Auxiliary Learning for Patch and WSI Pathology Image Classification.

Haoyu He

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

September 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Haoyu He**

Entitled: **Auxiliary Learning for Patch and WSI Pathology Image Classification.**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Dr.Yiming Xiao

_____ External Examiner
Dr. Dr.Yiming Xiao

_____ Examiner
Dr. Dr.Abdelhak Bentaleb

_____ Supervisor
Dr. Dr. Yang Wang

_____ Co-supervisor
Dr. Dr. Mahdi S Hosseini

Approved by _____
Dr. Denis Pankratov, Graduate Program Director, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Auxiliary Learning for Patch and WSI Pathology Image Classification.

Haoyu He

Early and accurate cancer detection through pathology imaging plays a critical role in improving patient outcomes. Despite promising advances from deep learning (DL) models, their real-world deployment is hindered by significant challenges.

To address the performance degradation caused by domain shifts—variations from different imaging devices, staining protocols, and patient demographics—we introduce PathTTT. This novel framework enhances model robustness at the patch level by combining Test-Time Training (TTT) with Model-Agnostic Meta-Learning (MAML). This bi-level optimization strategy leverages MAML to train a model for rapid adaptation, while TTT dynamically fine-tunes its parameters during inference, allowing for effective generalization to unseen distributions.

A separate, key challenge in computational pathology is the effective analysis of whole slide images (WSIs). While multi-instance learning (MIL) has become a widely adopted paradigm for WSI classification, it often suffers from overfitting due to the extremely high dimensionality and heterogeneity of WSIs, combined with the limited availability of annotated data. To address these challenges, we propose Masked Feature Embedding for Multi-Instance Learning (MFE-MIL), an innovative method designed to enhance existing end-to-end MIL pipelines. By introducing a masked feature embedding prediction task, our method provides a strong self-supervised signal that forces the model to learn highly discriminative and context-aware representations from instance features.

Extensive experiments on multiple benchmark pathology imaging datasets demonstrate that both PathTTT and MFE-MIL consistently outperform state-of-the-art methods in their respective domains. These results collectively underscore the potential of this work to facilitate more reliable and generalizable cancer detection systems in real-world clinical applications.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Yang Wang and my co-supervisor, Dr. Mahdi S. Hosseini, for your exceptional guidance, patience, insights, and encouragement. Without your supervision and invaluable support, this thesis would not have been possible.

Special thanks to my thesis examiners, Dr.Yiming Xiao and Dr.Abdelhak Bentaleb for their extremely valuable and constructive suggestions.

Lastly, I want to thank my family for their constant support and understanding during this period. Your love and encouragement have been my driving force.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer remains one of the leading causes of death globally, with breast and colorectal cancers being the most common types [1]. Breast and colorectal cancers have seen more than 310K and 152K new cases in the USA in 2024, resulting in more than 42K and 53K deaths, respectively [2]. Whole slide imaging (WSI) from tissue pathology plays a primary role in the early detection and diagnosis of cancer, which continues to pose significant challenges for clinicians, and researchers have been collaborating to develop Computer Aided Diagnosis (CAD) systems to automate the diagnostic evaluations. Despite the growing demand for Computational Pathology (CPath), analyzing these images to accurately assess the disease's progression remains a complex and time-consuming task. However, the deployment of these deep learning models in real-world clinical settings is hindered by two significant challenges: domain shift and the inherent limitations of Multi-Instance Learning.

This work introduces two novel frameworks designed to address key challenges in computational pathology. PathTTT offers a solution for enhancing model generalization in the presence of domain shift, while MFE-MIL provides a method for learning more robust instance-level features within the multiple instance learning (MIL) framework. Collectively, these contributions are intended to improve the reliability and performance of computational pathology models.

1.1 Challenge 1: Addressing Domain Shift with PathTTT

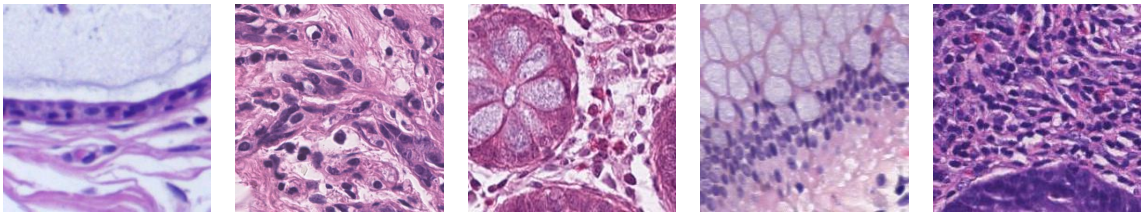
Deep learning has shown remarkable success in computational pathology, yet a major obstacle to clinical deployment is domain shift, where the distribution of training data differs from that of testing data. In pathology, domain shift frequently arises from staining variability (e.g., differences in hematoxylin and eosin [H&E] protocols across laboratories), scanner variation (e.g., Aperio vs. Hamamatsu devices producing distinct color profiles), and institutional bias (e.g., population differences or varying annotation standards). As a result, models trained on one dataset, such as breast cancer slides from TCGA, often exhibit substantial performance degradation when applied to an external dataset collected at other hospitals.

Figure 1.1 demonstrates two different CPath datasets, BACH [3] and BRACS [4], which focus on breast cancer classification, encompassing benign and malignant classes. Despite sharing this classification, they exhibit notable visual differences. Furthermore, these breast cancer datasets show certain visual similarities with datasets from other cancer types, such as Colorectal cancer datasets like NCT-CRC-HE-100K[5], Chaoyang[6], and MedFMC[7]. These cross-dataset visual variances and similarities highlight the challenges of domain shift in histopathological image analysis.

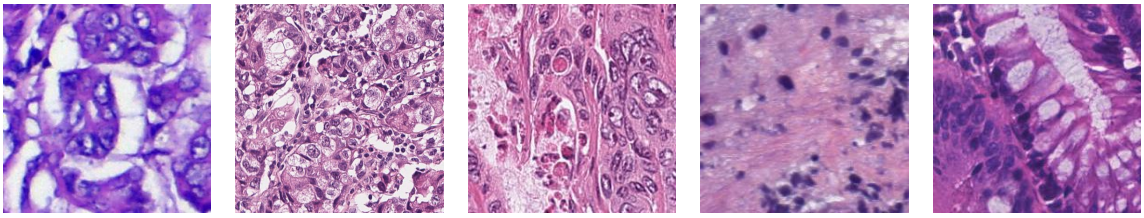
As shown in Figure 1.2, t-SNE diagrams[8] offer a powerful method for visualizing distribution shifts in high-dimensional feature spaces. The t-SNE visualization of the BACH and BRACS breast cancer datasets reveals a noticeable distribution shift between the two datasets. While there is some overlap, the points representing BACH (in blue) and BRACS (in orange) largely form separate clusters. This separation underscores the domain shift issue, where the same class labels (benign and malignant) are present in both datasets, but their visual representations differ significantly as shown Figure.1.1(a),(b),(f) and (g). Similarly, the t-SNE plot for three colorectal cancer datasets—NCT-CRC-HE-100K (in blue), Chaoyang (in orange), and MedFMC (in green)—shows significant overlap, suggesting that these datasets share more similar visual features, although distinct clusters are still present. These results align with the visual findings presented in Figure1.1.

To address this challenge, our work builds on the Test-Time Training (TTT) and Meta-Learning (MAML) paradigms. We propose PathTTT, a novel framework that integrates the strengths of these

approaches to improve model robustness. PathTTT treats each training instance as an individual task, employing a self-supervised Masked Autoencoder (MAE) loss to learn richer latent features. This bi-level optimization allows the model to dynamically adapt to new data distributions at test time. Our approach is the first to combine a meta-learning framework with an MAE-based auxiliary task to create a robust and generalizable model for cancer detection, bridging performance gaps across diverse clinical settings.



(a) BACH benign (b) BRACS benign (c) NCT benign (d) Chao benign (e) Med benign

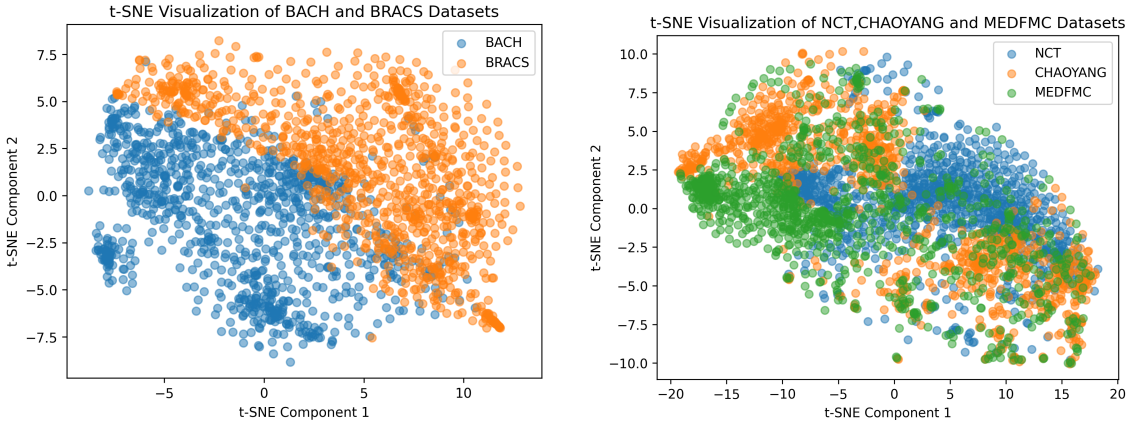


(f) BACH malignant (g) BRACS malignant (h) NCT malignant (i) Chao malignant (j) Med malignant

Figure 1.1: Sample images from two different classes of the BACH, BRACS, NCT-CRC-HE-100K, Chaoyang, and MedFMC datasets. The first row represents the benign class, while the second row represents the malignant class. The breast cancer datasets, BACH and BRACS, display distinct visual differences between these classes. Similarly, the colorectal cancer datasets, NCT-CRC-HE-100K, Chaoyang, and MedFMC, also exhibit clear visual differences between the benign and malignant cases, though some overlap is present.

1.2 Challenge 2: Improving Instance Feature Learning in MIL with MFE-MIL

Digital scans of pathology tissue slides, often referred to as WSIs, are widely used in computational pathology. They present a significant technical challenge due to the gigapixel scale of the images and the weakly labeled nature of available data. Traditional approaches are often bottlenecked by the sheer scale of the data and the lack of fine-grained, pixel-level annotations. To



(a) t-SNE visualization of BACH and BRACS

(b) t-SNE visualization of NCT, Chaoyang and MedFMC

Figure 1.2: Two t-SNE diagrams visualize the distribution shift in high-dimensional features for BACH, BRACS, NCT-CRC-HE-100K, Chaoyang, and MedFMC datasets. The t-SNE visualization shows a distribution shift between the BACH (blue) and BRACS (orange) breast cancer datasets, with some overlap but mostly distinct clusters. In contrast, the colorectal cancer datasets—NCT-CRC-HE-100K (blue), Chaoyang (orange), and MedFMC (green)—show more overlap, indicating shared visual features, though distinct clusters are still evident. These results are consistent with the visualization shown in Figure 1.1.

overcome these issues, Multi-Instance Learning (MIL) has become the prevailing approach, treating each WSI as a "bag" of image patches ("instances") with a single label assigned to the entire bag.

Despite their effectiveness, conventional MIL approaches face a key limitation: the inherent ambiguity at the instance level. Without fine-grained supervision, models receive weak learning signals for individual patches, hindering the development of highly discriminative and context-aware feature representations. This limitation poses a major bottleneck, as suboptimal embeddings can significantly degrade diagnostic accuracy and overall performance. In addition, standard MIL frameworks are prone to overfitting because they are trained only with slide-level labels, leaving patch-level predictions largely unconstrained. Consequently, the network can latch onto spurious patterns or dataset-specific artifacts—such as staining variations, scanner-specific characteristics, or repetitive tissue structures—rather than learning biologically relevant histopathological features. This over-reliance on superficial cues is especially pronounced when datasets are small or lack diversity, leading to poor generalization on unseen WSIs.

While self-supervised learning (SSL) holds promise for improving feature representations, its

direct application to WSI analysis is often impractical due to substantial computational demands. To overcome these limitations and contrast with common SSL methods, we introduce Masked Feature Embedding for Multi-Instance Learning (MFE-MIL), a novel and efficient self-supervised framework. A key distinction lies in its operational domain. Many popular SSL methods, such as contrastive learning frameworks like SimCLR or masked image modeling (MIM) approaches like MAE, operate directly on the raw image data. These methods typically rely on data augmentation (e.g., cropping, rotation, color shifts) or pixel-level masking to generate self-supervised signals. While highly effective on standard datasets, this paradigm is computationally infeasible for WSIs analysis. Applying these techniques to gigapixel-scale images would necessitate multiple rounds of feature extraction on enormous images, a process that is prohibitively expensive in terms of time, memory, and computational resources.

MFE-MIL elegantly sidesteps this computational bottleneck by applying its self-supervised task directly to the already-extracted feature vectors. After initial feature extraction from WSIs patches, we randomly mask a proportion of these feature embeddings. A Transformer-based decoder is then employed to reconstruct the original, unmasked features. This masked feature embedding prediction task provides a strong and efficient learning signal at the instance level, compelling the model to learn a more robust, context-aware representation for each instance. By operating on the compact feature space rather than the raw pixels, MFE-MIL provides a powerful and computationally efficient alternative to traditional SSL, effectively mitigating the ambiguity inherent in MIL and leading to significant improvements in downstream classification performance without the need for costly instance-level annotations.

1.3 Thesis organization

The rest of this thesis is organized as follows:

- Chapter 2: This chapter provides a comprehensive review of the existing literature, focusing on key concepts and methodologies relevant to our approaches, PathTTT and MFE-MIL.
- Chapter 3: This chapter presents the detailed frameworks of our two main approaches. We first describe PathTTT, our novel method for addressing domain shift. We then introduce

MFE-MIL, a multi-feature learning framework designed for improving feature learning. We discuss the architecture, key components, and rationale behind each approach.

- Chapter 4: This chapter presents the results of our experiments, comparing the performance of PathTTT and MFE-MIL against state-of-the-art methods. We also include a detailed ablation study to analyze the contribution of different components within our frameworks.
- Chapter 5: This chapter summarizes the key findings and contributions of this thesis. We conclude by discussing the potential limitations of our work and outlining promising directions for future research.

Chapter 2

Background and Literature Review

This work presents two novel deep learning frameworks that address distinct challenges in computational pathology. To provide the necessary context, this section first reviews prior research on domain generalization and test-time adaptation, which informs our PathTTT framework. We then discuss the foundational literature on Multi-Instance Learning and Self-Supervised methods, which are central to our MFE-MIL approach.

2.1 Background and Literature Review (Related to PathTTT)

This chapter provides an overview of key methodologies in computational pathology (CPath), critically examines existing approaches, and identifies gaps in current research. We present a comprehensive literature review of computer vision techniques, with a particular emphasis on their applications in pathology image analysis. Special attention is given to the limitations of contemporary models, particularly concerning generalizability and robustness under domain shift conditions.

2.1.1 Cancer Detection

The integration of deep learning (DL) into cancer detection has driven significant breakthroughs in computational pathology over the past decade. Convolutional neural networks (CNNs), in particular, have emerged as the cornerstone of automated image analysis. Architectures such as ResNet

[9, 10], DenseNet [11], and EfficientNet [12]have demonstrated exceptional performance in classifying histopathological images, often achieving expert-level accuracy in benchmark datasets.

However, despite their high performance on in-distribution data, these models often experience substantial performance degradation when deployed in unseen environments, such as data collected from different hospitals, staining protocols, or imaging devices. This phenomenon, commonly referred to as domain shift, poses a major limitation to the clinical translation of DL models. This issue undermines the generalizability of trained models and limits their practical deployment in real-world clinical workflows. To address this, various approaches have been proposed, including transfer learning, domain adaptation, and data augmentation. While these techniques can provide modest improvements, they often require access to labeled target domain data or assume relatively small distribution gaps. In more realistic, heterogeneous scenarios, their effectiveness remains limited, highlighting the need for more resilient and adaptive learning paradigms.

2.1.2 Domain Shift

Domain shift, where the distribution of training data differs from deployment data, is a major challenge in computational pathology. Sources of shift include staining variability (differences in H&E protocols across labs), scanner differences (e.g., Aperio vs. Hamamatsu), and institutional biases (population or annotation differences). Traditional approaches to mitigate domain shift include stain normalization and augmentation and domain-adversarial learning to encourage domain-invariant features. More recent methods employ self-supervised learning (e.g., contrastive learning, masked autoencoders) and test-time adaptation, such as Test-Time Training (TTT) and meta-learning strategies, to improve robustness under cross-institutional variability.

2.1.3 Masked Autoencoders

Masked Autoencoders (MAEs) represent a class of self-supervised learning methods that have gained considerable attention for their ability to learn robust visual representations without requiring labeled data. Introduced by [13], MAEs work by randomly masking a large portion (e.g., 75%) of an input image and training the model to reconstruct the missing pixels from the visible context. Unlike traditional autoencoders that reconstruct the entire image, MAEs focus on learning

semantic understanding and global structures, encouraging the encoder to extract informative and transferable features. The success of MAEs in natural image domains has prompted their adaptation to medical imaging, where labeled data is often scarce and domain variability is high. In computational pathology, MAEs offer a promising avenue for developing pretraining strategies that generalize well across diverse datasets and clinical settings.

2.1.4 Test-Time Training

Test-Time Training (TTT) is a technique designed to address the distribution shift problem that frequently occurs between training and test data. First introduced by [14], TTT allows models to adapt to test-time data by updating their parameters based on a self-supervised auxiliary task during inference. This adaptation can mitigate the effects of distribution shift by allowing the model to adjust its internal representations based on the characteristics of the test data, without needing access to labeled test samples. In the context of cancer detection, TTT provides a practical and lightweight solution for enhancing model robustness in real-world scenarios where retraining on new domains is impossible. When combined with powerful self-supervised tasks such as MAE-based reconstruction, TTT can dynamically adapt to diverse test-time data, offering improved generalization and clinical reliability. Despite these advances, current TTT and SSL approaches do not jointly address rapid adaptation and instance-level ambiguity, motivating our PathTTT frameworks.

2.1.5 Meta Learning

Meta-learning, often referred to as "learning to learn," is a framework that focuses on designing models capable of learning new tasks quickly with limited data. One of the most influential meta-learning algorithms is Model-Agnostic Meta-Learning (MAML), introduced by [15], which aims to learn an initialization of model parameters that can be rapidly fine-tuned using a few gradient steps on a novel task. MAML has been widely used in various fields, including computer vision, natural language processing, and reinforcement learning. Its applicability to computational pathology lies in its ability to facilitate quick adaptation to new cancer types, staining protocols, or patient populations with minimal labeled data. When integrated with test-time training or self-supervised learning, meta-learning can provide a principled framework for improving model adaptability and resilience

under domain shift.

2.1.6 Patch-Level Classification in Computational Pathology

Deep learning has been extensively utilized in patch-level analysis for pathology image classification, offering scalable and automated solutions for diagnostic support. Among recent contributions, FCCS-Net (FCCS) [16] presents a multi-level fully convolutional architecture equipped with dual attention mechanisms, specifically designed to enhance breast cancer classification by capturing both local and contextual features. CRCCN-Net (CRCCN) [17], a lightweight CNN model, addresses the challenge of automated colorectal tissue classification by striking a balance between computational efficiency and diagnostic accuracy—making it suitable for resource-constrained clinical settings. EfficientNetV2 (EV2) [12] extends the EfficientNet family, offering faster training and improved parameter efficiency, and has demonstrated strong performance across a variety of histopathological classification tasks.

2.1.7 Test-Time Adaptation Methods

Recent developments in test-time adaptation (TTA) have led to the emergence of frameworks designed to enhance model robustness in dynamic clinical environments. SAR [18] leverages entropy minimization and sharpness-aware optimization to enable stable model updates during inference, effectively mitigating distributional uncertainties. EATA [19] introduces parameter-efficient adaptation strategies combined with anti-forgetting mechanisms, thereby preserving model performance across diverse domain shifts while maintaining low computational overhead. These approaches represent promising directions for improving generalization in real-world deployments where access to labeled test data is limited or unavailable.

2.2 Background (Related to MFE-MIL)

This section reviews the core components that form our MFE-MIL approach and overfitting in WSIs: the Multi-Instance Learning (MIL) framework, the role of Self-Supervised Learning (SSL) for feature extraction, and the power of large-scale foundation models in computational pathology.

2.2.1 Self-Supervised Learning (SSL)

Self-supervised learning (SSL) is a powerful paradigm for learning valuable representations from large amounts of unlabeled data, which is especially critical in domains like medical imaging where manual annotation is costly and time-consuming. The core idea is to train a model on a "pretext task"—a supervised task created directly from the data itself—in order to learn features that are transferable to other, downstream tasks like image classification.

Modern SSL methods can broadly be categorized into two major families:

- **Contrastive Learning:** Methods such as SimCLR [20], MoCo [21], and DINO [22] learn representations by pushing dissimilar images apart in a feature space while pulling similar images (different augmented views of the same image) closer together. This process forces the model to learn what makes images distinct, leading to robust and discriminative features.

- **Masked Image Modeling (MIM):** Inspired by language models, MIM methods like MAE and BEiT [23] randomly mask parts of an image and train a model to reconstruct the missing information. This pretext task compels the model to learn a rich contextual understanding of the image. A key variant, Masked Feature Prediction (MFP), reconstructs high-level semantic features rather than raw pixels, often leading to more efficient and semantically meaningful representations.

These SSL approaches have proven highly effective for bridging the "domain gap" between models pre-trained on general datasets (like ImageNet) and specialized medical imagery, often outperforming traditional pre-training in various downstream tasks.

2.2.2 Application of MIL in WSIs classification

Whole Slide Images (WSIs) present a unique challenge for deep learning due to their immense size, often reaching gigapixel resolutions. Direct processing of these images is computationally infeasible. Consequently, WSIs are typically divided into a large collection of smaller patches. The primary challenge then shifts to classifying the entire WSI based on a single, global slide-level label, without specific annotations for each individual patch. This problem is perfectly suited for the

Multiple-Instance Learning (MIL) paradigm. While the MIL framework provides a robust solution, its effectiveness is often bottlenecked by two key stages: the quality of the feature extraction from individual patches and the sophistication of the aggregation mechanism.

In the MIL framework, each WSI is treated as a "bag", and the individual patches extracted from it are its "instances." The core assumption is that a bag is labeled "positive" if at least one of its instances is positive, while a "negative" bag implies all its instances are negative. This approach provides a robust solution for weak supervision in computational pathology, where patch-level annotations are often unavailable.

The standard MIL pipeline for WSIs classification generally consists of two stages:

Feature Extraction: A deep neural network, such as a Convolutional Neural Network (CNN) or a Vision Transformer (ViT), is used to extract a rich feature vector from each individual patch.

Aggregation: The feature vectors from all patches in a bag are combined by an aggregation function to produce a single, bag-level representation, which is then used for the final classification. Early MIL methods relied on simple aggregation functions like Max Pooling and Mean Pooling. While straightforward, these methods have drawbacks: Max pooling can miss important context by focusing only on the single most "positive" patch, while Mean pooling can dilute the signal from a small number of crucial patches with a large number of unimportant ones. To overcome this, modern MIL research has developed more sophisticated, learnable aggregation mechanisms. A key advancement is Attention-based MIL (ABMIL) [24], where the model learns to assign importance weights to each patch, focusing its attention on the most relevant instances. This not only boosts performance but also offers a degree of interpretability. Other advanced methods, like Transformer-based MIL (TransMIL)[25], model complex relationships between patches to capture global context, further enhancing classification accuracy.

2.2.3 Foundation Models in Computational Pathology (CPath)

Foundation Models represent the next generation of powerful self-supervised encoders, trained on massive datasets to learn highly generalizable representations. These models provide an ideal starting point for the feature extraction stage of a robust MIL pipeline. CONCH [26] and UNI [27] are two powerful foundation models that are pushing the boundaries of CPath.

CONCH is a vision-language model trained on 1.17 million image-caption pairs, enabling it to understand both visual pathology data and its textual descriptions. This allows it to achieve state-of-the-art results on 14 CPath tasks, including image classification and text-to-image retrieval.

UNI is the largest histopathology vision encoder, pretrained on 100 million images and 100,000 whole slide images. It excels at a wide range of tasks, showing state-of-the-art performance on 34 clinical tasks and notable gains in identifying rare and underrepresented cancer types.

2.2.4 Overfitting in WSI Classification

Overfitting is a major challenge in computational pathology, particularly for whole slide image (WSI) analysis. WSIs are extremely high-dimensional, often containing billions of pixels, while annotated datasets are typically small due to the labor-intensive nature of expert labeling. As a result, deep learning models can easily memorize dataset-specific patterns—such as staining variations, scanner artifacts, or repetitive tissue structures—rather than learning biologically meaningful histopathological features. This leads to high performance on training data but poor generalization to unseen slides or external cohorts. Overfitting is especially problematic in conventional Multiple-Instance Learning (MIL) frameworks, where only slide-level labels are available, leaving patch-level predictions unconstrained and prone to spurious correlations. Mitigating overfitting is therefore critical for building reliable and generalizable WSI classification models.

Chapter 3

Method

3.1 Introduction

This chapter discusses the methodology to be used to achieve the objectives of the thesis. Patch-level classification is a crucial task in histopathology, enabling detailed cancer diagnosis. However, it faces significant challenges due to domain shifts arising from variations in tissue preparation and imaging protocols. To address this, Test-Time Training (TTT) methods have emerged as a promising solution. These methods adapt a pre-trained model to individual test samples during inference using auxiliary self-supervised tasks, such as Masked Autoencoder (MAE) reconstruction. Despite their potential, conventional TTT approaches have a notable limitation: their initial model weights are not explicitly optimized for rapid adaptation, which can hinder their effectiveness and robustness when confronted with challenging or out-of-domain data.

We propose PathTTT, a novel framework that integrates meta-learning into the TTT paradigm to overcome this limitation. PathTTT explicitly optimizes the model's parameters for rapid adaptation by framing a self-supervised task as the inner-loop objective of a meta-learning process. This meta-training phase enables the model to learn meta-parameters that are primed for quick and effective adaptation to new test samples, significantly improving classification accuracy and robustness under domain shift. The framework leverages MAE-based self-supervision to dynamically adapt the feature extractor at test time, leading to more precise and reliable predictions.

Furthermore, we extend this idea to the challenging domain of MIL for WSIs classification. MIL

operates under a weak supervision paradigm, where only slide-level labels are available, which often leads to ambiguous instance-level supervision and weak feature representations. The aggregation of patch features in traditional MIL models frequently dilutes important discriminative signals, limiting overall predictive performance.

To solve these issues, we introduce MFE-MIL, a framework that combines MIL with a self-supervised Masked Feature Embedding (MFE) task. MFE works by randomly masking parts of patch-level feature vectors and training the model to reconstruct them. This process compels the feature extractor to learn context-aware, robust representations for each patch. This enriched feature space enhances feature discrimination and aggregation, enabling more accurate WSIs-level predictions under weak supervision.

3.2 PathTTT Framework

3.2.1 Problem Definition

In patch-level classification, we are given a dataset of N images, $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, where each image I_i is partitioned into M_i partially overlapping patches $P_i = \{p_{i1}, p_{i2}, \ldots, p_{iM_i}\}$. Each patch $p_{ij} \in \mathbb{R}^D$ represents the j-th patch extracted from image I_i, where D denotes the dimensionality of the patch representation (e.g., pixel values or feature embeddings). Every patch p_{ij} is associated with a patch-level label $y_{ij} \in \{0, 1, \ldots, C-1\}$, where C is the number of classes. The objective is to learn a mapping function

$$f : \mathbb{R}^D \to \{0, 1, \ldots, C-1\}$$

that accurately predicts the class label for unseen patches, enabling fine-grained analysis and interpretation at the local (patch) level.

3.2.2 Existing Limitation

Conventional Test-Time Training (TTT) methods [28, 14] adapt the model to each test sample by optimizing an auxiliary self-supervised objective (e.g., Masked Autoencoder reconstruction)

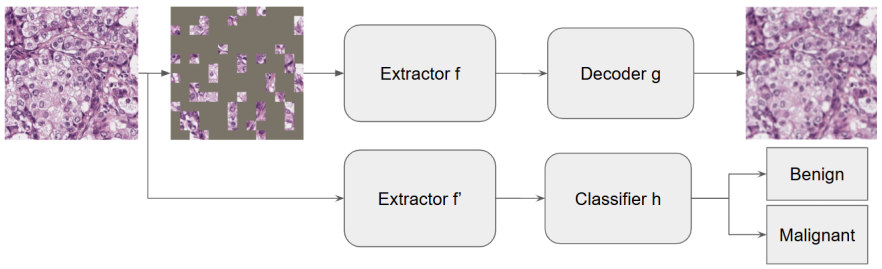


Figure 3.1: Illustration of the model architecture. The input images are first masked and processed through the shared feature extractor f and decoder g for the reconstruction task. This task updates the extractor, resulting in an enhanced version f', which is then used by the classification head h for cancer detection.

during inference. However, they suffer from two notable limitations. First, the initial weights of the encoder, decoder, and classifier are not explicitly optimized for rapid adaptation. As a result, the model may require more optimization steps or fail to adapt effectively when confronted with challenging test samples. Second, the absence of adaptation-oriented initialization reduces robustness to domain shifts, as the learned representations are primarily tuned for the training distribution. Consequently, standard TTT approaches often exhibit degraded performance when applied to data from unseen domains.

3.2.3 Methodology

Our method builds on the standard TTT framework [28, 14], as originally proposed. TTT involves adapting the model to each test sample during the test phase before making a prediction. We use the MAE [13] reconstruction task for the adaptation process. Unlike conventional training methods, TTT does not rely on the availability of the entire test dataset; instead, the model adapts to each test sample independently as it is encountered. After the network parameters are adjusted for a specific sample, the updated weights are immediately used to predict its class label. A key characteristic of the traditional TTT approach is that it does not modify the initial weights of the encoder, decoder, or classifier. However, optimizing these initial weights could significantly enhance the model's adaptability and performance. To address this, we integrate meta-learning into our training process. This integration equips the model with the ability to adapt more quickly and effectively to unseen samples, resulting in more accurate and robust predictions during the test phase.

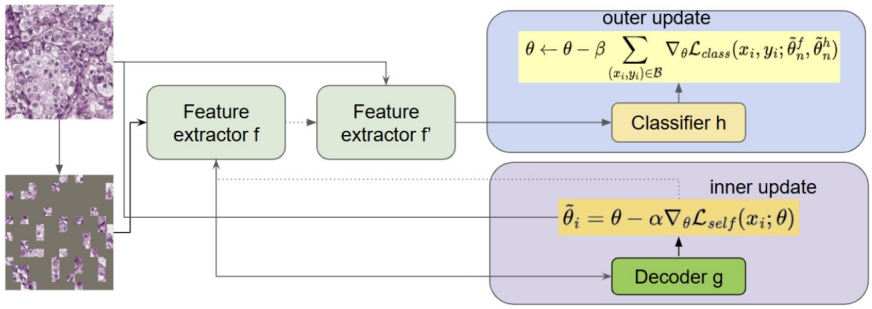


Figure 3.2: Illustration of the Meta-Auxiliary training. For a training example x_i with ground truth y_i, we update the model parameters θ using a self-supervised MAE loss in the inner loop to obtain an adapted model $\tilde{\theta}$. The adapted model is then used for the classification task. We use the supervised loss for the primary task as the meta-objective in the outer loop to update the model parameters θ.

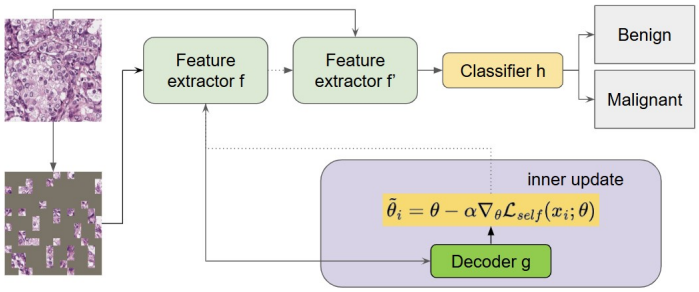


Figure 3.3: Illustration of the Meta testing. For a testing example x_i , we update the model parameters θ using a self-supervised MAE loss by the inner loop to obtain an adapted model $\tilde{\theta}$. The adapted model is then used for the classification task.

3.2.4 Architecture

The overall architecture is shown in Fig. 3.1 consisting of three components: a feature extractor f, a classification head h for the primary supervised classification, and a decoder head g for an auxiliary self-supervised task. Here, The feature extractor f acts as the encoder for the MAE, and the decoder head g corresponds to the decoder component of the MAE. Both the encoder and decoder are implemented using Vision Transformers (ViTs). The classification head h is implemented as a linear layer that maps the high-dimensional feature space generated by the encoder to the specific number of target classes. During TTT, the shared features extractor f is updated to enhance the quality of the extracted features, thereby improving the accuracy of the classification task. This adaptability allows the model to fine-tune its parameters for each test sample, optimizing performance dynamically.

3.2.5 Overall Process

Following standard practice, we begin with the MAE model [13], specifically using a ViT-Base encoder that has been pre-trained for reconstruction on the ImageNet-1k dataset. We explore four methods: supervised learning, joint training, TTT, and PathTTT.

Supervised Learning (SL). We use the same architecture as in Section 3.1, excluding the decoder. The model is trained by minimizing the cross-entropy loss function for classification, formulated as $\mathcal{L}_{\text{class}}(x_i, y_i; \theta) = l(\theta^h \circ \theta^f(x_i), y_i)$, where θ^f represents the parameters of the feature extractor, θ^h denotes the parameters of the classification head, x_i denotes the input images and the function l computes the cross-entropy loss between the predicted class probabilities and the true labels y_i.

Joint Training (JT). We compare PathTTT with joint training, using the same architecture. The self-supervised reconstruction loss is calculated by computing the pixel-wise mean squared error (MSE) between the decoded patches and their corresponding ground truth patches. The model is trained by minimizing the sum of the classification loss and self-supervised loss. The self-supervised reconstruction loss can be expressed as:

$$\mathcal{L}_{self}(x_i; \theta) = l(\theta^g \circ \theta^f(M(x_i)), x_i)) \tag{1}$$

Here θ^f represents the parameters of the feature extractor (encoder), and θ^g denotes the parameters of the decoder. $M(x_i)$ is a masked version of the input image x_i, where portions of the image are masked out to enable self-supervised learning.

During joint training, the model jointly optimizes both the main task (classification) and the auxiliary task (self-supervised reconstruction) via gradient descent:

$$\theta \leftarrow \theta - \alpha \nabla_\theta (L_{self}(x_i; \theta_0) + L_{class}(x_i, y_i; \theta_0)) \tag{2}$$

Test-Time Training. At test time, we start with the primary task head h_1 produced by joint training, as well as the encoder f_1 and decoder g_1. For each test input x, we perform test-time training (TTT) by optimizing the following loss function, focusing on the self-supervised reconstruction task. This

Algorithm 1 Test-Time Training

Require: Test dataset $\mathcal{D}_{\text{test}}$, adaptation criterion l_{self}, learning rate η.

Initialize the model with fine-tune weights:
$\theta = \{\theta^f, \theta^g, \theta^h\}$
for each test sample $x_i \in \mathcal{D}_{\text{test}}$ **do**
    Evaluate self-supervised reconstruction loss $\mathcal{L}_{self}$
    Update model parameters with gradient descent: $\tilde{\theta_i} = \theta - \alpha \nabla_\theta \mathcal{L}_{self}(x_i; \theta)$
    Make prediction $\hat{y}_i = f(x_i; \tilde{\theta}_t^f; \tilde{\theta}_t^h)$
end for

updated model is then used to make predictions:

$$\tilde{\theta}_i = \theta - \alpha \nabla_\theta \mathcal{L}_{self}(x_i; \theta). \tag{3}$$

Here, $\tilde{\theta}_i$ represents the adapted model parameters after TTT, where α is the learning rate, and L_{self} is the self-supervised reconstruction loss. The Test-Time Training algorithm is outlined in Algorithm 1.

Meta Auxiliary Training. PathTTT enhances the model's adaptability during test-time by framing the self-supervised task as a form of "learning to learn," or meta-learning. The objective of meta-training is to discover meta-parameters that enable the model to quickly adapt to new, unseen samples during testing, even in the face of unknown distribution shifts. This approach aims to ensure that the model can achieve more accurate classifications under varying conditions. The meta auxiliary training process consists of a bi-level optimization strategy:

1. Inner Loop: During the inner loop, the model is fine-tuned on individual training instances by minimizing the self-supervised loss l_{self} This process refines the model's feature extraction capabilities, making it more adept at handling the self-supervised task.

2. Outer Loop: Concurrently, the outer loop optimizes the meta-parameters across the entire training set. This alignment between the self-supervised task and the main task is crucial for ensuring that the model's meta-knowledge is robust and can generalize effectively to new data and tasks. The overall process as Figure 3.2 shown. This bi-level optimization ensures the model will have the ability to rapidly adapt during test-time, allowing it to deal with distribution shifts and varying

Algorithm 2 Meta Auxiliary Training

Require: Train dataset $\mathcal{D}_{\text{train}}$, learning rates β, α.

Initialize the model with fine-tune weights: $\theta = \{\theta^f, \theta^g, \theta^h\}$
while not converged **do**
    Sample a batch $\mathcal{B}$ of data from D_{train}
    for each training instance $x_i, y_i \in \mathcal{B}$ **do**
        Evaluate self-supervised reconstruction loss $\mathcal{L}_{self}(x_i; \theta)$
        Compute adapted parameters with gradient descent: $\tilde{\theta}_n = \theta - \alpha \nabla_\theta \mathcal{L}_{self}(x_i; \theta)$
    end for
    Compute task-specific classification loss and update the global model:
    $\theta \leftarrow \theta - \beta \sum_{(x_i, y_i) \in \mathcal{B}} \nabla_\theta \mathcal{L}_{class}(x_i, y_i; \tilde{\theta}_n^f, \tilde{\theta}_n^h)$
end while
return Optimized model parameters θ

data complexities more effectively. The overall procedure for meta-training is summarized in Algorithm 2. This approach combines the benefits of meta-learning with test-time training, making it highly versatile in dealing with distribution shifts and varying data complexities.

Meta Testing. During the test phase, the PathTTT procedure mirrors the Test-Time Training process, where the model is continually adapted to each test sample using the learned meta-parameters as Figure 3.3 shown.

3.3 MFE-MIL Framework

3.3.1 Problem Definition

In Multiple Instance Learning (MIL), we are given a training set of N bags, $\mathcal{B} = \{B_1, \ldots, B_N\}$, where each bag $B_i = \{x_{i1}, x_{i2}, \ldots, x_{iM_i}\}$ consists of M_i instances, and $x_{ij} \in \mathbb{R}^D$ is the j-th instance in bag B_i. Each bag B_i is associated with a bag-level label $Y_i \in \{0, 1, \ldots, C-1\}$, where C is the number of classes. The goal is to learn a function $f : \mathcal{B} \to \{0, 1, \ldots, C-1\}$ that predicts the label for unseen bags.

3.3.2 Existing Limitations

The Limitation of Standard MIL

Standard MIL models operate under a weak supervision paradigm, which presents significant limitations, particularly in the analysis of WSIs. A core challenge is that MIL models are only provided with a single "bag-level" label for an entire WSI—for example, a binary classification of "cancerous" or "benign." This approach fundamentally lacks the fine-grained supervision needed to learn highly discriminative features at the instance (or patch) level. The model is unable to definitively identify which specific patches are contributing to the positive or negative bag label, leading to inherent ambiguity.

Consequently, the learned features for each individual patch are often suboptimal. This "bottleneck" in feature learning directly impacts the overall performance of the model, as it struggles to make precise and accurate predictions on the WSI as a whole. While the model may learn to identify the most salient or obvious instances, this ambiguity makes it difficult to capture the subtle, yet important, features that may be present in other patches. Addressing these limitations requires a shift toward methods that can introduce more robust, patch-level supervision or better manage the inherent ambiguity of the MIL framework.

The Limitation of Combining SSL and MIL

Combining SSL with MIL offers a powerful paradigm for representation learning in weakly supervised settings like computational pathology. While this approach has shown significant promise, its effective application is contingent upon addressing several key limitations that are critical for discussion. These challenges can be broadly categorized into conceptual, technical, and practical considerations.

Conceptually, a primary limitation lies in the fundamental mismatch between the objectives of SSL and MIL. SSL pretext tasks are typically designed to learn robust instance-level features, such as through contrastive learning or masked image modeling. However, the final MIL decision relies on aggregating these patch-level signals to make a bag-level prediction. The features learned via

SSL, while powerful for individual instances, are not explicitly optimized to facilitate this aggregation, which can lead to suboptimal bag-level performance. Furthermore, SSL does not resolve the inherent ambiguity of MIL—the exact positive and negative instances within a bag remain unknown. This means that even with discriminative patch embeddings, the model can still be influenced by redundant or non-informative patches, particularly if they constitute a large portion of the bag.

Technically, the aggregation bottleneck remains a significant challenge. Strong, discriminative features learned through SSL can still be "washed out" by simplistic aggregation functions like mean or max pooling, especially in bags containing a large number of non-informative patches. This undermines the full potential of the learned representations. Another technical hurdle is the sensitivity to domain shift. While SSL is often lauded for its ability to learn generalizable features from large unlabeled datasets, in domains like pathology, it can inadvertently learn domain-specific low-level cues (e.g., staining patterns) rather than clinically relevant biological signals. This can degrade performance when transferring a pretrained model to a new dataset with a different domain, such as a different hospital or scanning protocol. Finally, the joint optimization of SSL and MIL can be unstable. For instance, an SSL task might prioritize invariance to transformations like rotation, whereas the MIL task might benefit from orientation-specific features, creating conflicting gradients that hinder stable convergence.

From a practical standpoint, the high computational cost of this combined approach is a notable limitation. SSL often necessitates extensive data augmentation and large batch sizes, while MIL involves processing large bags with hundreds or thousands of patches. Combining these two paradigms drastically increases GPU memory requirements and training time, making end-to-end training computationally prohibitive for many researchers. Moreover, the distribution of unlabeled data used for SSL may not align with the bag-level class distribution in the MIL task. Without careful sampling, this can lead to a bias in the learned representations, favoring patterns from the majority of the unlabeled data rather than the clinically relevant signals in the MIL bags. Lastly, while SSL can enhance feature quality, the learned representation space can be difficult to interpret from a biological perspective, especially when the pretext task is abstract (e.g., reconstruction). This limits the ability to use the model's outputs for actionable insights or to validate its biological relevance.

In conclusion, while the integration of SSL and MIL represents a promising direction for weakly supervised histopathology, its success is limited by challenges related to feature-task mismatch, aggregation bottlenecks, and domain shift, alongside practical constraints related to computational cost and optimization stability.

3.3.3 Methodology

The core methodology of MFE-MIL is a two-phase learning strategy designed to overcome the limitations of standard MIL on WSIs and Combining SSL and MIL. The primary challenge in MIL is the lack of fine-grained, instance-level supervision, which leads to ambiguous feature representations.

MFE-MIL utilizes a novel self-supervised task called Masked Feature Embedding (MFE), which is jointly optimized with the weakly supervised MIL classification task. This allows the model to simultaneously learn a highly discriminative feature space while also performing bag-level prediction. The MFE task operates by randomly masking a portion of a patch's feature vector and forcing the model to reconstruct the missing information. This creates a strong, predictive learning signal that compels the feature extractor to learn a context-aware representation for each patch. This enriched feature space provides a robust foundation for the MIL classification, enabling the model to make more precise and accurate predictions.

3.3.4 Architecture

The MFE-MIL framework is built upon a modular architecture consisting of the following key components:

- WSIs Pre-processing: This initial module is responsible for converting a gigapixel WSI into a manageable set of image patches. This involves tiling the image at a specific magnification and applying a tissue-filtering step to discard patches that contain no significant tissue information, thereby reducing computational load as shown in Figure 3.4.

- Feature Extractor : This is a neural network, typically a CNN or a Vision Transformer like UNI or CONCH, that maps each individual tissue patch to a high-dimensional feature vector

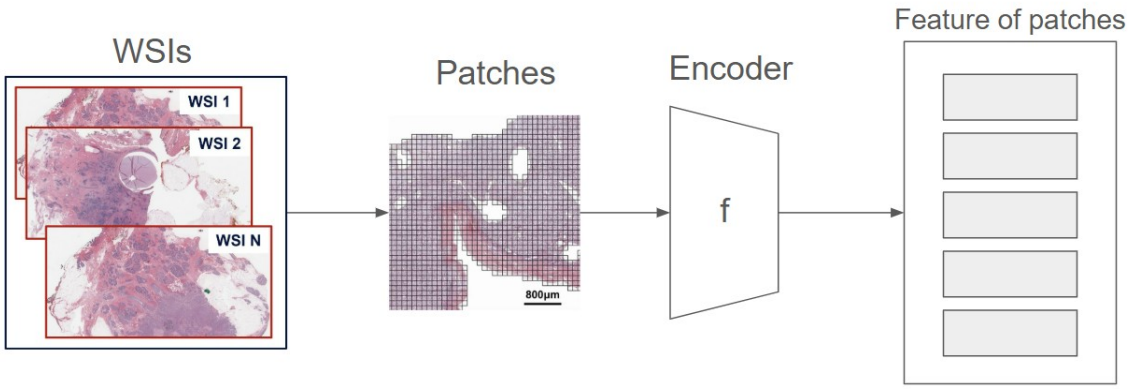


Figure 3.4: WSIs processing and Feature Extractor. Each WSI is divided into smaller patches to reduce computational complexity while preserving local tissue information exclude the vast, empty background. The pre-trained foundation encoder (Uni or Conch) extracts high-quality features for each patch. The encoder weights are frozen, ensuring stable and generalizable representations.

as shown in Figure 3.4.

- MFE Module: This is the self-supervised core. It is comprised of two parts:

  Feature Masking: A strategy that randomly replaces a percentage of the feature vectors in a bag with a special, learnable Mask Token.

  Transformer Decoder: A Transformer-based network that takes the partially masked bag of features (along with positional embeddings) and reconstructs the original, masked feature vectors.

- MIL Classification Head: This component is used in the final downstream task. It consists of an aggregation function that combines the bag of instance features into a single bag-level representation, followed by a simple classifier (e.g., a multi-layer perceptron) that predicts the final WSI label as shown in Figure 3.5.

3.3.5 Overall Process

The entire MFE-MIL process is a two-stage pipeline:

Phase 1: Encoding

(1) A WSI is pre-processed into a bag of tissue patches.

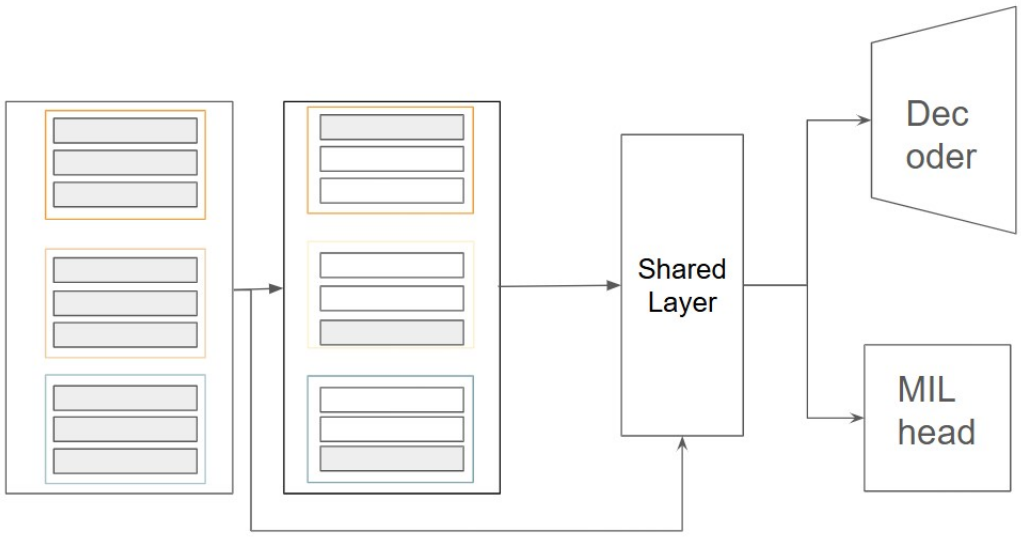


Figure 3.5: MFE and MIL. A portion of the instance features is randomly masked, and a reconstruction head predicts the masked features. Instance features are aggregated to form a slide-level representation using standard MIL pooling strategies. The aggregated representation is passed through a classification head to predict the slide-level label. The model is trained with a combined loss that includes the slide-level classification loss and the instance-level reconstruction loss, balancing discriminative learning and feature generalization.

(2) A pre-trained foundation model, such as UNI or CONCH, is used as the Feature Extractor (f_θ) to generate a bag of feature vectors

$$B = \{z_1, z_2, \ldots, z_N\}.$$

Phase 2: Downstream MIL Classification and SSL Task

(1) A predefined proportion of these feature vectors are randomly masked and replaced with the Mask Token (z_{mask}).

(2) The unmasked features and mask tokens are passed through the Transformer Decoder, which attempts to reconstruct the original values of the masked features.

(3) A Mean Squared Error (MSE) loss is calculated between the reconstructed features ($\hat{Z}_{\text{masked}}$)

and the original features (Z_{masked}):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{z}_i - z_i\|_2^2,$$

where $\mathcal{M}$ denotes the set of masked feature indices.

(4) The MIL Classification Head's aggregation function (e.g., attention pooling) aggregates these features into a single vector:

$$v = \text{Aggregate}(z).$$

(5) Both MSE reconstruction loss and MIL classification loss are combined to fine-tune the weights (θ) of the shared layer, decoder and MIL head, adapting them to learn powerful, context-aware representations specifically for the WSI domain.

(6) After training, The classifier makes a final prediction based on the aggregated representation, producing the WSI-level label $\hat{y}$.

Chapter 4

Experiment and Result

4.1 Experimental of PathTTT

To demonstrate the performance of the PathTTT with MAE in addressing domain shift, we performed binary classification experiments (benign vs. malignant) on two prevalent cancer types: Breast Cancer and Colon Cancer. The datasets utilized in this study, described in detail below and summarized in Table 4.1, were selected to assess cross-domain generalization.

A critical consideration in processing pathology whole slide images (WSIs) is ensuring a consistent field of view (FoV) across varying microns per pixel (Mpp) resolutions.

$$\text{FOV} = \text{Mpp} \times \text{Patch Size}$$

WSIs are often scanned at different Mpps, and extracting fixed pixel-sized patches without accounting for resolution differences would yield inconsistent physical coverage. To maintain fairness in model training and evaluation, we standardized patches to represent the same real-world area (FOV $= 112\mu m$) across all datasets, as shown in Table 4.1.

For each dataset, we applied a standard split, allocating 70% of the data for training, 10% for validation, and 20% for testing. All experiments were conducted on NVIDIA A100 GPUs.

4.1.1 Datasets

Breast Cancer Datasets

BACH: The Grand Challenge on Breast Cancer Histology Images (BACH) dataset [3] is a publicly available collection of H&E-stained breast histology microscopy images. he dataset includes images labeled as normal, benign, in situ carcinoma, or invasive carcinoma. Each image is provided in the RGB color model with a resolution of 2048 x 1536 pixels and a pixel scale of 0.42 µm x 0.42 µm. For our experiments, we focused on the Normal and Invasive Carcinoma classes, treating them as benign and malignant categories, respectively. To standardize the input, images were cropped to 267 × 267 pixels, and noise along with irrelevant background was removed to improve the quality of the data used for training and evaluation.

BRACS: The BReAst Carcinoma Subtyping (BRACS) dataset [4] is a large cohort of annotated H&E-stained breast histology images, designed to aid in the detailed characterization of breast lesions. BRACS contains 4539 Regions of Interest (ROIs) extracted from the Whole Slide Images (WSIs). The dataset is categorized into seven subtypes, with ROI images which easily exceed 4,000 × 4,000 pixels with a pixel scale of 0.25 µm/pixel at a 40× resolution. In our study, we selected the Normal and Invasive Carcinoma classes to represent benign and malignant categories, respectively. The images were cropped to 448 × 448 pixels, and noise, as well as background artifacts, were meticulously removed to ensure the data's suitability for the classification tasks.

Colorectal Cancer Datasets

NCT-CRC-HE-100K: The NCT-CRC-HE-100K dataset [5] consists of 100,000 non-overlapping image patches derived from 86 H&E stained human cancer tissue slides and normal tissue samples. Each image measures 224 × 224 pixels at a resolution of 0.5 µm/pixel. Expert pathologists have meticulously annotated the tissue regions in the whole slide images, categorizing them into nine distinct tissue classes. For our experiments, we use the Normal Colon Mucosa (NORM) class as the benign category and the Colorectal Adenocarcinoma Epithelium (TUM) class as the malignant category.

Table 4.1: Overview of datasets used in the experiments. Mpp (μm/pixel) stands for "Microns per Pixel". It refers to the spatial resolution of the images in the dataset, indicating how much real-world distance (in microns) corresponds to a single pixel in the image.

Dataset	Patch Size	Mpp(μm/pixel)	FOV(μm)	#Patch
BACH	267×267	0.42	112	26000
BRACS	448×448	0.25	112	30000
NCT	224×224	0.5	112	17400
Chaoyang	224×224	0.5	112	24000
MedFMC	236×236	0.475	112	200000

Chaoyang: The Chaoyang dataset [6] comprises colon slides obtained from Chaoyang Hospital, scanned at a ×20 objective magnification to produce WSIs, which were then divided into 512 × 512 pixel patches. The dataset contains four different classes. For our study, we selected the normal and adenoma samples as benign and malignant classes, he patches were cropped to 224 × 224 pixels, and noise along with background artifacts were removed to enhance the data quality for classification tasks.

MedFMC: The MedFMC Colon Path dataset [7] was originally gathered from Ruijin Hospital in China. The tissue slides, acquired between 2017 and 2019, were stained with H&E. After scanning, the WSIs were rescaled to a magnification of ×20, with a pixel resolution of 0.475 μm. This dataset includes 10,009 large tissue patches, each measuring 1024×1024 pixels, obtained from the colonoscopy pathology examinations of 396 patients, with both tumor and non-tumor labels. For our analysis, the images from the MedFMC dataset were cropped into 236 × 236 pixels, with noise and background artifacts removed to ensure a cleaner dataset for the classification process.

4.1.2 Implement detail

Both supervised learning and joint training, following [13], using the AdamW optimizer, setting a base learning rate of 1e-5 and using momentum values of 0.9 and 0.95. To prevent overfitting, a weight decay regularization strength of 0.05 is applied. Additionally, we use a cosine decay schedule to gradually reduce the learning rate over time, aiding in fine-tuning the model toward convergence. Both supervised learning and joint training are trained for 100 epochs with a batch size of 32. Early stopping is implemented to prevent overfitting, terminating the training if the

Table 4.2: Classification accuracy (%) for breast cancer classification. "BACH→BRACS" means training on the BACH dataset and testing on the BRACS dataset. The highest accuracy is in bold.

	SOTA			TTA		Ours			
Top1-ACC	FCCS	CRCCN	EV2	SAR	EATA	SL	JT	TTT	PathTTT
BACH→BRACS	56.93	57.90	65.93	68.37	56.80	66.8	66.42	67.00	**69.73**
BRACS→BACH	67.46	60.56	69.77	73.82	74.75	75.19	75.46	75.57	**76.07**
Avg	62.20	59.23	67.85	71.10	65.78	71.00	70.94	71.29	**72.90**

validation loss does not improve for 15 consecutive epochs.

In test-time training, we use the SGD optimizer with a momentum value of 0.9, and a base learning rate of 1e-4.

In meta-auxiliary training, we apply the AdamW optimizer with a base learning rate of 1e-4.

Unless specified otherwise, our experiments involve training using the following modifications [29] to the default method-specific augmentation scheme:

- Random vertical flipping with a probability of 0.5.
- Color dropping with a probability of 0.2, where the images are randomly converted to grayscale.
- Weak color jittering with a probability of 0.8, involving random adjustments to the brightness, contrast, saturation, and hue of the images, with respective strengths of 0.2, 0.2, 0.2, and 0.1.

4.1.3 Main result

Our main results for Breast Cancer and Colon Cancer datasets are presented in Tables 4.2 and 4.3. These experiments were conducted using the default training setup discussed in the previous section. In these experiments, models were trained on one dataset and tested on a different dataset to evaluate their robustness to domain shifts. PathTTT, applied on top of our baseline, demonstrates significant improvements in performance, effectively addressing the challenges posed by domain shifts.

The results presented in Table 4.2 and Table 4.3 provide a detailed comparison of classification performance under these conditions. Each row in the tables corresponds to a dataset transition, demonstrating how well the models handle different data distributions. These transitions simulate real-world scenarios where the testing data comes from a distribution distinct from the training data,

Table 4.3: Classification accuracy (%) for colorectal cancer classification. "NCT→Chaoyang" means training on the NCT dataset and testing on the Chaoyang dataset. The highest accuracy is in bold.

	SOTA			TTA		Ours			
Top1-ACC	FCCS	CRCCN	EV2	SAR	EATA	SL	JT	TTT	PathTTT
NCT→Chao	51.08	52.9	53.33	51.21	50.17	52.79	51.83	55.06	**56.33**
NCT→MedFMC	72.84	66.47	71.98	76.02	69.57	72.9	75.91	76.07	**77.17**
Chao→NCT	66.72	76.06	71.67	81.09	75.57	82.41	88.45	88.47	**88.59**
Chao→MedFMC	69.68	68.06	59.89	67.48	55.39	65.29	67.30	68.40	**69.74**
MedFMC→NCT	68.65	79.25	55.57	76.24	77.18	75.09	73.22	80.89	**83.10**
MedFMC→Chao	85.21	86.15	80.62	84.44	84.73	84.35	85.69	86.02	**86.17**
Avg	69.03	71.48	65.51	72.75	68.77	72.14	73.73	75.81	**76.85**

providing insights into the robustness of each method. These results were obtained using various SOTA methods, TTA strategies, baseline methods and our proposed PathTTT method. Across all transitions, PathTTT consistently outperforms other methods under domain shift scenarios. This suggests that PathTTT is particularly well-suited for tasks where the testing data distribution differs significantly from the training data distribution, highlighting its potential for improving generalization in real-world applications.

Further evidence supporting PathTTT's effectiveness is provided by the t-SNE visualizations in Fig. 1.2, there is noticeable overlap between clusters from different datasets, distinct dataset-specific structures remain evident. These differences reduce the generalization capability of models trained on one dataset when applied to another. PathTTT, however, demonstrates its ability to bridge these gaps effectively, as evidenced by its superior performance in transitions like BACH→BRACS (69.73%) and BRACS→BACH (76.07%). This ability to adapt to unseen test distributions, even under significant domain shifts, underscores the robustness and adaptability of PathTTT.

A deeper examination of the datasets also reveals varying levels of cross-dataset compatibility. For example, the relatively high cross-dataset performance between NCT and MedFMC can be attributed to the significant overlap in their feature distributions. In contrast, the limited overlap between Chaoyang and NCT highlights the difficulty of adapting to Chaoyang, where domain shifts are more pronounced.

Table 4.1 and Table 4.3 show larger datasets like MedFMC improve generalization by capturing diverse features, which capture more diverse feature representations, tend to improve generalization.

Table 4.4: Classification accuracy (%) for non-cross domain. The highest accuracy is in bold.

	SOTA			TTA		Ours			
Top1-ACC	FCCS	CRCCN	EV2	SAR	EATA	SL	JT	TTT	PTTT
BACH	99.33	96.46	88.60	99.59	97.98	**99.75**	99.54	99.59	99.61
BRACS	**89.00**	81.52	74.80	86.38	83.63	81.00	88.66	88.51	88.4
NCT	99.22	98.30	90.52	99.36	**99.54**	99.51	99.37	99.19	99.25
Chaoyang	98.04	91.52	94.77	98.22	99.18	**98.66**	98.10	98.19	98.17
MedFMC	98.92	95.90	95.77	99.52	97.57	99.46	**99.59**	98.82	98.85
Avg	96.90	92.74	88.89	96.61	95.58	95.68	**97.05**	96.86	**97.05**

However, even with such datasets, distinct clusters observed in the t-SNE plots and the corresponding performance drops emphasize the persistent challenge of adapting to domain shifts. PathTTT effectively addresses this by dynamically updating the model at test time, enabling robust performance across datasets with varying feature distributions.

When compared to existing SOTA methods, PathTTT provides a significant performance boost. PathTTT improves generalization across different datasets. For example, it outperforms EV2 with an accuracy of 69.73% in the BACH→BRACS setup and 76.07% in the BRACS→BACH setup. This superior performance highlights PathTTT's ability to adapt during testing, addressing domain shift issues more effectively than the SOTA methods.

We also compare our method to other test time adaptation (TTA) methods such as EATA and SAR, as shown in Table 4.2 and 4.3. PathTTT consistently outperforms all other methods across various datasets, proving its effectiveness in handling domain shifts. TTT often ranks second, highlighting its adaptability. In contrast, SAR and EATA sometimes show lower performance, further emphasizing the advantages of PathTTT in handling complex domain shifts.

4.1.4 Ablation Study

In this section, we conduct ablation studies on PathTTT to analyze various components of the proposed method.

Contribution of each component. We isolate the contributions of MAE, TTT and MAML to determine their individual impact.

As Tables 4.2 and 4.3 show, JT (equipped with MAE), TTT and PathTTT (Combines JT, TTT

and MAML) demonstrate progressive improvements, validating the distinct roles of each component. For breast cancer classification, JT achieves 66.42% and 75.46%, surpassing SL and highlighting MAE's reconstruction-driven alignment for domain invariance. PathTTT, combining TTT with MAML, further improves performance to 69.73% and 76.07%, demonstrating MAML's rapid adaptation to domain shifts. For colorectal cancer, JT shows strong generalization (e.g., 75.91% for NCT→MedFMC, 88.45% for Chaoyang→NCT), while PathTTT outperforms it with 77.17% and 88.59%, confirming MAML's adaptability. In the challenging MedFMC→NCT setting, PathTTT achieves 83.10% (+2.21% over TTT), emphasizing MAML's role in few-shot generalization.

Futhermore, the results in Table 4.4 demonstrate that PathTTT maintains competitive performance in non-cross-domain classification tasks. Across five datasets, PathTTT achieves accuracy comparable to or slightly lower than the best-performing methods. Notably, PathTTT attains the highest average accuracy (97.05%). This indicates that PathTTT does not negatively impact performance in scenarios without domain shift, ensuring its effectiveness across diverse settings.

Ablation studies further confirm the significance of our method's components. JT, integrating MAE, enhances adaptation through self-supervised learning, improving feature representations under distribution shifts. TTT further enhances adaptation by updating the encoder during inference, outperforming static baselines. Finally, PathTTT, which integrates MAML, optimizes the model to be more adaptable, ensuring rapid fine-tuning in novel domains. The combined effect of these components in PathTTT leads to superior performance compared to their individual contributions. While PathTTT requires approximately four additional hours of meta-training on the NCT dataset compared to JT, this computational cost is justified by its substantial gains in adaptability and performance.

4.2 Experimental of MFE-MIL

4.2.1 Datasets

We evaluate the proposed MIL and SSL methods on 3 different histopathological datasets– Camelyon16 [30], PANDA [31] and TCGA-BRCA. To reduce the impact of data split on model evaluation, we implement a 3-fold cross-validation approach, partitioning the data into training,

validation and testing subsets in a 8:1:1 ratio except CAMELYON16.

CAMELYON16: A widely used whole slide image (WSIs) dataset for breast cancer metastasis detection, comprising 400 WSIs split into 214 training. 54 validation and 128 testing slides.

PANDA : is a large-scale dataset for prostate cancer grading, containing 10307 digitized biopsy slides categorized into six Gleason grades: grade 0 (2783 slides), grade 1 (2615), grade 2 (1,324), grade 3 (1,212), grade 4 (1196), and grade 5 (1177).

TCGA-BRCA : is a breast cancer dataset containing 1,033 hematoxylin and eosin (H&E) whole slide images, categorized into two subtypes: invasive ductal carcinoma (822 slides) and invasive lobular carcinoma (211 slides).

4.2.2 Implementation Details

We conduct a comprehensive evaluation of the proposed MFE-MIL framework across three publicly available datasets, benchmarking its performance against a range of multiple instance learning (MIL) approaches. The comparison methods are categorized as follows:

- **Conventional Pooling Methods:** We adopt *MeanPooling* and *MaxPooling* as representative baseline strategies to evaluate the effectiveness of our proposed approach over traditional MIL pooling operations.
- **Attention-Based Methods:** This category includes the *Attention-Based MIL* (ABMIL) model and its advanced variants, specifically *CLAM-MB* [32] and *CLAM-SB*, which leverage attention mechanisms to identify and weight diagnostically relevant instances.

To ensure a fair and consistent comparison, we strictly follow the data pre-processing pipeline described in the original **CLAM** implementation. For all models, we employ the Adam optimizer with a fixed learning rate of 1×10^{-4} and default momentum parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). We maintain identical training hyperparameters across all methods, including the batch size, number of epochs, and weight initialization strategy, unless otherwise stated. Model selection is performed

based on the highest validation F1, and the final performance is reported as the average across three independent runs with different random seeds to mitigate variability due to stochastic optimization.

All experiments are conducted on an NVIDIA A100 GPU. The source code and implementation details for MFE-MIL will be made publicly available to facilitate reproducibility.

4.2.3 Main Result

As shown in Table 4.5, the proposed MFE-MIL method consistently and robustly improves the performance of a variety of MIL baselines across three benchmark datasets—Camelyon16, PANDA, and TCGA-BRCA—under both UNI and CONCH encoder settings. On Camelyon16, the largest gains are observed when MFE-MIL is applied to simpler aggregation strategies. For example, with the UNI encoder, the MEAN pooling baseline achieves only 68.23% ACC, 61.49% F1, and 62.45% AUC; integrating MFE-MIL boosts these scores by +9.90%, +13.17%, and +8.48%, respectively. The effect is even more pronounced with CONCH-MEAN, where the ACC, F1, and AUC increase by +10.41%, +13.68%, and a substantial +18.21%. This indicates that MFE-MIL is particularly effective in enhancing weaker aggregation mechanisms that may underutilize patch-level information. While high-performing attention-based models such as ABMIL and CLAM start from stronger baselines, they still show measurable improvements—e.g., UNI-CLAM-SB achieves +3.26% F1 and +2.87% AUC gains, and CONCH-ABMIL improves by +1.05% ACC and +1.21% F1—suggesting that MFE-MIL can complement even sophisticated attention pooling.

Performance gains on PANDA follow a similar trend, with all baselines benefiting from MFE-MIL. The largest improvement is seen in CONCH-CLAM-MB, which increases by +2.94% ACC and +3.35% F1, demonstrating that the method also generalizes to prostate cancer histology tasks with different data distributions. Smaller but consistent gains are found for other models, including attention-based ones, where even +0.16% ACC (UNI-CLAM-SB) may be meaningful given the dataset's size and difficulty. On TCGA-BRCA, where only the CONCH encoder is evaluated, MFE-MIL continues to show utility despite the already high baseline performance. Notably, CONCH-MAX improves by +3.22% ACC, +4.20% F1, and +1.75% AUC, and CONCH-CLAM-SB gains +2.61% ACC and +3.48% F1, underscoring its ability to refine decision boundaries even when initial classification accuracy is above 90%.

Overall, the results highlight three key strengths of MFE-MIL:

- It yields the largest benefits for simpler pooling strategies, making them more competitive with attention-based approaches;
- It still provides measurable improvements for strong baselines, demonstrating complementary effects; and
- It generalizes across datasets with distinct cancer types and image characteristics, confirming its robustness and adaptability to diverse MIL settings in computational pathology.

4.2.4 Ablation Study

This section presents an ablation study to analyze the contributions of various components within the proposed PathTTT method. We systematically evaluate the impact of the following key elements on the model's overall performance:

- Weighting for Joint Training: We analyze the sensitivity of the model to the weight assigned to each objective in the joint training framework.
- Masking Ratio: We investigate the effect of varying the percentage of masked feature embeddings during training.
- Shared Layer: We evaluate the role of the shared layer in the model's architecture and its effect on performance and parameter efficiency.

Weighting for Joint Training

We explore the effect of different weighting schemes for the joint training objective, which combines two losses: a contrastive loss for instance-level representation learning and a classification loss for slide-level prediction. A hyperparameter λ is used to control the balance between these two objectives. We evaluate three values for λ: 0.1, 0.2, and 0.3. The results, as shown in Table 4.6, indicate that a weight of $\lambda = \mathbf{0.3}$ yields the best average performance across both CAMELYON16

Table 4.5: Performance (%) Comparison on Camelyon16, PANDA, and TCGA-BRCA datasets. MFE-MIL improves multiple baseline MIL models.

Method		**Camelyon16**			**PANDA**			**TCGA-BRCA**		
		ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
UNI	MEAN	68.23	61.49	62.45	79.05	76.18	94.79	91.44	87.13	96.62
	+MFE-MIL	**78.13**	**74.66**	**70.93**	**80.08**	**76.7**	**95.01**	**94.40**	**92.09**	**96.81**
	Δ	+9.9	+13.17	+8.48	+1.03	+0.52	+0.22	+2.96	+4.96	+0.21
	MAX	91.93	91.13	96.33	70.45	64.59	91.61	92.69	89.96	97.05
	+MFE-MIL	**96.61**	**96.39**	**99.19**	**71.58**	**65.17**	**92.74**	**94.73**	**91.86**	**97.30**
	Δ	+4.68	+5.26	+2.86	+1.13	+0.58	+1.13	+2.04	+1.90	+0.25
	ABMIL	94.79	94.43	97.83	75.83	71.40	94.71	92.37	87.98	97.40
	+MFE-MIL	**95.83**	**95.61**	**97.68**	**76.53**	**72.48**	**94.65**	**93.06**	**89.48**	**97.77**
	Δ	+1.14	+1.18	-0.15	+1.30	+1.08	-0.06	+0.69	+1.50	+0.37
	CLAM-MB	91.93	91.28	95.9	76.88	72.99	94.54	92.74	89.15	97.15
	+MFE-MIL	**95.57**	**95.23**	**95.56**	**77.17**	**72.79**	**94.99**	**94.06**	**90.95**	**98.06**
	Δ	+3.64	+3.95	-0.34	+0.29	-0.20	+0.45	+1.32	+1.80	+0.91
	CLAM-SB	92.45	91.78	95.7	77.11	73.01	94.86	93.39	90.43	97.69
	+MFE-MIL	**95.31**	**95.04**	**98.57**	**77.27**	**73.59**	**95.03**	**93.71**	**90.94**	**98.10**
	Δ	+2.86	+3.26	+2.87	+0.16	+0.58	+0.17	+0.32	+0.51	+0.41
CONCH	MEAN	66.41	60.23	59.97	71.48	67.23	92.38	91.07	87.00	95.80
	+MFE-MIL	**76.82**	**73.91**	**78.17**	**74.10**	**69.51**	**92.46**	**91.37**	**87.90**	**95.99**
	Δ	+10.41	+13.68	+18.21	+2.62	+2.28	+0.08	+0.30	+0.90	+0.19
	MAX	95.83	95.54	98.53	63.98	57.69	88.91	88.48	83.85	95.50
	+MFE-MIL	**96.10**	**95.84**	**99.25**	**64.72**	**58.19**	**89.06**	**91.70**	**88.05**	**97.25**
	Δ	+0.27	+0.30	+0.72	+0.74	+0.50	+0.15	+3.22	+4.20	+1.75
	ABMIL	95.05	94.66	96.38	69.77	64.97	92.26	92.41	88.80	96.70
	+MFE-MIL	**96.10**	**95.87**	**97.86**	**71.64**	**66.16**	**92.61**	**94.03**	**91.29**	**97.71**
	Δ	+1.05	+1.21	+1.48	+1.87	+1.19	+0.35	+1.62	+2.49	+1.01
	CLAM-MB	93.49	92.87	94.87	69.70	64.16	92.36	92.07	88.48	97.12
	+MFE-MIL	**95.05**	**94.69**	**96.91**	**72.64**	**67.51**	**92.97**	**92.35**	**89.06**	**97.77**
	Δ	+1.56	+1.82	+2.04	+2.94	3.35	+0.61	+0.28	+0.58	+0.65
	CLAM-SB	95.06	94.81	97.54	70.16	64.80	92.27	91.10	87.20	97.47
	+MFE-MIL	**95.31**	**94.98**	**97.76**	**72.13**	**67.14**	**92.67**	**93.71**	**90.68**	**97.46**
	Δ	+0.25	+0.17	+0.22	+1.97	+2.34	+0.40	+2.61	+3.48	-0.01

and PANDA datasets. This suggests that placing slightly more emphasis on the slide-level classification objective, while still leveraging instance-level contrastive learning, is beneficial for the model's overall performance.

Table 4.6: Performance (AUC %) comparison of different weight settings on CAMELYON16 and PANDA datasets across five methods and their average.

Weight	Camelyon16						PANDA					
	MEAN+	MAX+	AB.+	.MB+	.SB+	Avg	MEAN+	MAX+	AB.+	.MB+	.SB+	Avg
0.1	73.96	97.42	94.08	98.48	94.73	91.73	94.86	91.30	94.20	93.88	94.51	93.75
0.2	70.06	94.94	99.61	97.00	91.23	90.97	95.15	91.06	94.56	94.35	94.55	93.93
0.3	70.93	99.19	97.68	95.56	98.57	**92.79**	95.01	92.74	94.65	94.99	95.03	**94.48**

Masking Ratio

The masking ratio is a crucial hyperparameter that determines the percentage of feature embeddings that are randomly masked during training. A higher masking ratio forces the model to learn more robust and contextual representations, as it must reconstruct a greater portion of the input from limited information. We test several masking ratios, ranging from 0.6 to 0.8. As shown in Table 4.7, the optimal performance is achieved with a masking ratio of 0.75. This finding indicates that a moderate to high masking ratio is effective in preventing the model from over-relying on local information and encourages the learning of global semantic features, which is essential for accurate classification in Whole Slide Images.

Table 4.7: Performance (AUC %) comparison of different mask ratio settings on CAMELYON-16 and PANDA datasets across five methods and their average.

Mask Ratio	CAMELYON-16						PANDA					
	MEAN+	MAX+	AB.+	.MB+	.SB+	Avg	MEAN+	MAX+	AB.+	.MB+	.SB+	Avg
0.6	70.07	96.06	96.72	97.93	95.14	91.98	95.22	91.48	94.51	94.05	94.79	94.41
0.65	68.32	97.39	97.88	93.85	97.70	91.83	94.71	91.32	94.66	93.72	94.52	93.79
0.7	71.40	98.50	97.42	99.04	97.18	92.71	95.46	91.89	94.46	93.55	94.59	93.99
0.75	70.93	99.19	97.68	95.56	98.57	**92.79**	95.01	92.74	94.65	94.99	95.03	**94.48**
0.8	73.65	99.82	97.93	90.46	97.08	91.99	94.88	91.14	94.48	94.68	94.04	93.84

Shared Layer

The shared layer in our architecture is designed to reduce the model's complexity and encourage the learning of generalizable features by sharing parameters between different stages. We evaluate the impact of the number of shared layers on performance and efficiency. We compare models with 0, 1, and 2 shared layers. Table 4.8 shows that the model with 2 shared layers achieves the highest average AUC on both datasets. This result suggests that parameter sharing, when appropriately implemented, not only improves computational efficiency but also enhances the model's ability to learn more robust and effective representations by enforcing a common representational bottleneck. The performance degradation with 0 or 1 shared layers highlights the importance of this architectural choice for the success of our method.

Table 4.8: Performance (AUC %) comparison of different shared layer settings on CAMELYON16 and PANDA datasets across five methods and their average.

Shared Layer	**CAMELYON16**						**PANDA**					
	MEAN+	**MAX+**	**AB.+**	**.MB+**	**.SB+**	**Avg**	**MEAN+**	**MAX+**	**AB.+**	**.MB+**	**.SB+**	**Avg**
0	54.79	99.46	94.11	94.78	96.56	87.94	95.42	92.04	94.99	94.33	95.05	94.37
1	63.19	98.58	99.56	97.13	95.74	90.84	94.98	91.35	94.27	94.27	94.73	93.92
2	70.93	99.19	97.68	95.56	98.57	**92.79**	95.01	92.74	94.65	94.99	95.03	**94.48**

The ablation study for the MFE-MIL method investigates the impact of key hyperparameters on model performance, focusing specifically on the weighting factor for joint training and the masking ratio. Within the joint training framework, a hyperparameter λ is introduced to balance the contrastive loss and the classification loss. Experimental results indicate that setting $\lambda = 0.3$ achieves the best average performance across both the CAMELYON16 and PANDA datasets, suggesting that placing a slight emphasis on the slide-level classification objective is beneficial.

Furthermore, the effect of the masking ratio, which controls the proportion of feature embeddings masked during training, was evaluated. The study found that an optimal masking ratio of 0.75 leads to the highest performance, implying that requiring the model to reconstruct a substantial portion of the input features promotes the learning of robust and global semantic representations, which are critical for WSIs analysis.

Finally, the use of two shared layers in the architecture was shown to be superior, both in terms of performance and computational efficiency, by enabling the model to learn generalizable features.

Chapter 5

Thesis Contributions and Future Work

5.1 Conclusion

In conclusion, both the PathTTT and MFE-MIL frameworks represent significant advancements in computational pathology, each addressing critical challenges in the field. PathTTT, by combining MAE with TTT and MAML, offers a robust solution to the pervasive problem of domain shift in patch level image analysis. Its superior performance in cross-dataset scenarios demonstrates a remarkable ability to generalize to unseen data, a crucial step toward the clinical deployment of reliable diagnostic models. The framework's ability to adapt to new domains makes it a highly effective approach for mitigating the variability introduced by different scanners, staining protocols, and patient populations.

The MFE-MIL framework, on the other hand, provides an effective method for enhancing the performance of various MIL models. By introducing an ensemble-based approach that combines features from different sources, MFE-MIL consistently improves a wide range of MIL baselines—from simpler pooling strategies to more complex attention-based networks. The framework's generalizability is a key strength, as it has been shown to be effective across different MIL architectures and on diverse cancer datasets. Its most notable impact is on models with lower baseline performance, where it provides a substantial boost, but it also offers a complementary performance gain for state-of-the-art models.

5.2 Future Work

5.2.1 PathTTT

Building upon the promising results of PathTTT in addressing distribution shifts at the patch level through a meta-learning framework combined with self-supervised learning, several avenues for future research remain open:

- **Extension to Multi-scale and WSI Levels:** While PathTTT has demonstrated robust adaptation at the patch level, extending the framework to operate effectively on multi-scale representations or entire WSIs could further improve clinical applicability. Integrating spatial context and hierarchical features may enhance the model's ability to handle complex domain shifts present in large-scale histopathological data.

- **Exploration of Alternative Self-Supervised Objectives:** The current method leverages MAE as the self-supervised component. Investigating other self-supervised learning paradigms—such as contrastive learning or predictive coding—could provide additional robustness and adaptability to unseen distributions.

- **Incorporation of Uncertainty Estimation:** Incorporating uncertainty quantification within the meta-test-time adaptation process may improve decision reliability, especially under significant domain shifts or limited data availability. This could facilitate more informed clinical decisions and model trustworthiness.

5.2.2 MFE-MIL

Future work will focus on several important directions to further advance the capabilities and applicability of the MFE-MIL framework:

- **Development of Advanced Masking Strategies:** We plan to investigate more sophisticated masking approaches that go beyond simple random masking. Potential strategies include attention-guided masking, which leverages the model's internal attention mechanisms to selectively mask less informative features, and spatially-aware masking that incorporates tissue

context or morphological structures. These approaches aim to improve the model's ability to learn more discriminative and contextually relevant feature representations, ultimately enhancing downstream classification performance.

- **Investigation of Domain Shift Robustness:** Recognizing the prevalence of domain shifts in histopathological data arising from variations in staining protocols, scanners, and patient populations, we aim to evaluate and enhance the robustness of MFE-MIL under domain shift conditions. This will involve incorporating domain adaptation or domain generalization techniques to ensure consistent performance across diverse datasets and clinical settings.

- **Enhancement of Model Interpretability:** To increase the clinical utility of the model, we will explore methods to interpret the reconstructed feature embeddings generated by the model. By analyzing these embeddings, we aim to uncover biologically meaningful patterns and relationships that contribute to diagnostic decisions. This could facilitate better understanding of the model's decision-making process and provide insights into underlying disease mechanisms, thereby bridging the gap between computational analysis and clinical pathology.

Bibliography

[1] M. S. Hosseini, B. E. Bejnordi, V. Q.-H. Trinh, L. Chan, D. Hasan, X. Li, S. Yang, T. Kim, H. Zhang, T. Wu, K. Chinniah, S. Maghsoudlou, R. Zhang, J. Zhu, S. Khaki, A. Buin, F. Chaji, A. Salehi, B. N. Nguyen, D. Samaras, and K. N. Plataniotis, "Computational pathology: A survey review and the way forward," *Journal of Pathology Informatics*, p. 100357, 1 2024.

[2] E. Surveillance and E. R. S. Program, "Cancer stat facts." https://seer.cancer.gov/statfacts/, 2024. Accessed: 2024-08-14.

[3] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar, "Bach: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 8 2019.

[4] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti, *et al.*, "Bracs: A dataset for breast carcinoma subtyping in h&e histology images," *Database*, vol. 2022, p. baac093, 2022.

[5] J. N. Kather, N. Halama, and A. Marx, "100,000 histological images of human colorectal cancer and healthy tissue," 4 2018.

[6] C. Zhu, W. Chen, T. Peng, Y. Wang, and M. Jin, "Hard sample aware noise robust learning for histopathology image classification," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 881–894, 2022.

[7] D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen, Q. Duan, J. Zhao, K. Li, Y. Qiao, and S. Zhang, "A real-world dataset and benchmark for foundation model adaptation in medical image classification," *Scientific Data*, vol. 10, 9 2023.

[8] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[9] A. Maleki, M. Raahemi, and H. Nasiri, "Breast cancer diagnosis from histopathology images using deep neural network and xgboost," *Biomedical Signal Processing and Control*, vol. 86, p. 105152, 2023.

[10] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I. Zörnig, D. Jäger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLOS Medicine*, vol. 16, p. e1002730, 1 2019.

[11] P. Gomathi, C. Muniraj, and P. Periasamy, "Digital infrared thermal imaging system based breast cancer diagnosis using 4d u-net segmentation," *Biomedical Signal Processing and Control*, vol. 85, p. 104792, 2023.

[12] F. Prezja, L. Annala, S. Kiiskinen, S. Lahtinen, T. Ojala, P. Ruusuvuori, and T. Kuopio, "Improving performance in colorectal cancer histology decomposition using deep and ensemble machine learning," *arXiv preprint arXiv:2310.16954*, 2023.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[14] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*, pp. 9229–9248, PMLR, 2020.

[15] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017.

[16] R. Maurya, N. N. Pandey, M. K. Dutta, and M. Karnati, "Fccs-net: Breast cancer classification using multi-level fully convolutional-channel and spatial attention-based transfer learning approach," *Biomedical Signal Processing and Control*, vol. 94, p. 106258, 8 2024.

[17] A. Kumar, A. Vishwakarma, and V. Bajaj, "Crccn-net: Automated framework for classification of colorectal tissue using histopathological images," *Biomedical Signal Processing and Control*, vol. 79, p. 104172, 2023.

[18] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, "Towards stable test-time adaptation in dynamic wild world," *arXiv preprint arXiv:2302.12400*, 2023.

[19] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International conference on machine learning*, pp. 16888–16905, PMLR, 2022.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[22] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[23] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[24] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[25] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147, 2021.

[26] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, p. 863–874, 2024.

[27] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban, *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, 2024.

[28] Y. Gandelsman, Y. Sun, X. Chen, and A. Efros, "Test-time training with masked autoencoders," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29374–29385, 2022.

[29] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira, "Benchmarking self-supervised learning on diverse pathology datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3344–3354, 2023.

[30] B. Ehteshami Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, and the CAMELYON16 Consortium, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[31] W. Bulten, K. Kartasalo, P. H. C. Chen, and et al., "Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge," *Nature Medicine*, vol. 28, pp. 154–162, 2022.

[32] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.