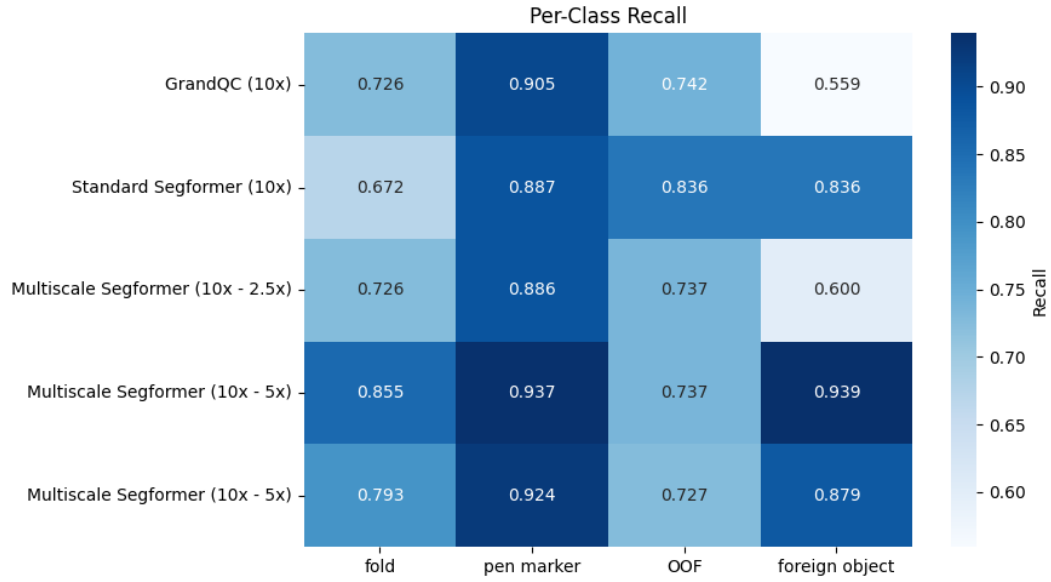


Figure 5.1: Per-class recall for the macro artifacts test set (T1-T8).

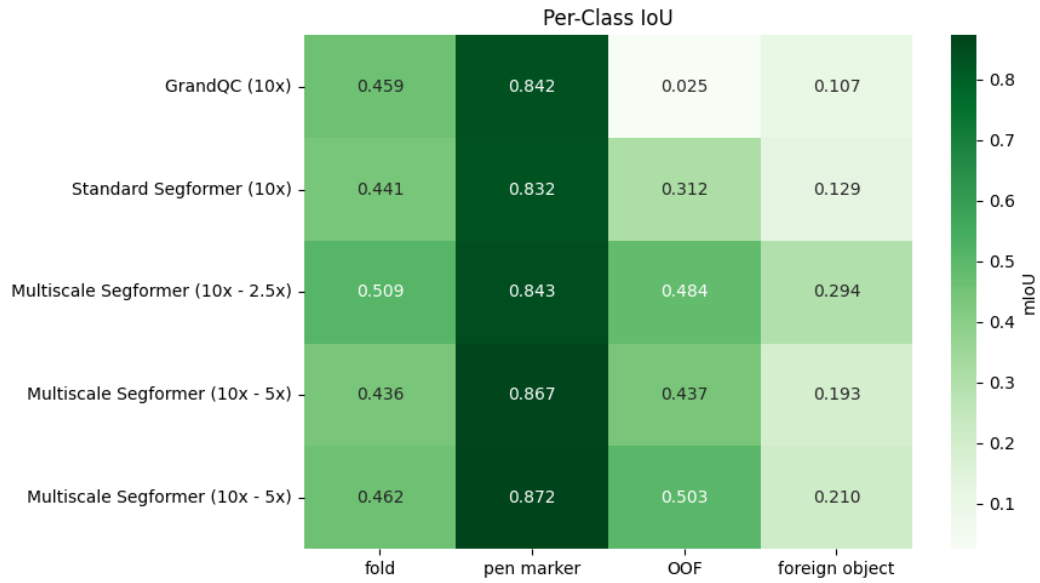


Figure 5.2: Per-class IoU for the macro artifacts test set (T1-T8).

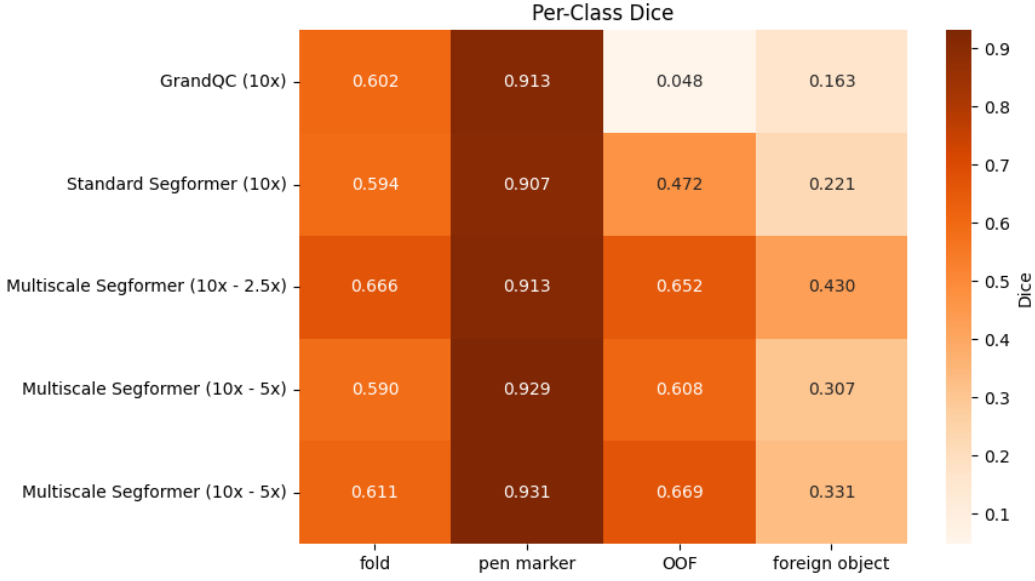


Figure 5.3: Per-class Dice for the macro artifacts test set (T1-T8).

Figure 5.4 presents the confusion matrix corresponding to the macro-artifact test set (T1-T8) evaluated using the proposed multitask multiscale SegFormer model. Because the ground truth exclusively provides annotations for artifact classes, the normal tissue and background classes were aggregated into a singular category. Notably, the most substantial cross-misprediction rate occurs between the out-of-focus and tissue fold artifact classes. This phenomenon can likely be attributed to the physical nature of folding artifacts, which inherently alter the focal plane and consequently render the adjacent tissue out of focus. The segmentation of these boundary regions is thus subject to ambiguity, as they simultaneously exhibit morphological characteristics of both artifact types; a representative visual example of this spatial overlap is illustrated in Figure 5.5. Furthermore, the misclassification rate of normal tissue and background as artifacts (first row) remains remarkably low. Consequently, diagnostically viable tissue—essential for subsequent downstream analysis—is rarely erroneously discarded. This high retention rate ensures that clinically relevant tissue regions are reliably preserved for both manual histopathological examination and automated computational pathology pipelines.

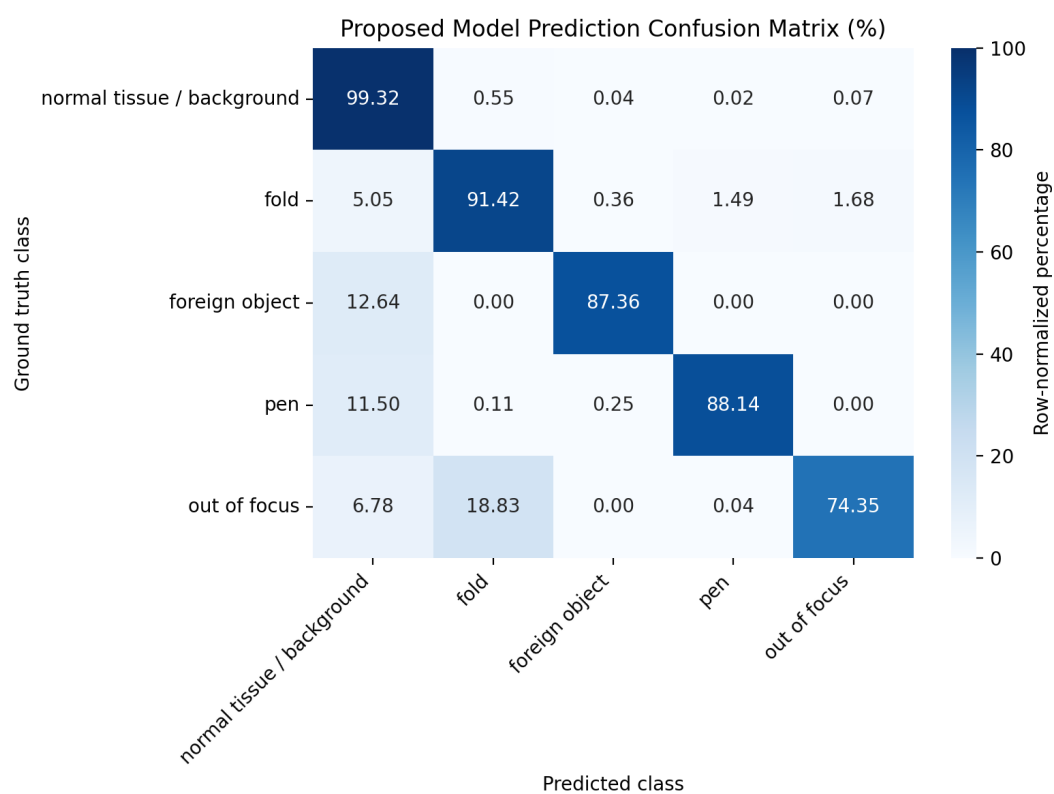
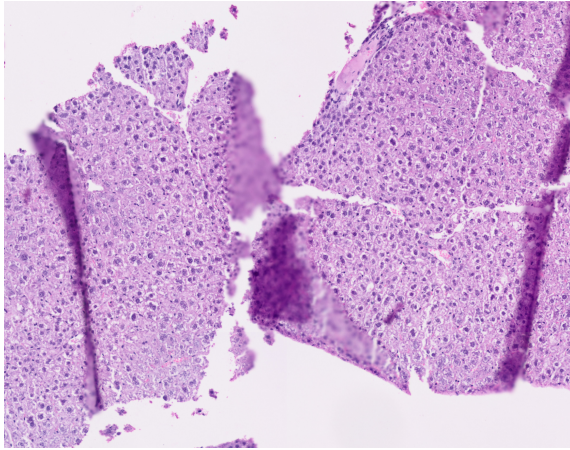
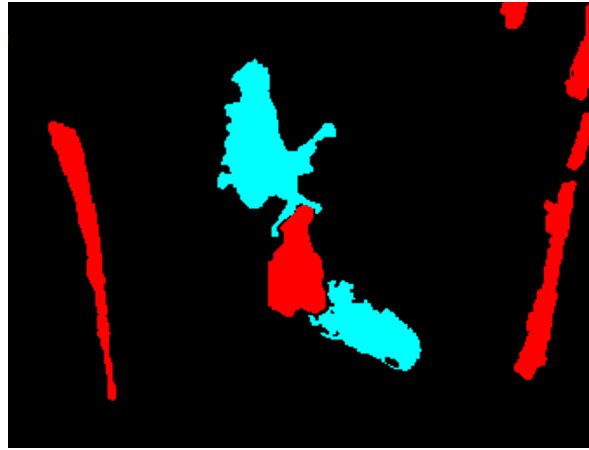
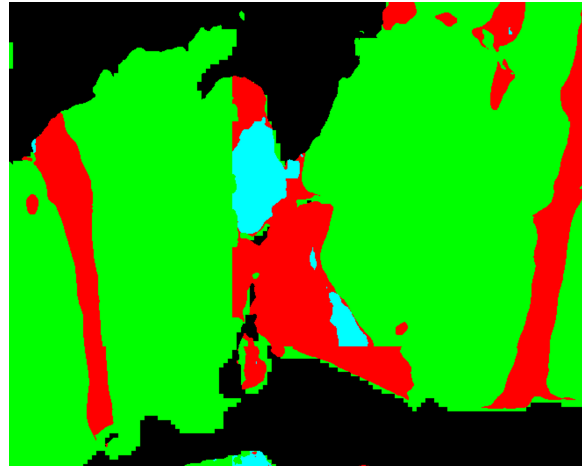


Figure 5.4: Proposed architecture confusion matrix.



(a) Input Image

(b) Ground Truth (*red*: fold, *cyan*: out of focus)

(c) Our prediction

Figure 5.5: Slide T6 (cropped)

Figures 5.6 to 5.13 present a visual comparison of the segmentation masks predicted by GrandQC, HistoQC and the proposed multitask multiscale SegFormer model for each of the slides in the macro artifacts test set (T1-T8). Each figure includes the input image, the ground truth mask, and the predicted masks from each model. HistoQC predictions report the “use” mask, which highlights the areas of the slide containing artifact-free tissue and that are considered usable for downstream analysis. GrandQC and the proposed model predictions contain the complete segmentation masks, containing both the artifact classes and the artifact-free tissue class.

Slide T1

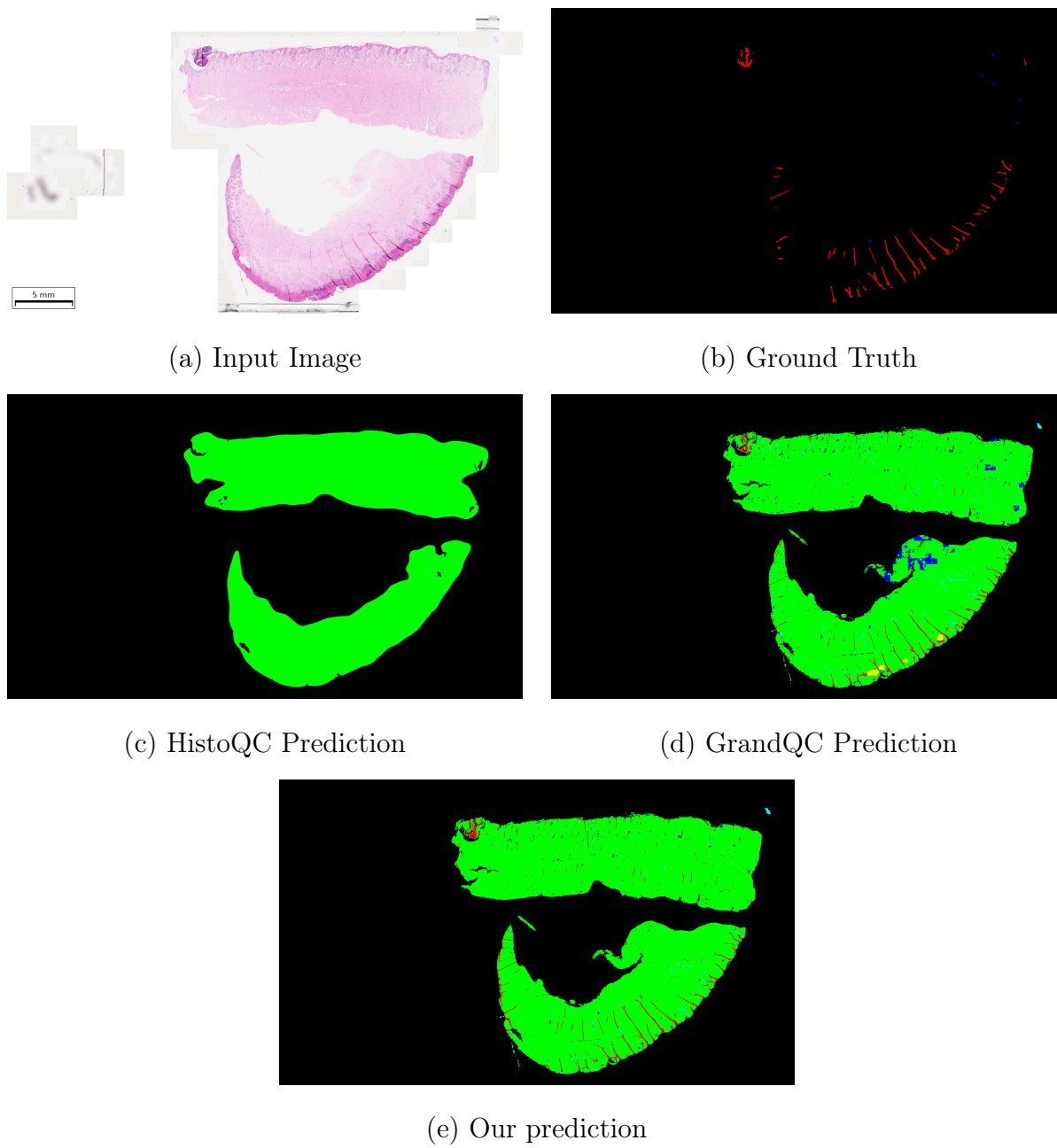
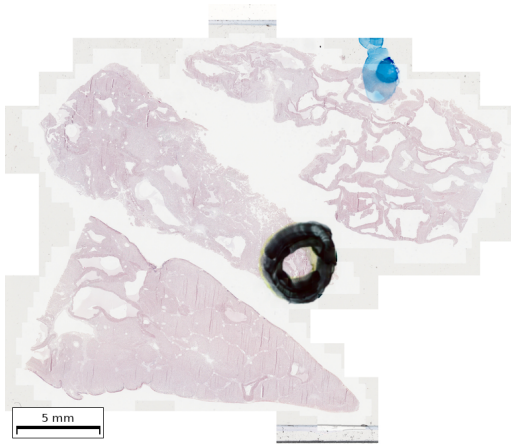
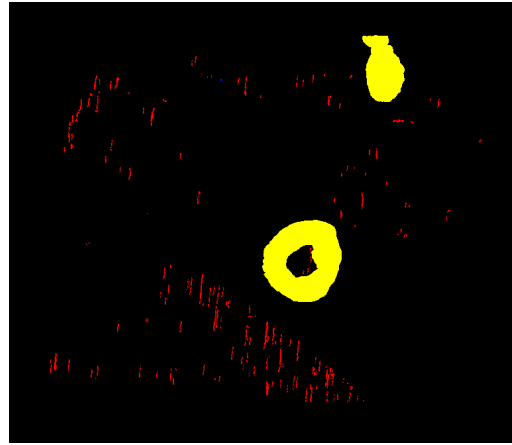


Figure 5.6: Slide T1

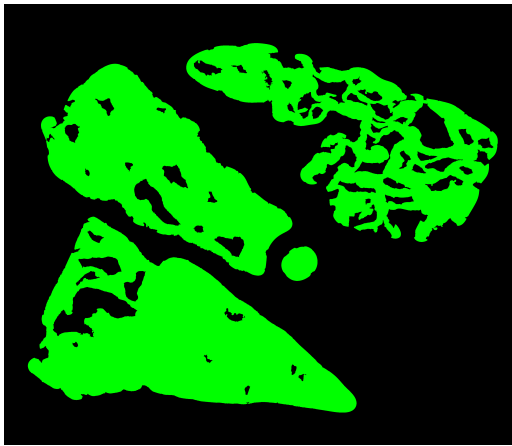
Slide T2



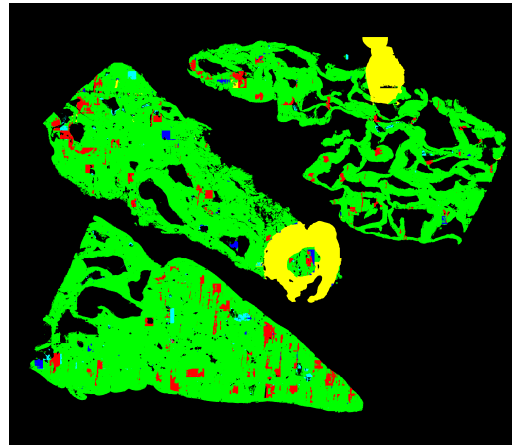
(a) Input Image



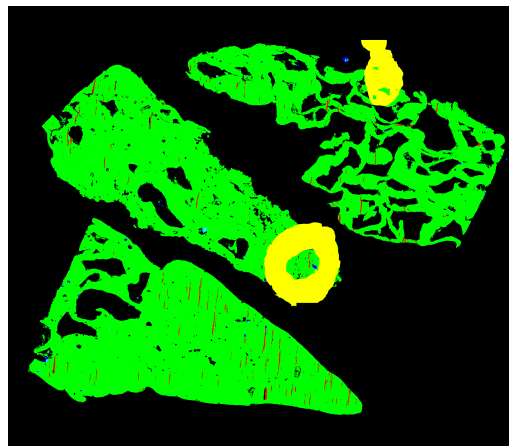
(b) Ground Truth



(c) HistoQC Prediction



(d) GrandQC Prediction



(e) Our prediction

Figure 5.7: Slide T2

Slide T3

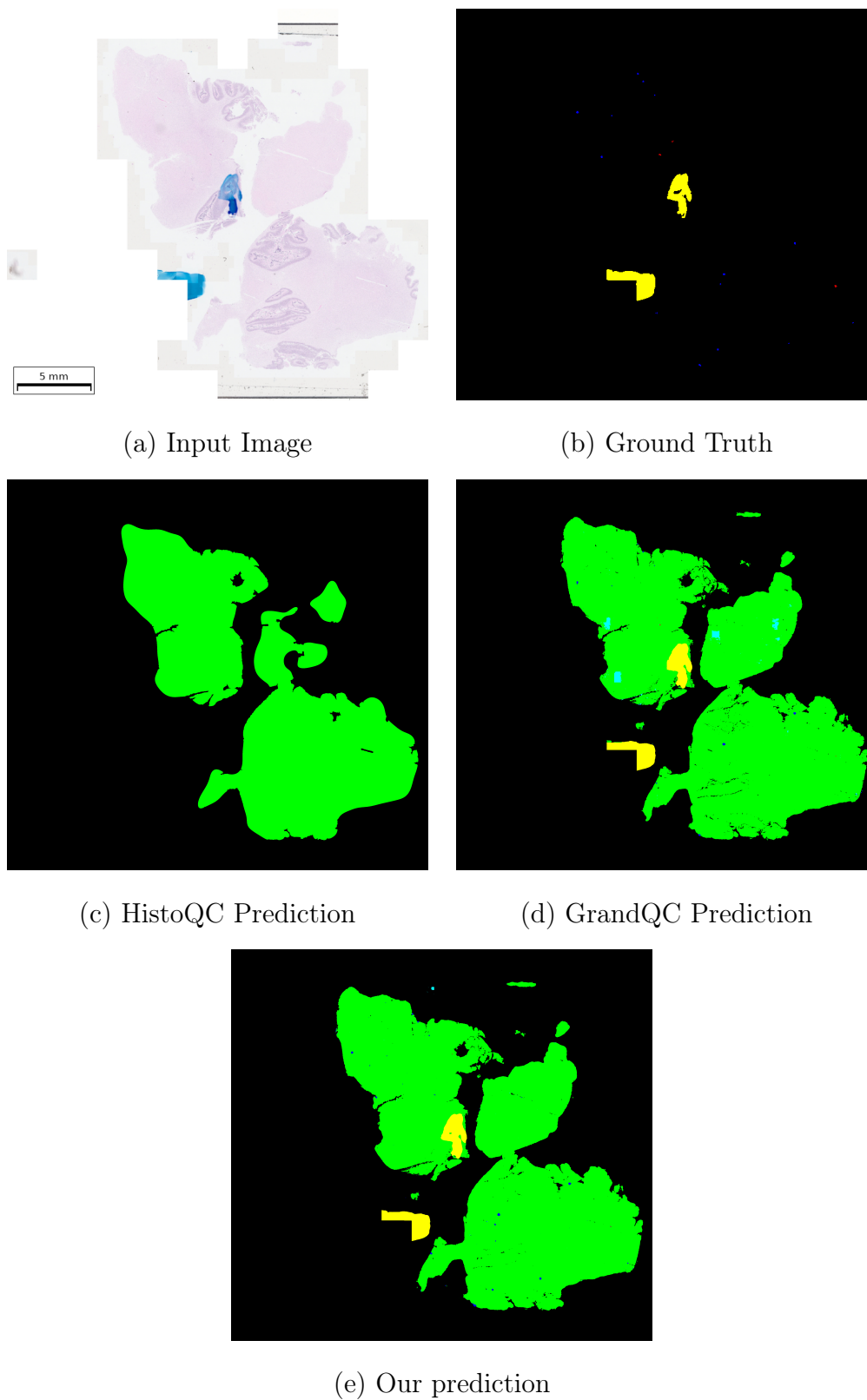
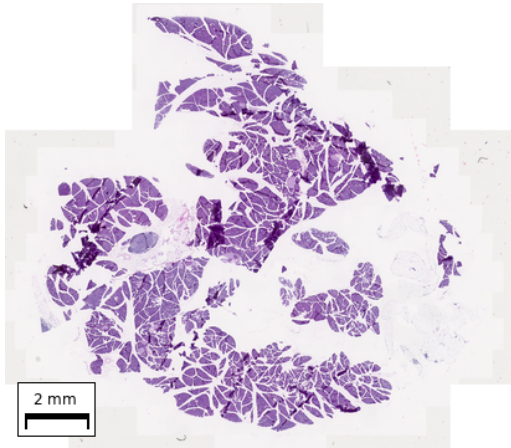
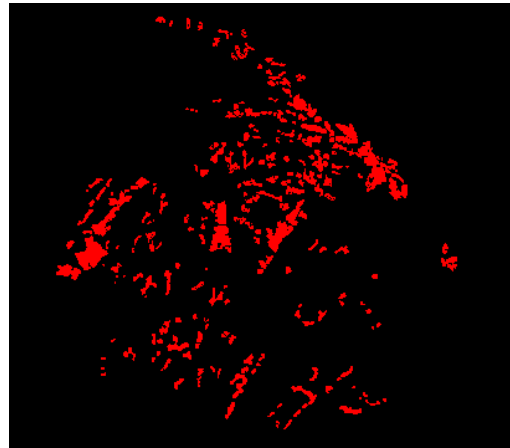


Figure 5.8: Slide T3

Slide T4



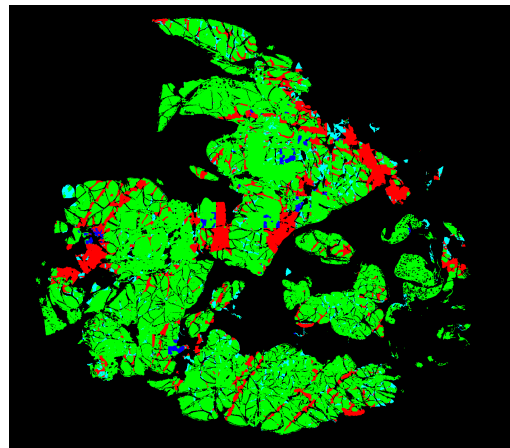
(a) Input Image



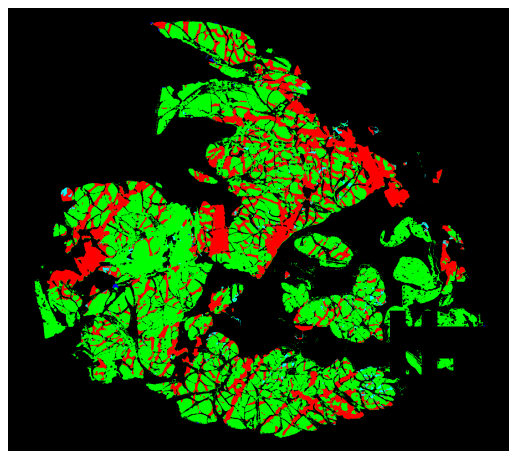
(b) Ground Truth



(c) HistoQC Prediction



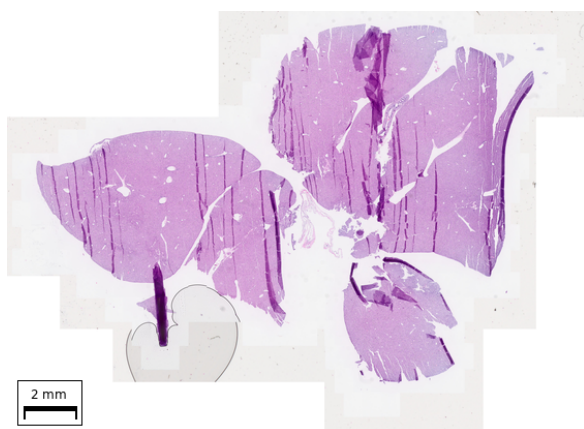
(d) GrandQC Prediction



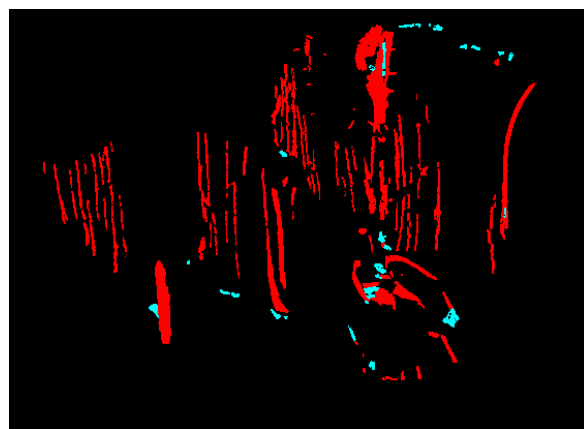
(e) Our prediction

Figure 5.9: Slide T4

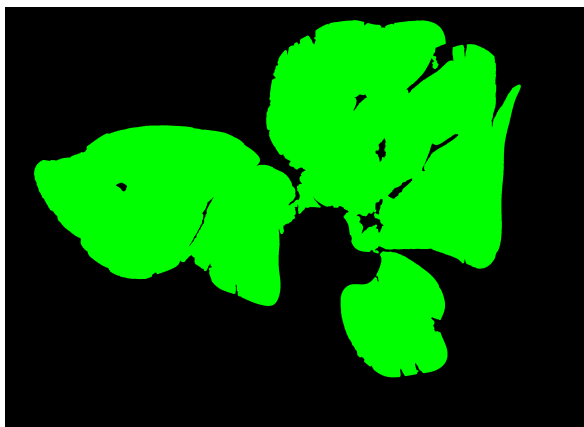
Slide T5



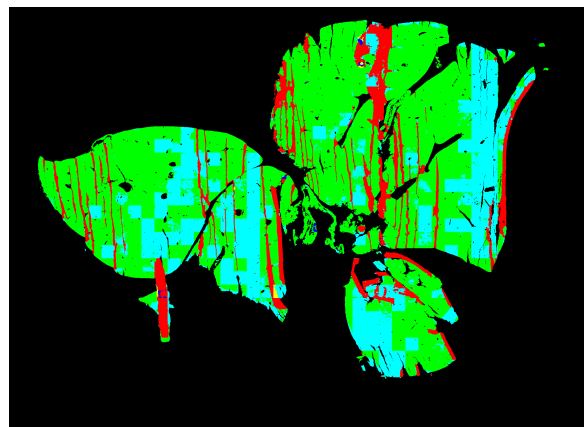
(a) Input Image



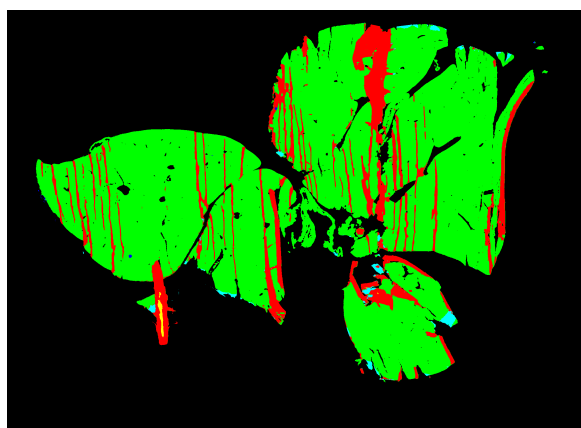
(b) Ground Truth



(c) HistoQC Prediction



(d) GrandQC Prediction



(e) Our prediction

Figure 5.10: Slide T5

Slide T6

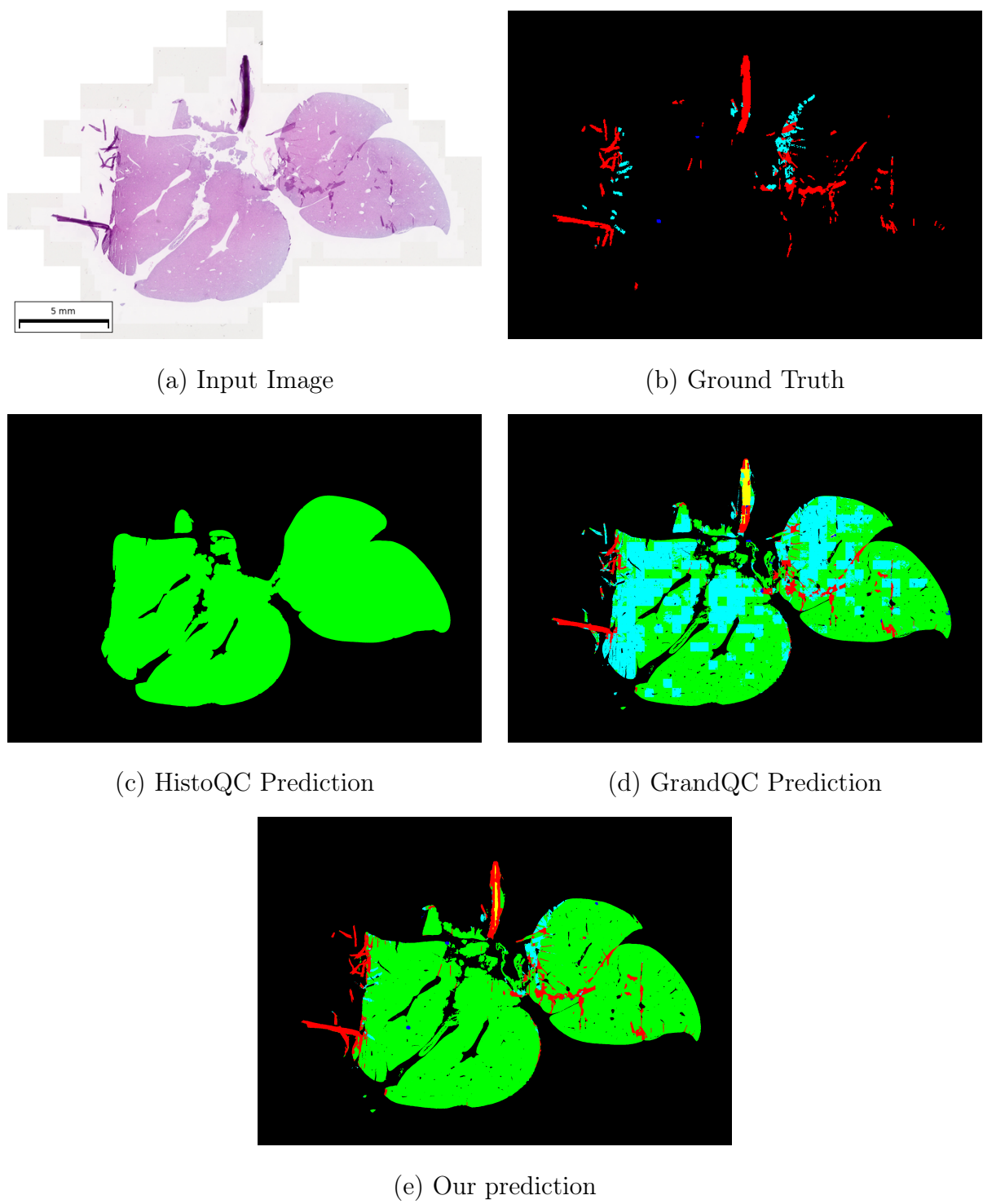
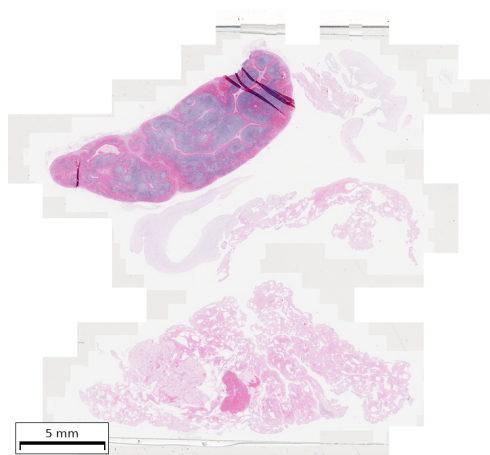
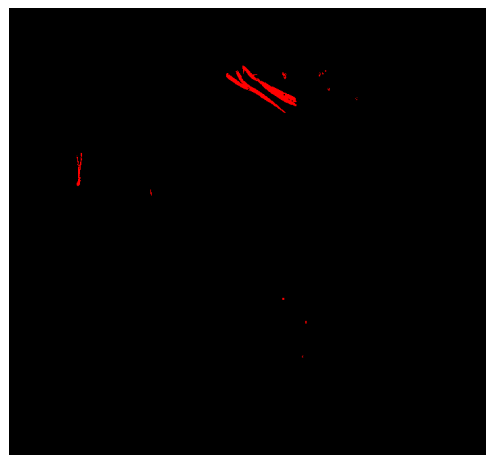


Figure 5.11: Slide T6

Slide T7



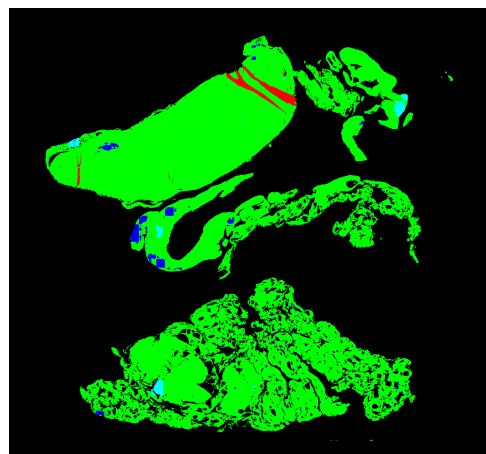
(a) Input Image



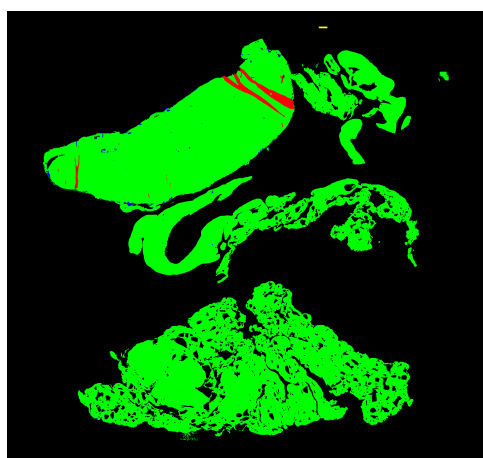
(b) Ground Truth



(c) HistoQC Prediction



(d) GrandQC Prediction



(e) Our prediction

Figure 5.12: Slide T7

Slide T8

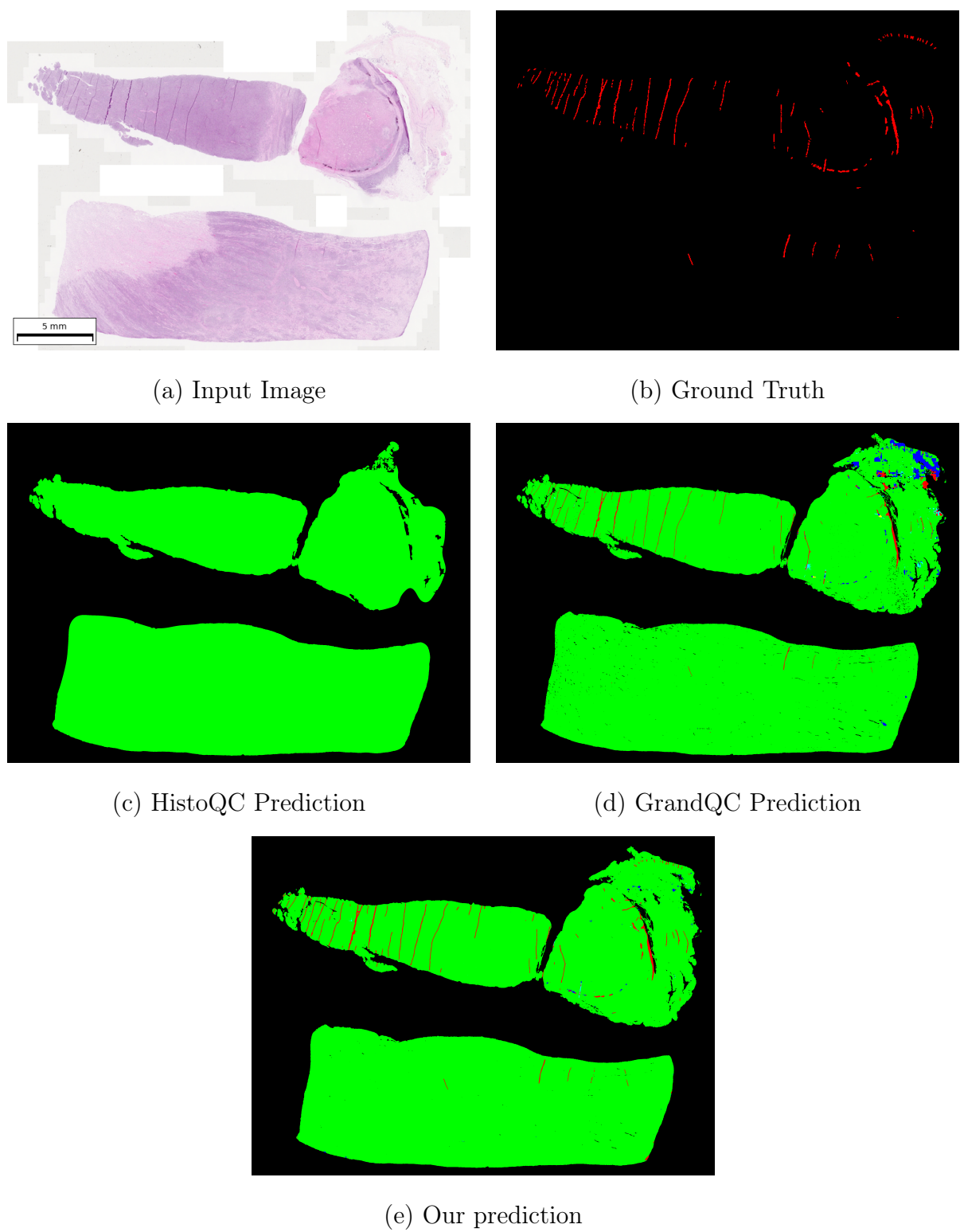
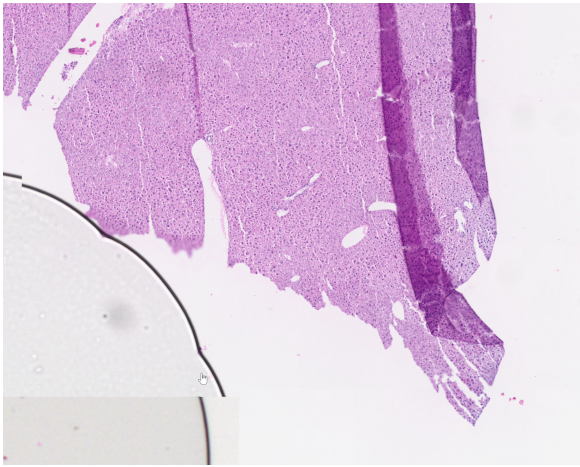
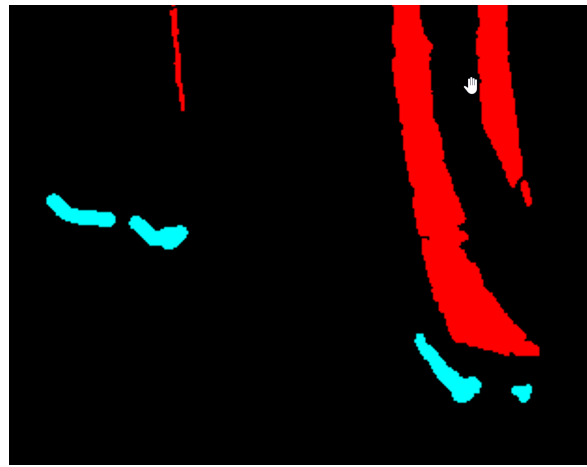


Figure 5.13: Slide T8

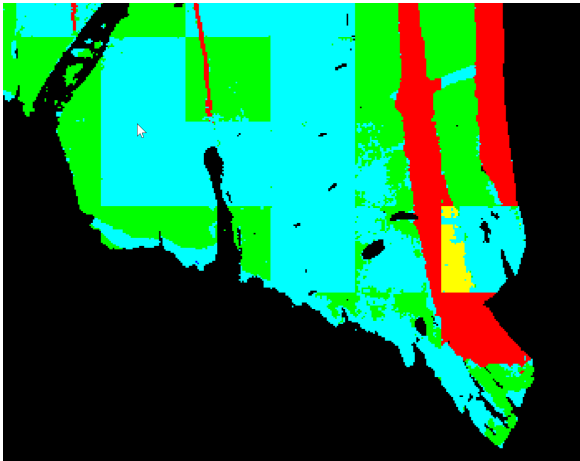
Visual inspection of the predicted segmentation masks complements the quantitative results, providing insights into the types of errors made by each model and the specific artifact classes that may be more challenging to segment. HistoQC predictions show a tendency to over-smooth boundaries and miss sparse tissue, incorrectly classifying it as background. This behaviour highlights the limitation of traditional image processing techniques in separating tissue from background, especially in cases where the tissue is thin or has low contrast with the background. Deep learning based tissue segmentation (GrandQC, Ours) shows improved performance in this regard. Figures 5.10 and 5.11 (cropped in 5.14) show poor precision in out of focus segmentation for GrandQC, with large areas of the slide being incorrectly classified as out of focus. This behaviour explains the extremely low IoU and Dice values for this artifact class for GrandQC in the heatmaps of Figures 5.1-5.3, compared to a much higher recall. Moreover, GrandQC predictions tend to be more fragmented and less smooth compared to the proposed model (Figures 5.7 (d) and 5.8 (d), cropped in 5.15), which may be a consequence of the model architecture and the lack of multi resolution processing, which is implemented in the SegFormer based models.



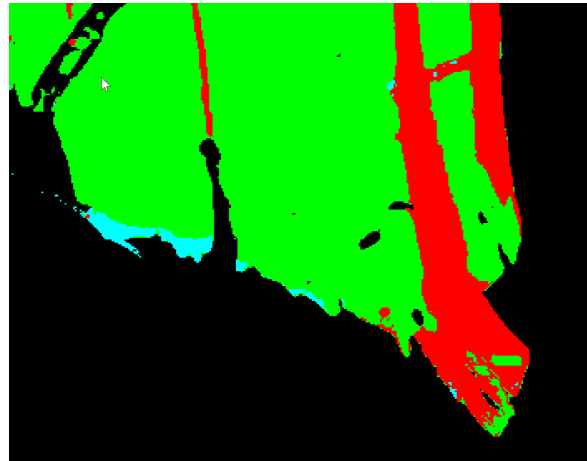
(a) Input Image



(b) Ground Truth

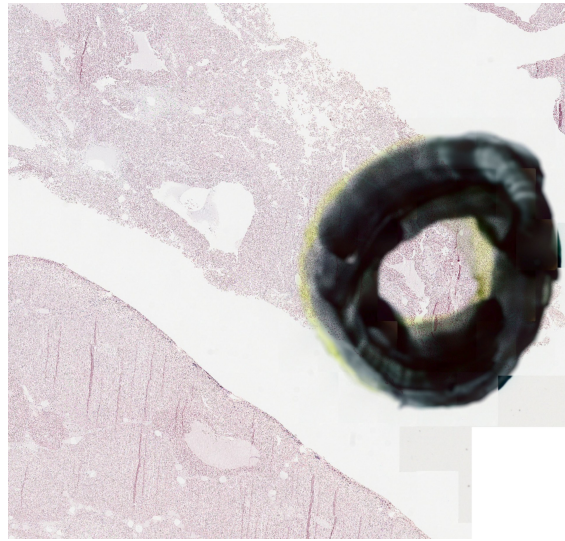


(c) GrandQC Prediction

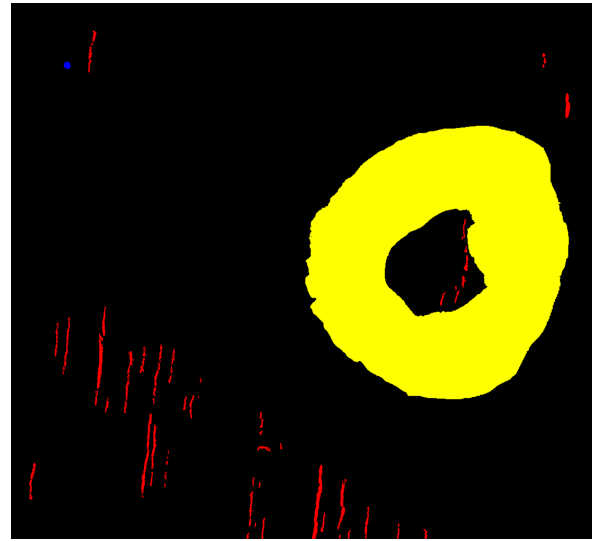


(d) Our prediction

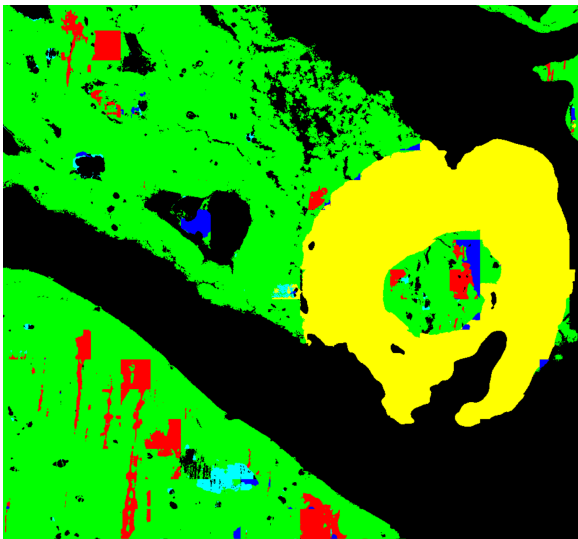
Figure 5.14: Slide T5 (cropped). See Figure 5.10 for full image. GrandQC mispredicts large areas of the slide as out of focus, while our model provides a more precise segmentation of the out of focus artifact, showcasing the improved generalization capabilities on different scanner devices.



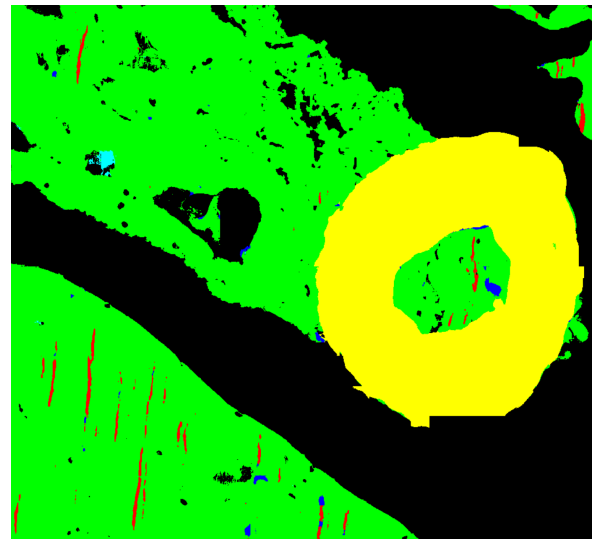
(a) Input Image



(b) Ground Truth



(c) GrandQC Prediction



(d) Our prediction

Figure 5.15: Slide T2 (cropped). See Figure 5.7 for full image. GrandQC correctly identifies tissue folds (*red*) but does not accurately delineate the artifact boundaries, leading to the misclassification of surrounding tissue as folded. In contrast, our model achieves a more precise segmentation of the tissue-fold artifact. However, some misclassification remains in our prediction for the foreign-object artifact (*blue*).

5.2. Evaluation of stitching artifact

In this section, we present the results of the proposed method for stitching artifact detection on the stitching artifact test set (T9-T13). The evaluation is performed separately

for each slide and then aggregated across the test set. The performance is assessed using the recall and false discovery rate metrics. The impact of the integration of a strip pooling module in the stitch detection branch is also analyzed, comparing the results of the proposed method with and without strip pooling.

Table 5.2 summarizes the aggregated performance of the proposed method for stitching artifact detection on the stitching artifact test set (T9-T13). Average recall and average false discovery rate are computed as the mean of recall and false discovery rate (FDR) across the 5 slides in the test set. The table includes the results of the proposed method with and without the integration of a strip pooling module in the stitch detection branch, right before the CNN decoder. The integration of strip pooling is hypothesized to improve the model’s ability to capture long-range dependencies and contextual information, which may be beneficial for stitch detection, as stitching artifacts can span large areas of the slide and have a characteristic straight shape, vertical or horizontal depending on the stitching direction.

Table 5.2: Model aggregated performance results - Stitching artifact test set (T9-T13)

Model	Input magnification	Stitch Detection	Strip pooling	Avg Recall	Avg FDR
Multiscale SegFormer	10x - 5x	Y	N	0.173	0.066
Multiscale SegFormer	10x - 5x	Y	Y	0.810	0.374

The results in Table 5.2 show a significant improvement in average recall with the integration of strip pooling (0.810 vs 0.173), indicating that the model is much more effective at identifying stitching artifacts when strip pooling is included in the stitch detection branch. This improvement comes with an increase in average false discovery rate (0.374 vs 0.066). This high false discovery rate has to be interpreted in the context of the annotation process described in Section 3.3.2, and it will be further analyzed in the next Chapter 6. A visual comparison of the raw stitch detection logits predicted by the stitch detection branch is shown in Figure 5.16. The images show the raw stitch detection logits for three different crops from the test set, with and without strip pooling. The integration of strip pooling results in a more continuous and less fragmented stitch detection output, consistent with the recall improvement observed in Table 5.2.



Figure 5.16: Impact of strip pooling on the raw logits predicted by the stitch detection branch.

Figures 5.17 to 5.21 present a visual comparison of the stitch detection results for each of the slides in the stitching artifact test set (T9-T13). Each figure includes the ground truth mask, the predicted stitch mask from the proposed method without strip pooling and the predicted stitch mask from the proposed method with strip pooling.

Slide T9



(a) Ground truth



(b) Our prediction (no strip pooling)



(c) Our prediction (with strip pooling)

Figure 5.17: Slide T9

Slide T10

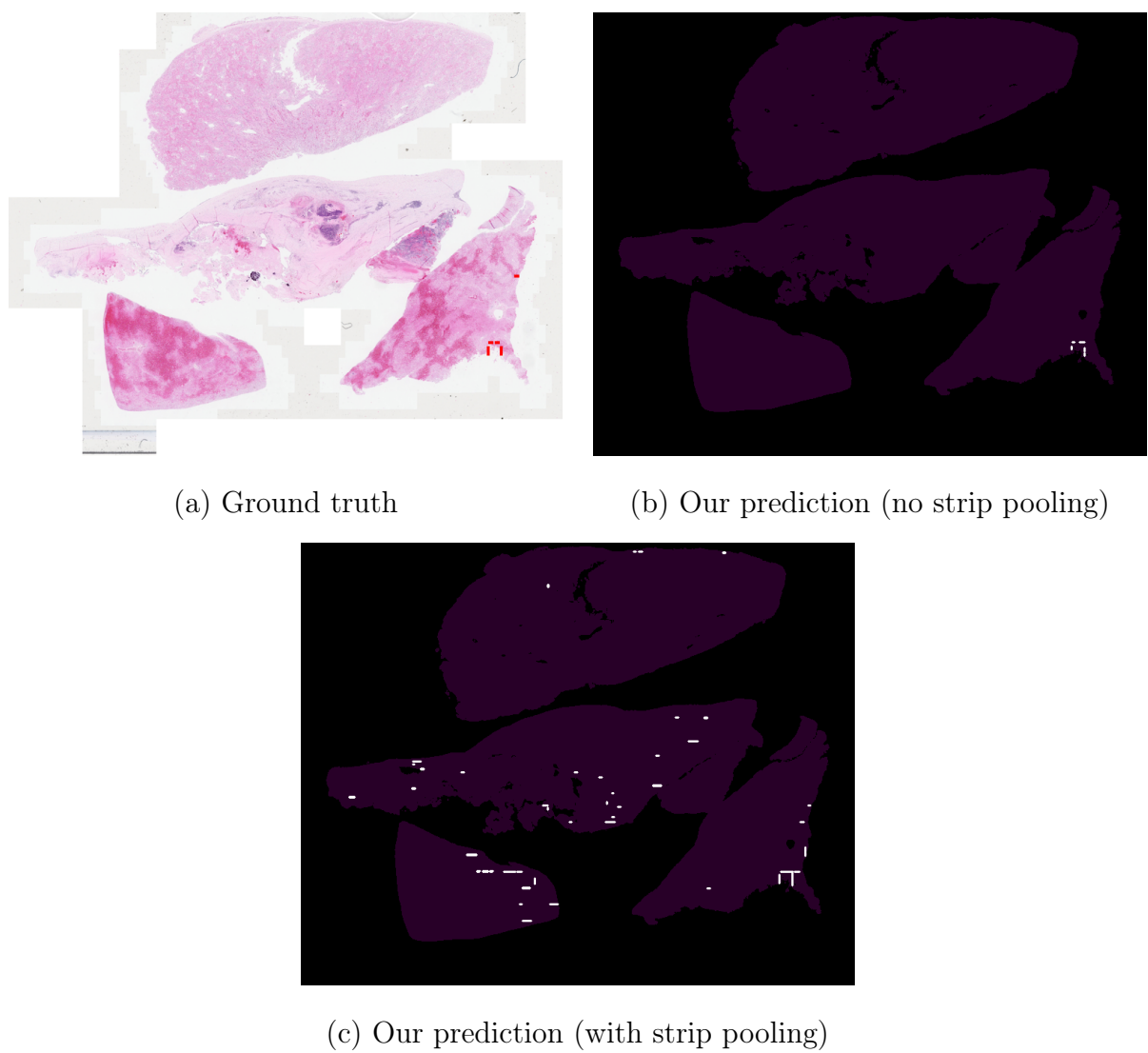
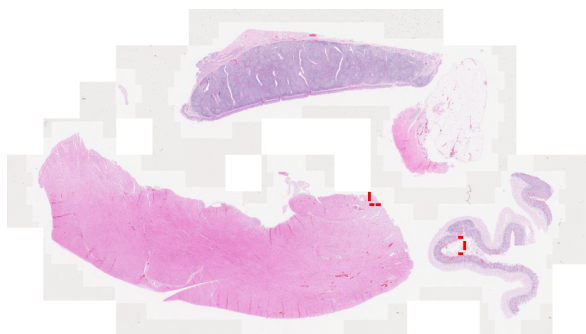


Figure 5.18: Slide T10

Slide T11



(a) Ground truth



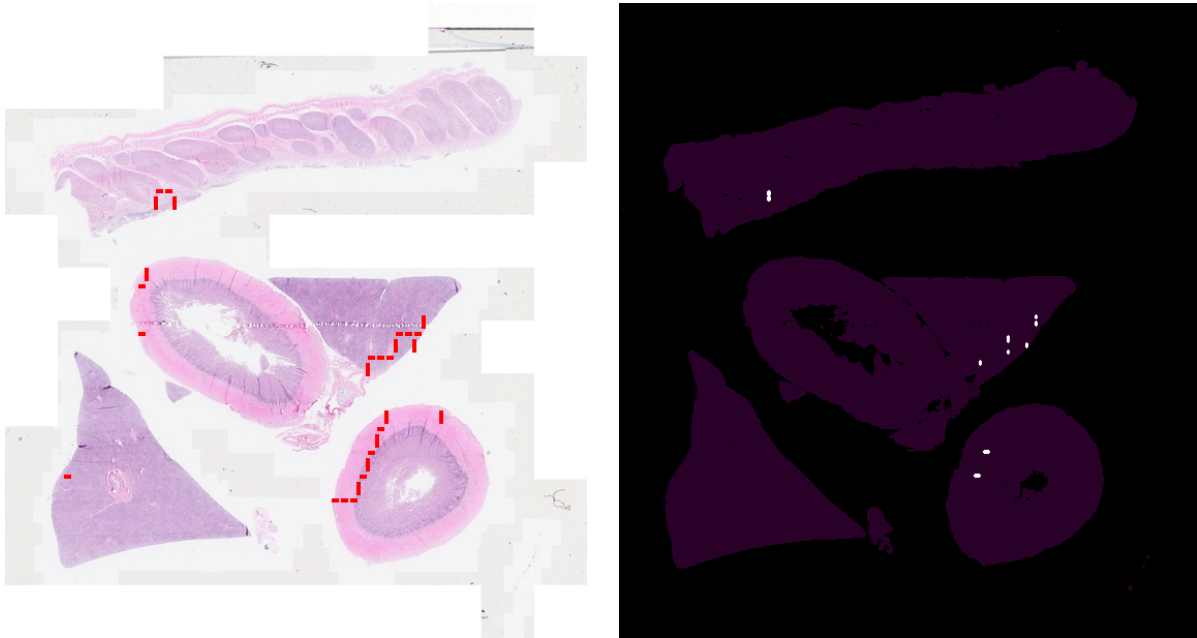
(b) Our prediction (no strip pooling)



(c) Our prediction (with strip pooling)

Figure 5.19: Slide T11

Slide T12



(a) Ground truth

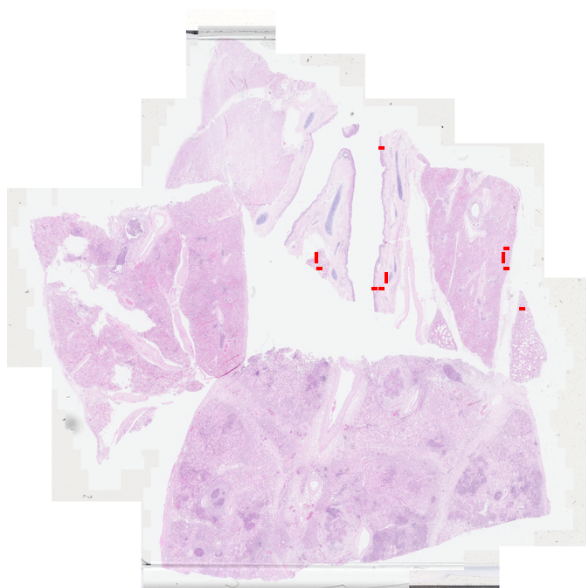
(b) Our prediction (no strip pooling)



(c) Our prediction (with strip pooling)

Figure 5.20: Slide T12

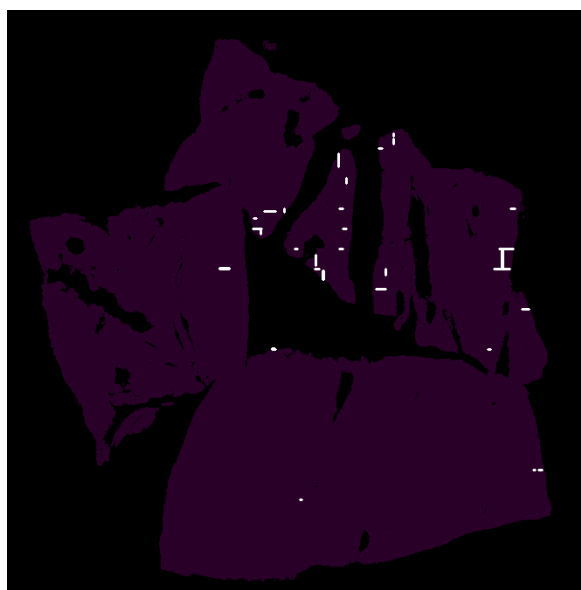
Slide T13



(a) Ground truth



(b) Our prediction (no strip pooling)



(c) Our prediction (with strip pooling)

Figure 5.21: Slide T13

5.3. Evaluation of the inference speed

Table 5.3 presents a comparison of the computational efficiency of models tested in this study, including GrandQC, the standard SegFormer model, the multiscale SegFormer models with different magnification combinations, and the proposed multitask multiscale SegFormer model with stitch detection. The system specifications are described in Section 4.1. The table includes the number of parameters for each model, the average throughput in terms of area processed per second (mm^2/s), and the average inference time per slide in seconds.

Table 5.3: Model inference speed results.

Model	Input magnification	Stitch Detection	Parameters	Avg Throughput (mm^2/s)	Avg Inference Time (s)
GrandQC	10x	N	6.25M	1.7	162.0
Standard SegFormer	10x	N	3.46M	6.4	42.0
Multiscale SegFormer	10x - 2.5x	N	6.91M	2.6	106.2
Multiscale SegFormer	10x - 5x	N	6.91M	2.2	126.3
Multiscale SegFormer	10x - 5x	Y	6.97M	2.1	141.5

All the SegFormer based models, including the proposed method, show a significant improvement in inference speed compared to GrandQC, with an average throughput ranging from 2.1 to 6.4 mm^2/s and an average inference time per slide ranging from 42.0 to 141.5 seconds, compared to 1.7 mm^2/s and 162.0 seconds for GrandQC. The standard SegFormer model with a single magnification input of 10x shows the highest throughput and lowest inference time, at the cost of a lower performance in terms of segmentation metrics compared to the multiscale models. The Multiscale SegFormer 10x - 2.5x model shows a slightly higher throughput and lower inference time compared to the Multiscale SegFormer 10x - 5x model. This can be attributed to the fact that the pyramidal WSI acquired by Grundium Ocus scanner already includes the 2.5x magnification level, which can be directly used as input for the model without the need for additional downsampling, as opposed to the 5x magnification level which needs to be generated from the 10x level through downsampling, which adds computational overhead and increases inference time. Despite the lower speed, the proposed multiscale architecture exploits the 10x - 5x magnification combination, which has been selected as the best trade off between performance and speed. The integration of the stitch detection branch in the proposed method results in a slight increase (0.8%) in the number of parameters, which translates into a decrease in average throughput and an increase in average inference time of 0.1 mm^2/s and 15.2 seconds respectively, compared to the multiscale SegFormer 10x - 5x model without

stitch detection. This decrease in speed is expected due to the additional computations required for the stitch detection task, but it is still within an acceptable range, supporting the multitask learning approach of the proposed method, with the introduction of a stitch detection branch providing the additional functionality with a minimal impact on computational cost.

6 | Conclusions

This chapter presents a comprehensive conclusion to the study. First, it discusses the experimental results in the context of the core research questions, specifically examining the efficacy of knowledge distillation for macro-artifact detection and the architectural integration of stitching artifact detection. Following this analysis, the chapter identifies the primary limitations of the current framework and outlines potential opportunities for future research to address these challenges. Finally, the chapter concludes with ethical considerations on the work and on the research field.

6.1. Discussion

The results presented in Chapter 5 are further analyzed in the following sections. Each section connects the results of the tested architectures with the identified research questions that motivated the study and that were presented in Chapter 1

6.1.1. Addressing RQ1: Artifact Detection & Knowledge Distillation

The first research question aimed to determine how effectively and efficiently Vision Transformers could be adapted through knowledge distillation for automated macro-artifact detection. The results definitively demonstrate that a Transformer-based architecture can successfully learn from noisy, distilled knowledge, and outperform the CNN-based approaches in terms of both accuracy and runtime.

Specifically, the Standard SegFormer (student) outperformed the GrandQC model (teacher) across all key metrics, yielding a notable improvement in average mIoU (0.429 vs. 0.358) and average Dice score (0.548 vs. 0.431). These results demonstrate both the effectiveness of the Human-In-The-Loop supervision during the training dataset building process and the already-established advantages of Vision Transformers architectures over traditional CNNs in terms of accuracy.

Furthermore, the integration of multiscale information proved highly beneficial for seg-

mentation performance. Artifacts in histopathology vary drastically in size and appearance; for instance, tissue folds require macroscopic context, whereas out of focus areas or small foreign objects are best identified at higher resolutions. By adopting a dual-branch architecture (both 10x - 2.5x and 10x - 5x magnification), the model achieved a superior average artifact balanced accuracy, mIoU and Dice score compared to the single scale architecture.

All the SegFormer-based models demonstrated higher efficiency in terms of throughput (mm^2/s) and inference time (s) compared to the teacher model (GrandQC). Specifically, the Standard SegFormer resulted in almost 4x faster inference than the GrandQC model, while the multiscale architectures showed a more conservative 1.2x to 1.5x improvement. The multiscale architecture inference time does not simply scale linearly with the number of processed patches (2 times more patches compared to the single scale architecture), but rather takes into account the increased complexity of the overall pipeline, which includes multi resolution patch extraction and downsampling, feature maps cropping, upsampling, and concatenation, that are necessary steps to enable multiresolution processing. These factors explain the slower inference speed of the multiscale architecture compared to the single scale SegFormer (2.2-2.6 vs 6.4 mm^2/s).

6.1.2. Addressing RQ2: Stitching Artifact Detection & Architecture Integration

The second research question addressed the formal modeling and integration of stitching artifact detection within the pipeline. Because stitching artifacts manifest as thin, elongated lines, standard square receptive fields struggle to capture their long-range spatial dependencies.

To overcome this, a Strip Pooling module was integrated into a secondary, lightweight CNN decoder head. The quantitative impact of this architectural choice was substantial: the inclusion of Strip Pooling increased the average recall for stitching artifacts from 0.173 to 0.810. By aggregating contextual information along distinct horizontal and vertical dimensions, the network successfully bridged small gaps between fragmented predictions, resulting in highly continuous stitch detection masks.

While the average False Discovery Rate (FDR) increased to 0.374 alongside the recall improvement, this metric must be contextualized within the limitations of the ground truth generation. The ground truth masks flag as stitching artifacts only seams with an estimated jump magnitude of at least 10 pixels. By visually inspecting the predicted stitch masks and the original slide images, we can notice that many of the false positives

are indeed seams with a jump magnitude smaller than 10 pixels but still visible in the slide, which may be considered as stitching artifacts. This suggests that the model may be detecting stitching artifacts that are not included in the ground truth due to the thresholding applied during the annotation process, rather than generating false positives in the traditional sense. Consequently, the elevated FDR highlights the model’s high sensitivity to sub-threshold alignment errors rather than a failure to distinguish between artifacts and normal tissue.

The stitching artifact detection has been successfully integrated into the pipeline, and the empirical results are promising. However, a critical limitation arises from the severe scarcity of training data for stitching artifacts: of the 18,432 total patches in the training dataset, only 504 contain stitch annotations, representing the 2.7% of the total. This profound class imbalance significantly disrupts the training dynamics when combined with the model’s strict spatial masking strategy and a relatively small batch size of 8. Because the stitching loss ($\mathcal{L}_{stitch}$) is computed exclusively on patches containing actual stitch pixels, the statistical probability of drawing a batch with zero stitching data is remarkably high. In practice, approximately 80% of the training batches consist entirely of macro-artifact or background patches, meaning the stitch decoder receives absolutely no gradient updates for the vast majority of the training process.

This extreme sparsity in the stitching gradient introduces significant training instability and creates a profound multitask optimization imbalance. While the stitching decoder sits idle during 80% of the optimization steps, the shared Vision Transformer encoders are continuously updated in every single batch, driven entirely by the dominant macro-artifact loss ($\mathcal{L}_{macro}$). As a result, the shared feature representations become overwhelmingly biased toward macro artifacts detection. This continuous, one-sided optimization heavily hinders the encoder’s capacity to learn the spatial features required to accurately resolve stitching lines.

6.2. Limitations and future work

Although the proposed pipeline achieved promising results for both macro-artifact and stitching artifact detection, several limitations must be acknowledged. These limitations concern the size and diversity of the training data, the range of artifact categories considered, the knowledge distillation strategy adopted, and the limited availability of stitching artifact annotations. Addressing these aspects would represent an important direction for future research and could further improve the robustness and generalizability of the proposed framework.

A first limitation concerns the size and heterogeneity of the dataset used to train the macro-artifact detection models. The training dataset employed in this study contains 75 whole-slide images (WSIs) selected from 12 TCGA cohorts and was extracted from the prediction masks generated by GrandQC. While this dataset was sufficient to demonstrate the feasibility of adapting Vision Transformer-based architectures through knowledge distillation, it remains substantially smaller than the dataset originally used to train GrandQC, which included 420 WSIs. Moreover, the complete GrandQC-generated dataset contains annotations for 13,543 WSIs from 32 cohorts. Therefore, the student models developed in this work were trained on only a limited subset of the available data distribution. This may constrain their ability to generalize across the full diversity of tissue types and organ-specific morphology encountered in large-scale histopathology repositories. Future work should therefore focus on expanding the training dataset by including a larger number of WSIs and by covering a broader range of TCGA cohorts. Increasing the number of training samples would likely improve the stability of the optimization process and reduce the risk of overfitting to cohort-specific characteristics. In addition, incorporating more tissue types and organs would allow the model to learn a more diverse representation of artifact appearance across different histological contexts.

A second limitation is related to the number of artifact categories considered in this study. The proposed model was designed to identify tissue folds, out-of-focus regions, foreign objects, pen markings, and stitching artifacts. These categories cover several of the most common and visually disruptive artifacts in digitized histopathology slides, and their detection is highly relevant for quality control workflows. Nevertheless, they do not exhaustively represent the full spectrum of artifacts that may occur during slide preparation, staining and scanning. Other artifact types, such as air bubbles and coverslip-related artifacts may also negatively affect downstream computational pathology pipelines. Extending the proposed framework to additional artifact categories would therefore be a natural continuation of this work.

Another important limitation concerns the knowledge distillation strategy used in this study. The distillation process was performed using hard labels, namely the final class predictions produced by the GrandQC teacher model. This approach has the advantage of simplicity and allows the student model to be trained as a standard supervised segmentation model. However, hard-label distillation discards potentially useful information contained in the teacher model's output probabilities. In particular, the relative confidence assigned by the teacher to each class may encode meaningful information about uncertain regions, ambiguous artifact boundaries, or visually overlapping artifact categories. By converting the teacher predictions into discrete labels, this uncertainty is lost.

Future research could therefore investigate soft-label-based knowledge distillation. In this setting, the student model would not only learn the final class assigned by the teacher, but also the full probability distribution over classes.

A further limitation concerns the training data available for stitching artifact detection. As discussed in the previous section, stitching artifacts are severely underrepresented in the current dataset. Only a small fraction of the training patches contain stitch annotations compared with the number of patches available for macro-artifact segmentation. This imbalance affects both the quality of the learned representations and the stability of the optimization process. Future work should therefore prioritize the enrichment of the stitching artifact dataset. A larger and more balanced stitching dataset would allow the model to receive more consistent supervision for the stitch detection task and would likely improve both recall and precision.

6.3. Ethical considerations

The deployment of deep learning architectures within the pathology workflow introduces significant ethical responsibilities that must be carefully considered. As automated systems are increasingly integrated into the pre-analytical and analytical phases of the clinical workflow [16], they directly influence diagnostic decision-making, patient safety, and data governance [33].

The development of robust and generalizable deep learning models is fundamentally dependent on the availability of diverse, high-quality datasets. Whole Slide Images (WSIs) are highly detailed, gigapixel-sized representations of a patient’s biological tissue. Although these images are routinely de-identified by removing explicit text metadata, the unique morphological structures, specific staining variations, and rare disease patterns can act as biometric identifiers, exposing patients to re-identification risks. Strict data sharing agreements and ethical guidelines are essential to mitigate these risks, but they also heavily restrict the sharing of annotated datasets, which are crucial for training and validating deep learning models against prior work. The scarcity of publicly available, well-annotated datasets creates a significant barrier to entry for researchers, particularly those from smaller academic institutions or independent research groups that lack direct affiliations with major healthcare networks capable of providing access to large-scale proprietary data.

As deep neural networks shift from traditional hand-crafted image processing techniques to highly complex, data-driven models, they inherently become “black boxes” whose internal decision-making processes are highly non-linear and difficult for humans to inspect.

Under the European Union Artificial Intelligence Act (EU AI Act) [11], AI systems deployed in healthcare that directly drive clinical decisions or act as computer-aided diagnosis tools are strictly classified as High-Risk AI Systems. However, the quality control module developed in this thesis operates exclusively within the pre-analytical phase, functioning as a technical triage mechanism rather than a diagnostic medical device. Because the system strictly identifies hardware or preparation-induced flaws without generating diagnostic predictions, it aligns with a minimal or low-risk classification. Moreover, by utilizing a multi-class pixel-level segmentation design, the model provides granular, spatially localized explanations for its artifact detection rather than outputting an opaque, unexplainable slide usability score. This ensures that, even as a low-risk supportive tool, its operations remain fully verifiable and intuitively interpretable to clinical users, improving the explainability and interpretability of the system.

Bibliography

- [1] A. Abdollahi, H. Saffar, and H. Saffar. Types and frequency of errors during different phases of testing at a clinical medical laboratory of a teaching hospital in tehran, iran. *North American journal of medical sciences*, 6(5):224, 2014.
- [2] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984.
- [3] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.
- [4] L. T. Blessing and A. Chakrabarti. *DRM, a design research methodology*. Springer, 2009.
- [5] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.
- [6] G. Campanella, A. R. Rajanna, L. Corsale, P. J. Schüffler, Y. Yagi, and T. J. Fuchs. Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. *Computerized medical imaging and graphics*, 65:142–151, 2018.
- [7] X. Chen, B. Zheng, and H. Liu. Optical and digital microscopic imaging techniques and applications in pathology. *Analytical Cellular Pathology*, 34(1-2):5–18, 2011.
- [8] M. Cui and D. Y. Zhang. Artificial intelligence and computational pathology. *Laboratory investigation*, 101(4):412–422, 2021.
- [9] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

- [11] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), 2024. URL <http://data.europa.eu/eli/reg/2024/1689/oj>.
- [12] A. Foucart, O. Debeir, and C. Decaestecker. Artifact identification in digital pathology from weak and noisy supervision with deep residual networks. In *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, pages 1–6. IEEE, 2018.
- [13] M. Haghighat, L. Browning, K. Sirinukunwattana, S. Malacrino, N. K. Alham, R. Colling, Y. Cui, E. Rakha, F. C. Hamdy, C. Verrill, et al. Pathprofiler: automated quality assessment of retrospective histopathology whole-slide image cohorts by artificial intelligence—a case study for prostate cancer research. *medRxiv*, pages 2021–09, 2021.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [16] M. S. Hosseini, B. E. Bejnordi, V. Q.-H. Trinh, L. Chan, D. Hasan, X. Li, S. Yang, T. Kim, H. Zhang, T. Wu, et al. Computational pathology: a survey review and the way forward. *Journal of Pathology Informatics*, 15:100357, 2024.
- [17] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4003–4012, 2020.
- [18] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
- [19] S. W. Jahn, M. Plass, and F. Moinfar. Digital pathology: advantages, limitations and emerging perspectives. *Journal of clinical medicine*, 9(11):3697, 2020.
- [20] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi. Histoqc: an

- open-source quality control tool for digital pathology slides. *JCO clinical cancer informatics*, 3:1–7, 2019.
- [21] A. R. Jara-Lazaro, T. P. Thamboo, M. Teh, and P. H. Tan. Digital pathology: exploring its applications in diagnostic surgical pathology practice. *Pathology*, 42(6): 512–518, 2010.
- [22] A. Jurgas, M. Wodzinski, M. D’Amato, J. Van Der Laak, M. Atzori, and H. Müller. Improving quality control of whole slide images by explicit artifact augmentation. *Scientific Reports*, 14(1):17847, 2024.
- [23] S. Kahaki, A. R. Webber, G. Zamzmi, A. Subbaswamy, R. Deshpande, and A. Badano. Histoart: Histopathology artifact detection and reporting tool. *arXiv preprint arXiv:2507.00044*, 2025.
- [24] N. Kanwal, F. Pérez-Bueno, A. Schmidt, K. Engan, and R. Molina. The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. *IEEE Access*, 10:58821–58844, 2022.
- [25] N. Kanwal, F. Khoraminia, U. Kiraz, A. Mosquera-Zamudio, C. Monteagudo, E. A. Janssen, T. C. Zuiverloon, C. Rong, and K. Engan. Equipping computational pathology systems with artifact processing pipelines: a showcase for computation and performance trade-offs. *BMC Medical Informatics and Decision Making*, 24(1):288, 2024.
- [26] J. N. Kather et al. GrandQC: QC Masks for TCGA whole-slide images, Nov. 2024. URL <https://doi.org/10.5281/zenodo.14041577>.
- [27] S. Khan, M. Tijare, M. Jain, and A. Desai. Artifacts in histopathology: A potential cause of misinterpretation. *Res Rev J Dent Sci*, 2(2):23–31, 2014.
- [28] T. Kohlberger, Y. Liu, M. Moran, P.-H. C. Chen, T. Brown, J. D. Hipp, C. H. Mermel, and M. C. Stumpe. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. *Journal of pathology informatics*, 10(1):39, 2019.
- [29] S. Kothari, J. H. Phan, and M. D. Wang. Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. *Journal of pathology informatics*, 4(1):22, 2013.
- [30] N. Kumar, R. Gupta, and S. Gupta. Whole slide imaging (wsi) in pathology: current perspectives and future directions. *Journal of digital imaging*, 33(4):1034–1040, 2020.
- [31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object

- detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [33] G. A. Matthews, C. McGenity, D. Bansal, and D. Treanor. Public evidence on ai products for digital pathology. *NPJ Digital Medicine*, 7(1):300, 2024.
- [34] S. Palokangas, J. Selinummi, and O. Yli-Harja. Segmentation of folds in tissue section images. In *2007 29th annual international Conference of the IEEE Engineering in Medicine and biology society*, pages 5641–5644. IEEE, 2007.
- [35] M. Scotto, R. Patti, V. L’imperio, F. Fraggetta, F. Molinari, and M. Salvi. Slideinspect: From pixel-level artifact detection to actionable quality metrics in digital pathology. *International Journal of Imaging Systems and Technology*, 36(1):e70292, 2026.
- [36] C. Senaras, M. K. K. Niazi, G. Lozanski, and M. N. Gurcan. Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning. *PloS one*, 13(10):e0205387, 2018.
- [37] H. Shakhawat, S. Hossain, A. Kabir, S. H. Mahmud, M. M. Islam, and F. Tariq. Review of artifact detection methods for automated analysis and diagnosis in digital pathology. In *Artificial Intelligence for Disease Diagnosis and Prognosis in Smart Healthcare*, pages 177–202. CRC Press, 2023.
- [38] H. M. Shakhawat, T. Nakamura, F. Kimura, Y. Yagi, and M. Yamaguchi. Automatic quality evaluation of whole slide images for the practical use of whole slide imaging scanner. *ITE Transactions On Media Technology And Applications*, 8(4):252–268, 2020.
- [39] G. Smit, F. Ciompi, M. Cigéhn, A. Bodén, J. Van Der Laak, and C. Mercan. Quality control of whole-slide images through multi-class semantic segmentation of artifacts. In *Medical Imaging with Deep Learning*, 2021.
- [40] T. Sørensen. A method of establishing groups of equal amplitude in plant sociologybased on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skr., K. danske Vidensk*, 5:1–34, 1948.
- [41] Z. Swiderska-Chadaj, T. Markiewicz, S. Cierniak, and R. Koktysz. Automatic quantification of vessels in hemorrhoids whole slide images. In *2016 17th International*

- Conference Computational Problems of Electrical Engineering (CPEE)*, pages 1–4. IEEE, 2016.
- [42] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.
 - [43] U.S. Food and Drug Administration. Technical performance assessment of digital pathology whole slide imaging devices, apr 2016. Guidance for Industry and Food and Drug Administration Staff. Docket No. FDA-2015-D-0230.
 - [44] M. Van Rijthoven, M. Balkenhol, K. Siliņa, J. Van Der Laak, and F. Ciompi. Hooknet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical image analysis*, 68:101890, 2021.
 - [45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
 - [46] N. C. Wang, J. Kaplan, J. Lee, J. Hodgins, A. Udager, and A. Rao. Stress testing pathology models with generated artifacts. *Journal of Pathology Informatics*, 12(1): 54, 2021.
 - [47] Z. Wang, M. S. Hosseini, A. Miles, K. N. Plataniotis, and Z. Wang. Focuslitenn: High efficiency focus quality assessment for digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–413. Springer, 2020.
 - [48] Z. Wang, Z. Zhou, Z. Wen, J. H. Kook, J. B. Wojcik, and J. Kang. Diffusionqc: Artifact detection in histopathology via diffusion model. *arXiv preprint arXiv:2601.12233*, 2026.
 - [49] Z. Weng, A. Seper, A. Pryalukhin, F. Mairinger, C. Wickenhauser, M. Bauer, L. Glamann, H. Bläker, T. Lingscheidt, W. Hulla, et al. Grandqc: A comprehensive solution to quality control problem in digital pathology. *Nature Communications*, 15(1):10685, 2024.
 - [50] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.
 - [51] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

List of Tables

2.1	Summary of related work in Deep Learning-based WSI artifact detection .	11
3.1	Sources of the artifact segmentation training dataset.	19
3.2	Sources of the artifact segmentation testing dataset.	22
4.1	Source projects of the selected training WSIs.	32
4.2	Training class distribution measured on labeled pixels.	32
4.3	Training class distribution measured on labeled pixels before and after applying the balancing strategies.	34
4.4	Summary of the artifact segmentation training dataset.	34
4.5	Weights assigned to each class in the macro-artifact segmentation branch, calculated based on the pixel distribution in the training set.	42
4.6	Default hyperparameters utilized for the network training phase.	45
5.1	Model aggregated performance results - Macro artifacts test set (T1-T8) .	48
5.2	Model aggregated performance results - Stitching artifact test set (T9-T13)	64
5.3	Model inference speed results.	72
6.1	List of Symbols	89

List of Symbols

Abbreviation	Description
CAD	Computer-Aided Diagnosis
CCD	Charge-Coupled Device
CMOS	Complementary Metal-Oxide Semiconductor
CMYK	Cyan, Magenta, Yellow, Key
CNN	Convolutional Neural Network
CPath	Computational Pathology
DCT	Discrete Cosine Transform
DRM	Design Research Methodology
DS-I	Descriptive Study I
DS-II	Descriptive Study II
FDR	False Discovery Rate
FOV	Field of View
H&E	Hematoxylin and Eosin
HITL	Human-in-the-Loop
IoU / mIoU	Intersection over Union / Mean Intersection over Union
MLP	Multi-Layer Perceptron
MoE	Mixture-of-Experts
PS	Prescriptive Study
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
VAE	Variational Autoencoder
ViT	Vision Transformer
WSI	Whole Slide Imaging (or Whole Slide Image)

Table 6.1: List of Symbols

Acknowledgements

I owe my deepest gratitude to the individuals and institutions that made this research possible. To my supervisors, Emin Zerman, Matti Pellikka and Giacomo Boracchi: thank you for your mentorship and constructive feedback. To my colleagues at Grundium, especially Jenna, for their support throughout this journey. To my family, for their endless love and encouragement, despite the physical distance from home. To my friends: thank you to the wonderful new people I met across the countries I lived in who brightened my daily life, and to my friends at home who were always there waiting to welcome me back. Finally, I would like to acknowledge the Erasmus Mundus Joint Master in Imaging committee and the European Union for the invaluable opportunity to participate in this academic program.

