

M3amba: Memory Mamba is All You Need for Whole Slide Image Classification

Tingting Zheng^{1,2} Kui Jiang¹ Yi Xiao³ Sicheng Zhao⁴ Hongxun Yao^{1,2*}

¹Harbin Institute of Technology

²China Mobile 5G Application Innovation Joint Research, Harbin Institute of Technology

³Wuhan University ⁴Tsinghua University

{23b903051@stu.hit, jiangkui@hit, xiao.yi@whu, schzhao@tsinghua, h.yao@hit}.edu.cn

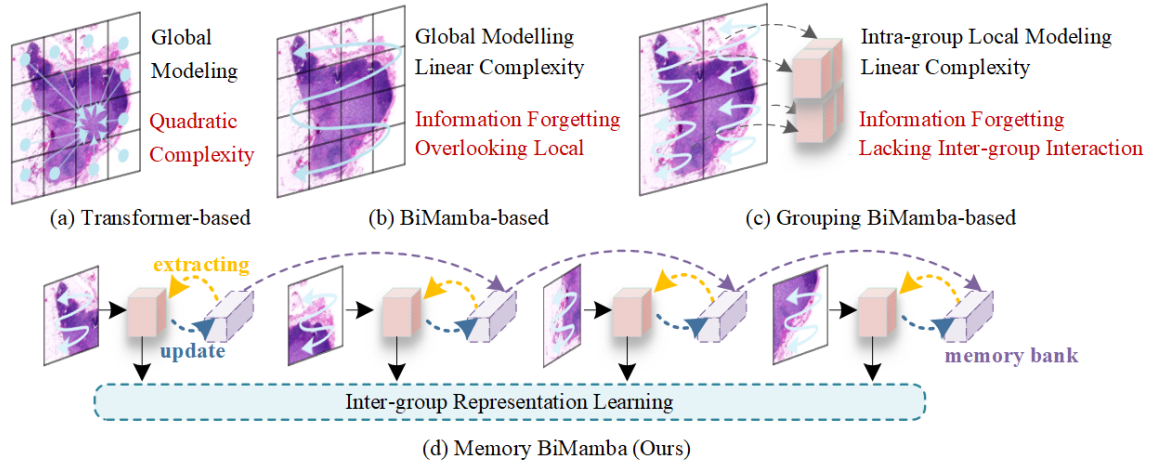


Figure 1. Comparison of different methods. (a)-(c) Transformers, bidirectional Mamba (BiMamba), and grouping BiMamba suffer from quadratic complexity, overlooking local or intra-group interactions, and forgetting information, respectively. (d) We propose a memory bank with grouping BiMamba to summarize and incorporate past group features into the current group for better bag prediction.

Abstract

Multi-instance learning (MIL) has demonstrated impressive performance in whole slide image (WSI) analysis. However, existing approaches struggle with undesirable results and unbearable computational overhead due to the quadratic complexity of Transformers. Recently, Mamba has offered a feasible solution for modeling long-range dependencies with linear complexity. However, vanilla Mamba inherently suffers from contextual forgetting issues, making it ill-suited for capturing global dependencies across instances in large-scale WSIs. To address this, we propose a memory-driven Mamba network, dubbed M3amba, to fully explore the global latent relations among instances. Specifically, M3amba retains and iteratively updates historical information with a dynamic memory bank (DMB), thus overcoming the catastrophic forgetting defects of Mamba for long-term context representation. For better feature representation, M3amba involves an intra-group bidirectional Mamba (BiMamba) block to refine local interactions within groups. Meanwhile, we additionally perform cross-attention fusion to incorporate relevant historical informa-

tion across groups, facilitating richer inter-group connections. The joint learning of inter- and intra-group representations with memory merits enables M3amba with a more powerful capability for achieving accurate and comprehensive WSI representation. Extensive experiments on four datasets demonstrate that M3amba outperforms the state-of-the-art by 6.2% and 7.0% in accuracy on the TCGA BRCA and TCGA Lung datasets while maintaining low computational costs.

1. Introduction

Histopathology slides with rich biological structures are crucial for accurate cancer diagnosis [2]. The advancement of digital technology opens new avenues for traditional slide pathology by scanning slides into whole slide images (WSIs). With the remarkable development of deep learning, pathologists can now recruit deep models to explore texture information from WSIs, outperforming traditional hand-crafted feature-based approaches [3, 22, 28, 51]. However, due to the gigapixel resolution and complex mi-

crostructures of WSI, collecting pixel-level annotations is a laborious task [37, 40, 50]. To address this, multi-instance learning (MIL) has been proposed as a promising solution. MIL crops patches (instances) from WSI and associates them using bag-level labels, resulting in semi-supervised learning for bag label prediction.

Despite these advantages, MIL does not perform as well as expected when modeling long-distance dependency among instances, as convolutional neural networks (CNNs) inherently emphasize local connectivity. Although some non-local attention [30, 47] and Transformer [36] methods could enhance the capability for global in-context communications, they often face limited performance and suffer from quadratic computational costs due to the self-attention mechanism, as illustrated in Figure 1 (a). Recent efforts have attempted to reduce the long-range sequence N by marking random ratios of input instances [17, 48]. However, dropping instances inevitably lose information and damage the intrinsic spatial patterns of the tumor regions, which is crucial for global bag representation. Another approach, the MMIL-Transformer [48], decreases the original complexity from $O(N^2)$ to $O(\frac{N^2}{g})$ by dividing N into g groups. However, it inevitably undermines the distinctiveness within each group, and the inherent quadratic complexity in Transformers remains unresolved.

The recent-popular Mamba [13, 32] provides a promising solution to overcome this issue by modeling long sequences with linear complexity. However, the vanilla Mamba lacks explicit consideration of the characteristics in WSI analysis, including memory decay and suboptimal local representation. More precisely, **there exist sparse tumors and longer instances within a large-scale WSI**, where the long-range correlation among instances is inevitably forgotten by sequence modeling. This long-context deficiency issue of Mamba is inherently attributed to its RNN-like architecture. As evidenced in previous works [42, 45], this architecture exhibits limitations in preserving crucial information from earlier input sequences when the context length increases, due to its fixed-size memory constraints [38]. In addition, **WSIs exhibit obvious spatial hierarchies and intricate local patterns**, which are barely explored by the Mamba elaborated for unidirectional 1-D continuous data, leading to undesired suboptimal local representation (Figure 1 (b)). While some group-based Mamba approaches promote local modeling in vision tasks [29], relying only on cross-channel communication results in limited information exchange among groups (Figure 1 (c)). These aspects have been completely overlooked by previous Mamba-based methods, which intuitively complicate the direct application of Mamba to WSI classification tasks.

To this end, this study marries group sampling and a memory bank with Mamba, developing an intra- and inter-group memory-driven Mamba (M3amba) fusion and rep-

resentation learning framework for efficient and effective WSI tasks. Instead of exploring global relations across all instances with self-attention, the M3amba architecture decomposes the bag-level classification into two sub-tasks: (1) memory Mamba-driven group representation learning and (2) inter-group bag representation learning. The former employs a memory Mamba block that utilizes a bidirectional Mamba and attention block (BiMamba Block) to fully capture instance correlations within the intra-group. In particular, considering that information forgetting and limited intra-group interactions lead to feature sparsity, we present a dynamic memory bank (DMB) to store and refine historical group representations. To incorporate more informative cues into the current group representation, we employ cross-attention among the DMB to facilitate memory updates. This helps retain critical information and promotes information exchange during long-sequence modeling. Moreover, to achieve comprehensive inter-group representation learning, the latter sub-task first consolidates all group representations using the BiMamba Block, and extracts group-specific local information via a convolutional block (ConvBlock). Different from previous work that employs *static* concatenation or summation for bag prediction, we harmonize the merits of global and local group representations using *dynamic* weights from a Mixture of Experts, leading to more accurate bag predictions. Experimental results demonstrate that our M3amba significantly outperforms state-of-the-art methods by 6.2%, 7.0%, and 3.8% in accuracy on the TCGA BRCA, TCGA Lung, and TCGA ESCA datasets, respectively, and by 12.8% and 6.5% in terms of AUC on the BRACS dataset.

Our main contributions can be summarized as follows:

- We propose the M3amba framework, which marries the power of Mamba with a memory bank to address the challenges of modeling long-range instance dependencies. M3amba mitigates quadratic complexity and information forgetting, achieving effective WSI classification.
- We develop the BiMamba Block to capture intra-group correlations and introduce a Dynamic Memory Bank (DMB) to store and refine long-range features, effectively alleviating catastrophic forgetting defects.
- To grasp overlooked inter-group correlations, M3amba employs cross-attention fusion to integrate relevant historical cues across groups within the DMB. This approach bridges local and global dependencies, ensuring a comprehensive feature representation for complex WSIs.

2. Related Work

2.1. Multi-instance Learning in WSI Classification

The multi-instance learning (MIL) [19] aims to predict WSI labels by learning local morphological and global spatial features. Early works employ CNNs to capture local tis-

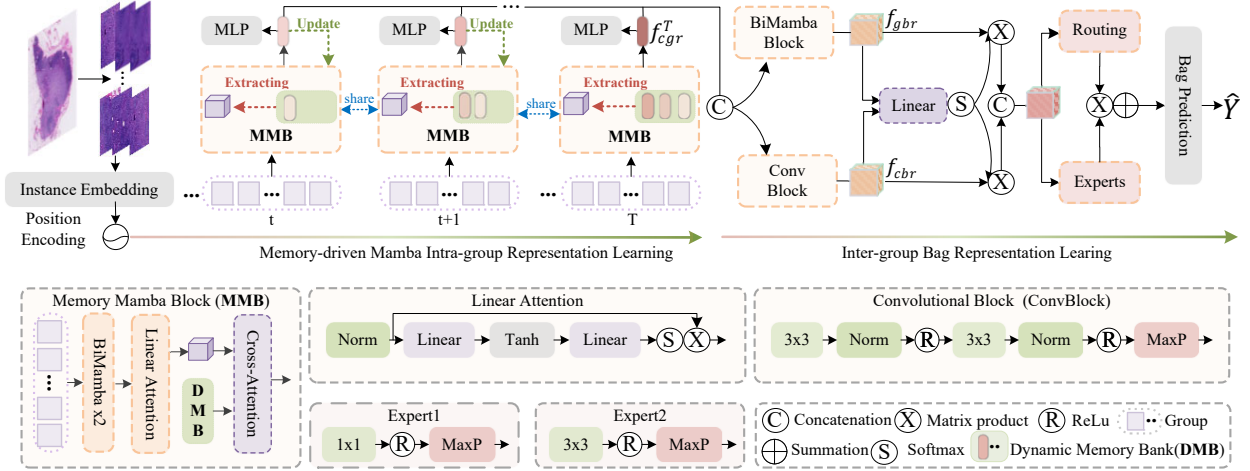


Figure 2. The architecture of our proposed memory-driven memory Mamba network (M3amba) for WSI classification. M3amba randomly samples group instance features at time t . And then, the memory Mamba block introduces a bidirectional Mamba (BiMamba) and a linear attention block (BiMamba Block) to explore relations among instances. To alleviate key information forgetting and enhance discriminative features within the group, we employ cross-attention to refine and fuse historical information from a dynamic memory bank (DMB) into the current group representation f_{cgr}^t , and then predict the group class probability using a multi-layer perceptron (MLP). Meanwhile, f_{cgr}^t helps the DMB update for $t+1$ group information exchange. Subsequently, a BiMamba Block aggregates global inter-group relations, and a convolutional block (ConvBlock) captures group-specific local information within each group. Finally, the Mixture of Experts harmonizes the merits of global and local group representations for more robust and stable bag prediction $\hat{Y}$.

sue hierarchical information equipped with MaxPooling or Expectation-Maximization strategies to identify tumor instances [16, 19]. Subsequently, the RNNMIL [5] introduces a Recurrent Neural Network (RNN) to model the correlations between Top-K instances from the MaxPooling classifier to improve performance. However, these approaches often struggle to produce reliable discriminative features due to limited exploration of relations between instances. To address these issues, some researchers have turned to attention mechanisms, GNNs, and Transformers to capture complex global and local tumor tissue features [3, 18, 31]. In particular, Transformer-based methods show excellent superiority due to their global in-context correspondence [30, 41]. However, the self-attention mechanism demands quadratic complexity, leading to a substantial computational burden for high-resolution WSIs in practice [48]. Therefore, we propose a memory-driven Mamba network to model long-range dependencies among instances, facilitating better performance with computational efficiency.

2.2. State Space Model

State Space Models (SSMs) [32] show significant effectiveness in modeling continuous long-range dependencies, but are impractical in information-intensive data due to excessive computational and memory demands. To address these challenges, the Mamba [13] introduces selective scanning mechanisms and hardware-aware algorithms, achieving better performance with linear complexity. Subsequently, Mamba-based variants, ranging from various scan-

ning strategies to combinations of CNNs, GNNs, and Transformers with Mamba, have been developed in the visual domain [26, 39, 52] and WSI analysis tasks [11, 43]. Unlike approaches that directly wrap all instances into one or multiple models, which incur extreme learning and computation burdens [11, 12], we propose memory-driven Mamba to enhance intra-group instance fusion and inter-group representation learning. Additionally, we design an informative bank and a cross-attention fusion scheme to facilitate representation and global modeling

2.3. Mixture of Experts

The Mixture of Experts (MoE) [20] typically decomposes tasks into multiple parts and uses a routing network to integrate the outputs from part-specific experts into a final result. Due to powerful adaptive capability and efficient dynamic structure, MoE techniques have been widely applied to multi-task vision [9] and multi-modal fusion [6]. In particular, recent research is growing interest in employing MoE approaches to facilitate diverse WSI representations learning and fusion [14]. For example, CNN-based multiple experts exploit spatial structure features between different magnification instances for better cancer sub-type region segmentation [34]. In addition, existing efforts devise class-specific experts leading to more accurate multi-class lymphoma WSI classification [14]. Recognizing the merits of MoE, we pioneer the coupling of MoE with Mamba for robust WSI label inference.

*Corresponding authors

3. Method

This section starts with the formulation of MIL and SSMs and an overview of our proposed M3amba network. It then details the memory-driven Mamba group representation learning and inter-group bag representation learning modules and their essential components.

3.1. Preliminaries

MIL Formulation. Given a single WSI X , the corresponding label Y , we partition the non-background tissue regions of the “bag” X into B instances $\{(x_b \in \mathbb{R}^{W \times H \times 3}, y_b) | 1 \leq b \leq B\}$, where H , W and y_b denote the height, width and class label of the instance x_b , respectively.

The classic MIL-based approach embeds an instance x_b into a $1 \times D$ -dimensional feature vector v_b by pre-trained encoders [25, 47, 51]. By exploring the global and local correlations among instance features $\{f_b\}_{b=1}^B$ to form bag representation f , a multi-layer perceptron (MLP) classifier is employed to predict the bag label. Taking binary classification as an intuitive example in Eq.(1), if an instance x_b is labeled 1, the bag prediction is “positive”. Otherwise, it will be labeled 0 (negative).

$$Y = \begin{cases} 1, & \text{if } \exists x_b \in X : y_b = 1, \\ 0, & \text{if } \forall x_b \in X : y_b = 0. \end{cases} \quad (1)$$

State Space Models. The State Space Models (SSMs) map input sequences $x(t)$ at time t into output sequences $y(t)$ through latent states $h(t)$ from a deep sequence model. This process typically uses linear ordinary differential equations or their discrete difference equations to describe the continuous or discrete dynamics.

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad (2)$$

$$h(t) = \bar{\mathbf{A}}h(t-1) + \bar{\mathbf{B}}x(t), \quad (3)$$

$$y(t) = \mathbf{C}h(t) + \mathbf{D}x(t), \quad (4)$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are continuous parameters learned by gradient descent, and $\mathbf{D}$ can be viewed as a skip connection. $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are the $\mathbf{A}$ and $\mathbf{B}$ to discrete parameters. The discrete-based Mamba introduces a selectively propagating and forgetting algorithm, which facilitates long-term dependency and global modeling with linear complexity. To enhance comprehensive instance communication and computational efficiency while addressing key information forgetting, we couple the characteristics of the MIL-based WSI tasks with the merits of Mamba and memory bank to propose a memory Mamba network.

3.2. Overview for M3amba

Figure 2 outlines our proposed M3amba network. Our core goal is to effectively and efficiently explore massive instance relationships by introducing intra- and inter-group

BiMamba with a memory bank for better bag label inference.

To promote the exploration of spatial and structural relations between instances, we introduce a learnable positional encoding f_p for all instance features $\{f_b\}_{b=1}^B$. Unlike existing approaches [43, 48], which aggregate all instances via the Mamba or self-attention mechanism, leading to undesirable local correspondence and computational complexity, M3amba randomly samples the larger bag $\{f_b\}_{b=1}^B$ into T smaller groups $\{f_{ins}^t = \{f_m^t\}_{m=1}^M | t \in [1 \leq T = \frac{B}{M}]\}$, where M is the number of instances in a group. We design a memory Mamba block to introduce a bidirectional Mamba and a linear attention block (BiMamba Block) to capture semantics and spatial features into the group representation f_{gr}^t . To alleviate key information forgetting while enhancing discriminative information within a group, we equip with a dynamic memory bank (DMB) and employ cross-attention to extract label-dependent information from the DMB into f_{cgr}^t and predict the group category probability $\hat{y}_{cgr}^t$. To further extract label-related components, we introduce the bag label Y to supervise the group prediction. In particular, $\hat{y}_{cgr}^t$ and f_{cgr}^t assist in updating and refining the DMB to alleviate discriminative information memory decay while promoting group information exchange. To facilitate bag representation, a BiMamba block distills a global bag representation across groups, while the convolutional block (ConvBlock) preserves the group-specific local feature within each group to form a bag representation. Subsequently, the Mixture of Experts harmonizes global and local bag representations, resulting in more accurate bag predictions $\hat{Y}$.

3.3. Memory-driven Mamba Group Representation Learning

For effective and efficient modeling of instance relations, we propose the memory Mamba block (MMB) that couples Mamba with a dynamically updated and refined memory bank. MMB addresses information forgetting and limited group interactions. Taking the f_{ins}^t operation as an example, the BiMamba Block fully explores complex relations within group instances into group representation f_{gr}^t . To alleviate the sparsity of disease features in the current group and information forgetting, f_{gr}^t is equipped with a cross-attention to integrate crucial past knowledge $\mathbf{M}$ from a DMB and augmented representations $\{\bar{\mathbf{f}}_a^k\}_{k=2}^K$ (More details in following section) into f_{cgr}^t . The f_{cgr}^t is passed to the MLP to infer current group category probability $\hat{y}_{cgr}^t \in [0, 1]$. The procedures are expressed as

$$f_{gr}^t = \text{BiMamba Block}(f_{ins}^t), \quad (5)$$

$$f_{cgr}^t = \text{Cross-Attention}\left(f_{gr}^t, \text{Cat}\left(\mathbf{M}, \{\bar{\mathbf{f}}_a^k\}_{k=2}^K\right)\right), \quad (6)$$

$$\hat{y}_{cgr}^t = \text{MLP}(f_{cgr}^t). \quad (7)$$

Dynamic Memory Bank. The DMB aims to tackle information forgetting associated with longer instances and enhance group representations by extracting and updating the K discriminative features $\mathbf{M} = \{(\{f_{cgr}^k\}_{k=1}^K, \{\hat{y}_{cgr}^k\}_{k=1}^K)\}$, where $\{f_{cgr}^k\}_{k=1}^K$ features are ranked from highest to lowest according to the label confidence $\{\hat{y}_{cgr}^k\}_{k=1}^K$. To preserve the most salient features promptly, at time $t - 1$, the DMB takes $\hat{y}_{cgr}^{t-1}$ and f_{cgr}^{t-1} as inputs to rearrange $\mathbf{M}$ by comparing $\hat{y}_{cgr}^{t-1}$ with the stored $\{\hat{y}_{cgr}^k\}_{k=1}^K$. Since the information that facilitates label prediction may be dispersed across groups, we design feature *filtering*, *quantization*, and *aggregation* to extract class-related information from $\{f_{cgr}^k\}_{k=2}^K$ into $f_{cgr}^{k=1}$. Specifically, we introduce $\hat{y}_{cgr}^k$ as the category-aware score to distill the category-aware representation in f_{cgr}^k using the dot product while producing a category-aware attention map $\mathbf{A}_{cgr}^k$ via Softmax. Subsequently, we generate weights by quantifying the category-aware attention maps to aggregate dispersed discriminative information among $\{f_{cgr}^k\}_{k=2}^K$. The operations can be expressed as

$$\bar{f}_{cgr}^k = \hat{y}_{cgr}^k \times f_{cgr}^k, k = [1, 2, \dots, K], \quad (8)$$

$$\mathbf{A}^k = \text{Softmax}(\bar{f}_{cgr}^k), k = [1, 2, \dots, K], \quad (9)$$

$$\bar{\mathbf{f}}_{\mathbf{a}}^k = f_{cgr}^{k=1} + \frac{\mathbf{A}^k}{\mathbf{A}_{k=1}} \times f_{cgr}^k, k = [2, 3, \dots, K]. \quad (10)$$

We concatenate $\{\bar{\mathbf{f}}_{\mathbf{a}}^k\}_{k=2}^K$ and $\{f_{cgr}^k\}_{k=1}^K$ to provides more label-related features for the current group f_{cgr}^t by Eq.(6).

3.4. Inter-group Bag Representation Learning

To fully integrate rich group representations, we design two parallel sub-branches: BiMamba Block and ConvBlock. The former grasp the global relations across groups. To avoid homogenization across group representations, the latter employs a ConvBlock to retain and integrate the salience feature within each group. The procedures are encapsulated as

$$f_{gbr} = \text{BiMamba Block}(\{f_{cgr}^t\}_{t=1}^T), \quad (11)$$

$$f_{cbr} = \text{ConvBlock}(\{f_{cgr}^t\}_{t=1}^T). \quad (12)$$

To promote comprehensive and accurate bag representation, instead of directly summing or concatenating f_{gbr} and f_{cbr} , we generate the attention scores using a linear layer followed by a Softmax function, and then refine and concatenate f_{gbr} and f_{cbr} . The procedures are expressed as:

$$S_{gbr}, S_{cbr} = \text{Softmax}(\text{Linear}(f_{gbr}, f_{cbr})), \quad (13)$$

$$f_{gcb} = \text{Cat}(S_{gbr} \times f_{gbr}, S_{cbr} \times f_{cbr}). \quad (14)$$

Bag label Prediction. To improve label prediction robustness and applicability, we are inspired by a Mixture of Experts, introducing two different convolutional modules and

using attention scores to integrate their outputs into an MLP for the final bag prediction. The procedures are depicted as

$$f_{\text{expert1}} = \text{MaxPooling}(\text{Conv}_{1 \times 1}(f_{gcb})), \quad (15)$$

$$f_{\text{expert2}} = \text{MaxPooling}(\text{Conv}_{3 \times 3}(f_{gcb})), \quad (16)$$

$$w_1, w_2 = \text{Softmax}(\text{Linear}(f_{gcb})), \quad (17)$$

$$\hat{Y} = \text{MLP}(w_1 \times f_{\text{expert1}} + w_2 \times f_{\text{expert2}}). \quad (18)$$

3.5. Model Optimization

To achieve efficient intra- and inter-group instances modeling and high-quality feature representation for better bag label inference, we explore practical composite cross-entropy losses, involving group-label probability $\hat{y}_{cgr}^t$, bag-label probabilities (S_{gbr} , S_{cbr} , and $\hat{Y}$). These loss functions are formulated as follows

$$\begin{aligned} \mathcal{L}(\mathcal{J}) &= -[Y \log \mathcal{J} + (1 - Y) \log(1 - \mathcal{J})], \\ \mathcal{J} &\in \{S_{gbr}, S_{cbr}, \hat{Y}\}, \end{aligned} \quad (19)$$

$$\mathcal{L}(\hat{y}_{cgr}) = -\frac{1}{T} \sum_{t=1}^T [Y \log \hat{y}_{cgr}^t + (1 - Y) \log(1 - \hat{y}_{cgr}^t)]. \quad (20)$$

The aforementioned loss function is given by

$$\mathcal{L} = \mathcal{L}(S_{gbr}) + \mathcal{L}(S_{cbr}) + \mathcal{L}(\hat{Y}) + \lambda \mathcal{L}(\hat{y}_{cgr}), \quad (21)$$

where λ is adjusted from 0 to 0.2 via integrating epoch and cosine curve to guide the model from coarse to label-related information exploration [24].

4. Experiments

4.1. Datasets

To comprehensively evaluate our M3amba, we conduct extensive experiments on four public WSI datasets: TCGA Breast Cancer (BRCA) [35], TCGA Lung Cancer [35], TCGA Esophageal Cancer (ESCA) [35], and Breast Adenocarcinoma Subtypes (BRACS) [4].

TCGA Breast Cancer Dataset. We collect 952 breast slides from the TCGA BRCA project, including 749 for Invasive Ductal Carcinoma (IDC) and 203 for Invasive Lobular Carcinoma (ILC). Following [23], the TCGA BRCA dataset is randomly split into training, validation, and testing sets with ratios of 65:10:25, and CLAM [25] is used to crop each WSI into 256×256 non-overlapping instances at $10 \times$ magnification, resulting in approximately 3.1 million instances.

TCGA Lung Cancer Dataset. The TCGA Lung cancer dataset [35] includes two cancer subtypes: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). It comprises 541 LUAD slides from 478 cases and

Table 1. Quantitative comparison of our results and MIL methods on TCGA BRCA and TCGA Lung datasets using Resnet18-ImageNet feature extractor. The numbers in **black** and underline indicate the best and second performance. Following the DTFD method, the FLOPs are measured with the 120 instances in a bag and an instance feature dimension $D = 1024$. We set $K = 1$ and $M = 120$, respectively.

Methods	TCGA BRCA				TCGA Lung				FLOPs	Parameters
	Accuracy	AUC	Precision	Recall	Accuracy	AUC	Precision	Recall		
ABMIL [18]	0.862±0.025	0.882±0.038	0.888±0.019	0.945±0.020	0.844±0.023	0.919±0.026	0.845±0.028	0.853±0.016	78.1M	655.3K
DSMIL [3]	0.823±0.021	0.820±0.033	0.861±0.029	0.928±0.041	0.783±0.041	0.856±0.040	0.797±0.067	0.783±0.022	117.6M	855.7K
CLAM-MB [25]	0.865±0.020	0.890±0.029	0.890±0.016	0.947±0.031	0.844±0.023	0.919±0.026	0.845±0.028	0.853±0.016	94.8M	791.1K
CLAM-SB [25]	0.858±0.011	0.877±0.029	0.877±0.018	0.955±0.014	0.834±0.030	0.912±0.034	0.840±0.054	0.840±0.047	94.8M	790.7K
TransMIL [30]	0.847±0.021	0.846±0.036	0.883±0.019	0.929±0.024	0.819±0.038	0.885±0.030	0.830±0.053	0.819±0.033	64.4M	2670.6K
DTFD-MaxMin [47]	0.816±0.023	0.810±0.033	0.874±0.028	0.897±0.014	0.832±0.031	0.907±0.031	0.850±0.047	0.821±0.057	80.1M	986.7K
DTFD-AFS [47]	0.823±0.028	0.824±0.034	0.863±0.027	0.924±0.025	0.852±0.020	0.917±0.022	0.861±0.042	0.854±0.058	79.4M	986.7K
DTFD-MaxS [47]	0.828±0.038	0.826±0.049	0.885±0.018	0.897±0.048	0.764±0.010	0.837±0.017	0.790±0.029	0.738±0.044	79.4M	986.7K
MHIM-ABMIL [33]	0.858±0.004	0.883±0.020	0.891±0.021	0.935±0.020	0.867±0.031	0.931±0.026	0.868±0.044	0.878±0.024	141.4M	1182.0K
MHIM-TransMIL [33]	0.848±0.022	0.872±0.013	0.889±0.032	0.924±0.014	0.832±0.035	0.911±0.026	0.852±0.035	0.817±0.081	1228.6M	5342.2K
IBMIL-ABMIL [21]	0.859±0.018	0.897±0.028	0.890±0.017	0.937±0.025	0.816±0.027	0.901±0.024	0.822±0.051	0.825±0.052	157.1M	1444.4K
SSMMIL [12]	0.863±0.006	0.896±0.032	0.886±0.014	0.949±0.023	0.843±0.033	0.919±0.022	0.845±0.028	0.850±0.047	126.1M	1052.16K
MambaMIL [43]	0.868±0.017	0.878±0.032	0.903±0.031	0.933±0.019	0.856±0.027	0.910±0.024	0.843±0.046	0.889±0.033	71.5M	594.69K
ACMIL [49]	0.869±0.017	0.900±0.019	0.889±0.030	0.955±0.023	0.844±0.023	0.919±0.026	0.845±0.028	0.853±0.016	78.7M	662.4K
M3amba (Ours)	0.931±0.024	0.971±0.005	0.956±0.035	0.957±0.022	0.937±0.027	0.975±0.003	0.938±0.028	0.947±0.016	72.4M	1242.1K

512 LUSC slides from 478 distinct cases. Following the settings in previous work [21, 33, 47], the TCGA Lung dataset is randomly split into training, validation, and testing sets with ratios of 65:10:25, and CLAM [25] is used to crop each WSI into 256×256 non-overlapping instances at $20 \times$ magnification, resulting in approximately 4.1 million instances.

TCGA Esophageal Cancer Dataset. The TCGA ESCA dataset consists of 156 diagnostic WSIs, including 90 squamous cell carcinoma and 66 adenocarcinoma cases [35]. As in MuRCL [53], the TCGA ESCA dataset is randomly split into training, validation, and testing sets with ratios of 3:1:1, and CLAM [25] is employed to crop non-overlapping 256×256 instances at $20 \times$ magnification, resulting in approximately 0.5 million instances.

Breast Carcinoma Subtyping Dataset. The BRACS dataset involves three categories: 265 benign, 89 atypical, and 193 malignant breast tumors [4]. Similar to [49], we also adhere to the official dataset split [4, 49], with 537 WSIs available: 395 for training, 65 for validation, and 87 for testing. CLAM [25] is used to extract non-overlapping 256×256 instances at $10 \times$ magnification, resulting in approximately 1.4 million instances.

4.2. Implementation Details

Similar to [21, 23, 53], for the TCGA BRCA, Lung and ESCA datasets, We use a ResNet18 [15] encoder pre-trained on ImageNet [10] (Resnet18-ImageNet) to obtain 512-dimensional feature vectors from each instance. For the BRACS dataset, we directly used the public features from [49], pre-trained with a 512-dimensional on Resnet18-ImageNet and a 384-dimensional ViT-S/16 model using self-supervised learning through DINO [7] (ViT-S/16-SSL). The AdaMax optimizer [1] with a weight decay of $1e - 5$ and an initial learning rate of $1e - 4$ was used. With the above settings, we trained M3amba for 100 epochs with a batch size of 1 on a single NVIDIA RTX 3090Ti GPU.

4.3. Baseline and Evaluation Metrics

Baseline. We compare our proposed M3amba with representative MIL methods, including attention-based approaches (ABMIL [18], CLAM [25], DSMIL [3], MHIM-ABMIL [33], IBMIL-ABMIL [21], ACMIL [49]), transform-based approaches (TransMIL [30], DTFD [47], MHIM-TransMIL [33]), and Mamba-based approaches (SSMMIL [12] and MambaMIL [43]). The results for all other methods are reproduced using the official code provided under the same settings.

Evaluation Metrics. The widely used performance metrics in the literature [3, 21, 51], such as area under the receiver operating characteristic curve (AUC), accuracy, F1 score (F1), precision, and recall, are employed for evaluation, with a threshold set at 0.5. Due to the imbalance between normal and positive samples, we follow established work [33, 43, 44] that uses 5-fold cross-validation in each main experiment to evaluate model performance more accurately and reliably. For each training, the training and validation data are split at the patient level according to the ratios for model selection. The mean classification performance and standard deviation across the five test sets are reported.

4.4. Comparison with State-of-the-arts

Quantitative Comparison. We compare our M3amba with representative methods on four widely used WSI classification datasets. Quantitative results are presented in Tables 1, 2, and 3. As expected, our M3amba achieves competitive performance across all metrics, outperforming the attention-based methods ACMIL [49] and MHIM-ABMIL [49] by 6.2%, 7.0%, and 3.8% in accuracy, respectively. Notably, compared to the transformer-based methods MHIM-TransMIL and TransMIL [30], our M3amba significantly reduces computational complexity and model parameters while maintaining remarkable classification performance. Benefiting the efficient long-range modeling,

Table 2. Quantitative comparison of our results and MIL results on the ESCA dataset using Resnet18-ImageNet feature extractor.

Methods	Accuracy	AUC	Precision	Recall
ABMIL [18]	0.827±0.092	0.914±0.066	0.817±0.088	0.911±0.097
DSMIL [3]	0.808±0.065	0.882±0.084	0.845±0.090	0.844±0.124
CLAM-MB [25]	0.821±0.078	0.902±0.088	0.845±0.064	0.844±0.102
CLAM-SB [25]	0.834±0.061	0.927±0.064	0.848±0.091	0.889±0.086
TransMIL [30]	0.796±0.101	0.895±0.083	0.807±0.109	0.867±0.097
DTFD-MaxMin [47]	0.834±0.110	0.881±0.145	0.806±0.113	0.967±0.027
DTFD-AFS [47]	0.872±0.054	0.911±0.046	0.872±0.028	0.911±0.083
DTFD-MaxS [47]	0.777±0.116	0.820±0.092	0.779±0.128	0.900±0.082
MHIM-ABMIL [33]	0.859±0.082	0.940±0.046	0.848±0.109	0.944±0.035
MHIM-TransMIL [33]	0.853±0.054	0.911±0.040	0.842±0.057	0.922±0.057
IBMIL-ABMIL [21]	0.859±0.115	0.915±0.086	0.861±0.095	0.900±0.124
SSMMIL [12]	0.809±0.092	0.910±0.069	0.826±0.098	0.856±0.083
MambaMIL [43]	0.821±0.098	0.908±0.074	0.869±0.092	0.822±0.129
ACMIL [49]	0.885±0.078	0.948±0.042	0.883±0.063	0.922±0.083
M3amba (Ours)	0.923±0.026	0.989±0.009	0.918±0.049	0.956±0.022

Table 3. Quantitative comparison of our results and MIL results from ACMIL [49] on the BRACS dataset.

Methods	Resnet18-ImageNet		ViT-S/16-SSL	
	F1	AUC	F1	AUC
MeanPooling	0.484±0.029	0.685±0.011	0.522±0.038	0.739±0.007
MaxPooling	0.489±0.047	0.738±0.014	0.596±0.029	0.823±0.033
ABMIL [18]	0.523±0.028	0.723±0.035	0.680±0.051	0.866±0.029
DSMIL [3]	0.511±0.052	0.751±0.028	0.577±0.028	0.816±0.028
CLAM [25]	0.521±0.046	0.750±0.039	0.631±0.034	0.863±0.005
TransMIL	0.444±0.040	0.732±0.043	0.631±0.003	0.841±0.006
DTFD [47]	0.469±0.016	0.717±0.032	0.612±0.080	0.870±0.022
MHIM [33]	0.511±0.022	0.774±0.021	0.625±0.060	0.865±0.017
IBMIL [21]	0.510±0.043	0.726±0.034	0.645±0.041	0.871±0.014
SSMMIL [12]	0.549±0.052	0.759±0.018	0.658±0.031	0.853±0.015
MambaMIL [43]	0.596±0.041	0.811±0.017	0.705±0.025	0.882±0.019
ACMIL [49]	0.552±0.048	0.754±0.008	0.722±0.030	0.888±0.010
M3amba (Ours)	0.800±0.046	0.939±0.015	0.834±0.044	0.953±0.012

the SSMMIL [12] and MambaMIL [43] methods show improvements on the TCGA Lung dataset, where more than 80% of the area belongs to tumor. However, they still lag behind our method by 9.7% and 5.6% in Recall, respectively. Moreover, for the BRACS dataset, our M3amba outperforms MambaMIL [43] by 20.4%, 11.2%, 12.8%, and 6.5% in terms of F1 and AUC, respectively. The main reason for these improvements is that most comparative methods simultaneously pack all instances into the model, which lacks explicit exploration of the critical relationships among global and local instances. Besides, they suffer from the limitation of long-context information forgetting, leading to ambiguous feature representations and suboptimal performance. In contrast, our M3amba employs BiMamba Block for inter-group instance fusion and intra-group representation learning and develops a novel dynamic memory bank (DMB) to retain long-range information, thus achieving efficient yet effective performance.

Stability Comparison. Figure 3 highlights the high accuracy and robustness of our proposed M3amba. Notably, the CLAM and MambaMIL methods excel on BRCA by leveraging relationships among instances via attention and Mamba. However, they underperform on the more challenging ESCA WSIs. By contrast, our approach demonstrates greater stability by introducing intra- and inter-group

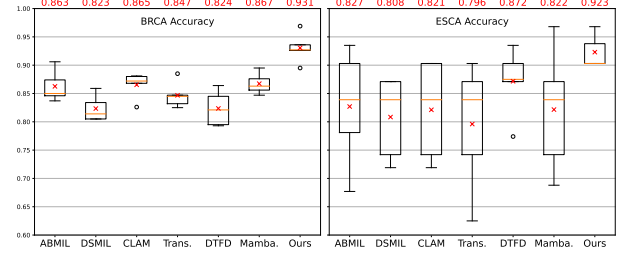


Figure 3. Comparative results regarding accuracy box plots on the TCGA BRCA and TCGA ESCA datasets.

Mamba, supported by a dynamically refined memory bank. This facilitates a comprehensive exploration of correlations among instances and global bag representations, leading to improved performance.

4.5. Ablation Studies

Validation on Basic Components. We conduct ablation studies to validate the contributions of individual components, including the bidirectional Mamba (BiMamba) block, convolutional block (ConvBlock), dynamic memory bank (DMB), bag label for group cross-entropy (GCE) loss, and Mixture of Experts (MoE) to the final classification performance. For simplicity, we denote our model as M3amba and create a baseline model (w Mamba) by removing all components except the bag cross-entropy loss $\mathcal{L}(S_{gbr})$. We then design three models (w Mamba, w BiMamba, and w ConvBlock) to investigate the effectiveness of intra- and inter-group fusion and representation learning. To verify that the DMB improves group representation, we create w GCE and w DMB models. Finally, we evaluate the robustness and adaptability of the MoE by comparing w DMB with M3amba.

Quantitative results are presented in Table 4 (a) and (b), revealing that the complete model M3amba significantly outperforms its incomplete versions. Specifically, the grouping strategy marries the power of BiMamba to facilitate intra-group instance fusion and inter-group representation learning, resulting in a 4.6% accuracy gain (comparing w Mamba and w BiMamba). In addition, the ConvBlock extracts group-specific local information within groups, leading to a 2.3% accuracy gain (comparing w BiMamba and w ConvBlock). Due to the information forgetting and the features sparsity in a group, the DMB enhances discriminative representation, achieving improvements of 11.5% in accuracy and 10.8% in recall (comparing w DMB and w ConvBlock). In particular, w DMB alleviates group representation and bag label bias, improving F1 and accuracy by 9.5% and 8.0%, respectively (comparing w GCE and w DMB). Finally, harmonizing the global and local bag representations results in a more comprehensive representation, increasing accuracy by 8.1%.

Ablations of DMB. To further validate our proposed

Table 4. Ablation studies are conducted on the BRACS dataset with ResNet18-ImageNet features. Apart from M3amba, the other results from $\hat{Y} = \frac{S_{cbr} + S_{gbr}}{2}$.

(a) Validation on basic components. K and M are set to 3 and 512, respectively.

Model	ConvBlock	GCE	DMB	MoE	Precision	Recall	F1	Accuracy
w Mamba	✗	✗	✗	✗	0.576	0.526	0.537	0.540
w BiMamba	✗	✗	✗	✗	0.579	0.568	0.572	0.586
w ConvBlock	✓	✗	✗	✗	0.608	0.581	0.582	0.609
w GCE	✓	✓	✗	✗	0.606	0.604	0.586	0.644
w DMB	✓	✓	✓	✗	0.725	0.689	0.681	0.724
M3amba	✓	✓	✓	✓	0.809	0.804	0.800	0.805

(b) Ablations of DMB. $^\#$ denote the use of random mixup [46] scheme to fuse features in Eq.(10). § denotes no augmented feature $\{\bar{\mathbf{f}}_a^k\}_{k=2}^K$ in Eq.(6).

Model	Precision	Recall	F1	Accuracy
w DMB §	0.686	0.612	0.559	0.667
w Mixup $^\#$	0.700	0.654	0.644	0.690
w DMB	0.725	0.689	0.681	0.724

(c) Effects of DMB size and group size are evaluated using w DMB and w ConvBlock models, respectively.

DMB Size (K)					Group Size (M)				
K	Precision	Recall	AUC	Accuracy	M	Precision	Recall	F1	Accuracy
0	0.608	0.581	0.788	0.609	256	0.543	0.543	0.525	0.586
3	0.686	0.612	0.888	0.667	512	0.608	0.581	0.582	0.609
10	0.554	0.560	0.772	0.609	1024	0.566	0.562	0.550	0.598
20	0.616	0.566	0.740	0.598	2048	0.540	0.539	0.516	0.586

DMB, we construct the w DMB § and w Mixup $^\#$ models. As shown in Table 4 (b), removing the feature *filtering*, *quantization*, and *aggregation* processes results in a 5.7% accuracy decline due to redundant and insufficient discriminative features. While using random ratios to fuse stored features enhances current group representation, it disrupts discriminative features, causing w Mixup $^\#$ to lag behind w DMB by 2.5% and 3.5% in Precision and Recall.

Effect of Dynamic Memory Bank and Group Size. We conduct experiments to analyze the effect of DMB size K and group size M . The numerical comparisons are displayed in Table 4 (c). A suitable DMB size K can store the most informative features, thereby facilitating the current group representation to alleviate salience features memory decay, yielding a significant 9.2% gain in AUC. However, a large K can lead to feature redundancy and interference, which impairs the identification of tumor WSI, with declining 4.6% and 6.9% in recall and accuracy. While BiMamba excels at modeling long-range dependencies, excessively long sequences can lead to insufficient exploration of instance dependencies, resulting in a 6.6% decrease in F1.

4.6. Analysis on Other Feature Extractors

To further validate robustness, we focus on experimental setups using the TCGA Lung dataset with two other commonly used encoders [3, 25, 30]: the ResNet50 pre-trained on ImageNet and the SimCLR [27] ResNet18. The quantitative results presented in Table 5 demonstrate that our M3amba achieves the highest scores across all met-

Table 5. Quantitative comparison of our results and MIL results on the TCGA Lung with pre-trained Resnet50 on ImageNet and SimCLR ResNet18.

Methods	SimCLR ResNet18		ResNet50-ImageNet	
	Accuracy	F1	Accuracy	F1
ABMIL [18]	0.886±0.005	0.892±0.004	0.842±0.007	0.824±0.013
DSMIL [3]	0.940±0.007	0.935±0.008	0.809±0.004	0.787±0.007
CLAM-MB [25]	0.938±0.005	0.932±0.006	0.858±0.004	0.842±0.008
CLAM-SB [25]	0.941±0.003	0.935±0.004	0.855±0.011	0.840±0.007
TransMIL [30]	0.937±0.007	0.930±0.008	0.856±0.002	0.849±0.002
DTFD-MaxMin [47]	0.915±0.007	0.907±0.006	0.883±0.006	0.874±0.008
DTFD-AFS [47]	0.945±0.005	0.939±0.006	0.875±0.009	0.865±0.011
DTFD-MaxS [47]	0.882±0.013	0.877±0.011	0.823±0.009	0.814±0.009
MHIM-ABMIL [33]	0.943±0.007	0.940±0.007	0.891±0.007	0.883±0.007
MHIM-TransMIL [33]	0.938±0.006	0.933±0.006	0.878±0.008	0.871±0.007
IBMIL-ABMIL [21]	0.941±0.005	0.936±0.005	0.871±0.004	0.856±0.006
SSMMIL [12]	0.933±0.014	0.924±0.019	0.876±0.014	0.875±0.014
MambaMIL [43]	0.933±0.009	0.927±0.009	0.873±0.016	0.872±0.017
ACMIL [49]	0.925±0.010	0.920±0.009	0.884±0.017	0.876±0.015
M3amba (Ours)	0.979±0.015	0.977±0.016	0.934±0.024	0.928±0.027

rics and datasets. Specifically, M3amba significantly outperforms the transformer-based DTFD [8] and attention-based MHIM [21] approaches, with accuracy improvements of 3.4% and 4.3%, respectively. When comparing the SimCLR-ResNet18 and ResNet50 results, other methods exhibit a notable performance decline, while our M3amba maintains an accuracy of 93.4%. Our approach excels by effectively integrating memory-driven Mamba intra-group instance fusion and inter-group representation learning for better bag prediction.

5. Conclusion

This paper presents M3amba, a novel intra- and inter-group memory-driven Mamba framework that effectively addresses the challenges of whole slide image (WSI) classification tasks. By combining group sampling, memory Mamba blocks, and a dynamic memory bank (DMB), M3amba enhances both local and global feature representation while maintaining computational efficiency. Extensive experiments on multiple benchmark datasets demonstrate that M3amba outperforms state-of-the-art methods, achieving significant accuracy improvements while reducing computational overhead. This work provides a promising approach for efficient and accurate WSI analysis, with potential applications in medical image classification and other domains requiring large-scale, high-resolution image processing.

Acknowledgements

This research was supported by the National Science Foundation of China (No. 62476069), the Natural Science Foundation of Heilongjiang Province of China for Excellent Youth Project (YQ2024F006), and the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (GML-KF-24-09).

References

- [1] Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412, 2014. [6](#)
- [2] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermesen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. [1](#)
- [3] Kevin W. Eliceiri Bin, Li. Yin Li. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, pages 14318–14328, 2021. [1](#), [3](#), [6](#), [7](#), [8](#)
- [4] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022. [5](#), [6](#)
- [5] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. [3](#)
- [6] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *CVPR*, pages 23555–23564, 2023. [3](#)
- [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pages 9650–9660, 2021. [6](#)
- [8] Kaitao Chen, Shiliang Sun, and Jing Zhao. Camil: Causal multiple instance learning for whole slide image classification. In *AAAI*, pages 1120–1128, 2024. [8](#)
- [9] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *ICCV*, pages 17300–17311, 2023. [3](#)
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. [6](#)
- [11] Ruiwen Ding, Kha-Dinh Luong, Erika Rodriguez, Ana Cristina Araujo Lemos da Silva, and William Hsu. Combining graph neural network and mamba to capture local and global tissue spatial relationships in whole slide images. *arXiv preprint arXiv:2406.04377*, 2024. [3](#)
- [12] Leo Fillioux, Joseph Boyd, Maria Vakalopoulou, Paul-Henry Cournède, and Stergios Christodoulidis. Structured state space models for multiple instance learning in digital pathology. In *MICCAI*, pages 594–604, 2023. [3](#), [6](#), [7](#), [8](#)
- [13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. [2](#), [3](#)
- [14] Noriaki Hashimoto, Hiroyuki Hanada, Hiroaki Miyoshi, Mi-haru Nagaishi, Kensaku Sato, Hidekata Hontani, Koichi Ohshima, and Ichiro Takeuchi. Multimodal gated mixture of experts using whole slide image and flow cytometry for multiple instance learning classification of lymphoma. *Journal of Pathology Informatics*, 15:100359, 2024. [3](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [6](#)
- [16] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *CVPR*, pages 2424–2433, 2016. [3](#)
- [17] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yue-dong Yang, and Hejun Wu. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In *MICCAI*, pages 561–570, 2021. [2](#)
- [18] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2127–2136. PMLR, 2018. [3](#), [6](#), [7](#), [8](#)
- [19] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for digital histopathology. In *MICCAI*, pages 521–546. 2020. [2](#), [3](#)
- [20] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. [3](#)
- [21] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *CVPR*, pages 19830–19839, 2023. [6](#), [7](#), [8](#)
- [22] Kangning Liu, Weicheng Zhu, Yiqiu Shen, Sheng Liu, Narges Razavian, Krzysztof J Geras, and Carlos Fernandez-Granda. Multiple instance learning via iterative self-paced supervised contrastive learning. In *CVPR*, pages 3355–3365, 2023. [1](#)
- [23] Pei Liu, Luping Ji, Xinyu Zhang, and Feng Ye. Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. *TMI*, 2024. [5](#), [6](#)
- [24] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. [5](#)
- [25] Ming Y. Lu, Drew FK Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. [4](#), [5](#), [6](#), [7](#), [8](#)
- [26] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. [3](#)
- [27] Faisal Mahmood, Richard J. Chen, and Nicholas J. Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *TMI*, 37(12):2572–2581, 2018. [8](#)
- [28] Hao Quan, Xingyu Li, Weixing Chen, Qun Bai, Mingchen Zou, Ruijie Yang, Tingting Zheng, Ruiqun Qi, Xinghua Gao, and Xiaoyu Cui. Global contrast-masked autoencoders are powerful pathological representation learners. *PR*, 156:110745, 2024. [1](#)
- [29] Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. Groupmamba: Parameter-efficient and accurate group visual state space

- model. *arXiv preprint arXiv:2407.13772*, 2024. 2
- [30] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, and Xiangyang Ji. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *NeurIPS*, 34:2136–2147, 2021. 2, 3, 6, 7, 8
- [31] Jiangbo Shi, Lufei Tang, Zeyu Gao, Yang Li, Chunbao Wang, Tieliang Gong, Chen Li, and Huazhu Fu. Mg-trans: Multi-scale graph transformer with information bottleneck for whole slide image classification. *TMI*, 2023. 3
- [32] Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *ICLR*, 2022. 2, 3
- [33] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *CVPR*, pages 4078–4087, 2023. 6, 7, 8
- [34] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *CVPR*, pages 12597–12606, 2019. 3
- [35] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wizerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 5, 6
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2
- [37] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *TCYB*, 50(9):3950–3962, 2019. 2
- [38] Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. Rnns are not transformers (yet): The key bottleneck on in-context retrieval. *arXiv preprint arXiv:2402.18510*, 2024. 2
- [39] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Yuzeng Chen, Qiang Zhang, and Chia-Wen Lin. Frequency-assisted mamba for remote sensing image super-resolution. *IEEE Transactions on Multimedia*, 27:1–14, 2024. 3
- [40] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Xianyu Jin, Jiang He, Liangpei Zhang, and Chia-Wen Lin. Local-global temporal difference learning for satellite video super-resolution. *TCSVT*, 34(4):2789–2802, 2024. 2
- [41] Conghao Xiong, Hao Chen, Joseph J.Y. Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In *IJCAI*, pages 1587–1595, 2023. 3
- [42] Kai Yang, Jan Ackermann, Zhenyu He, Guhao Feng, Bohang Zhang, Yunzhen Feng, Qiwei Ye, Di He, and Liwei Wang. Do efficient transformers really save computation? *arXiv preprint arXiv:2402.13934v1*, 2024. 2
- [43] Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *MICCAI*, 2024. 3, 4, 6, 7, 8
- [44] Jin-Gang Yu, Zihao Wu, Yu Ming, Shule Deng, Qihang Wu, Zhongtang Xiong, Tianyou Yu, Gui-Song Xia, Qingping Jiang, and Yuanqing Li. Bayesian collaborative learning for whole-slide image classification. *TMI*, 2023. 6
- [45] Danlong Yuan, Jiahao Liu, Bei Li, Huishuai Zhang, Jinggang Wang, Xunliang Cai, and Dongyan Zhao. Remamba: Equip mamba with effective long-sequence modeling. *arXiv preprint arXiv:2408.15496*, 2024. 2
- [46] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 8
- [47] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, pages 18780–18790, 2022. 2, 4, 6, 7, 8
- [48] Ruijie Zhang, Qiaozhe Zhang, Yingzhuang Liu, Hao Xin, Yan Liu, and Xinggang Wang. Multi-level multiple instance learning with transformer for whole slide image classification. *arXiv*, 2023. 2, 3, 4
- [49] Yunlong Zhang, Honglin Li, Yuxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2311.07125*, 2023. 6, 7, 8
- [50] Tingting Zheng, Weixing Chen, Shuqin Li, Hao Quan, Mingchen Zou, Song Zheng, Yue Zhao, Xinghua Gao, and Xiaoyu Cui. Learning how to detect: A deep reinforcement learning method for whole-slide melanoma histopathology images. *CMIG*, 108:102275, 2023. 2
- [51] Tingting Zheng, Kui Jiang, and Hongxun Yao. Dynamic policy-driven adaptive multi-instance learning for whole slide image classification. In *CVPR*, pages 8028–8037, June 2024. 1, 4, 6
- [52] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3
- [53] Zhonghang Zhu, Lequan Yu, Wei Wu, Rongshan Yu, Defu Zhang, and Liansheng Wang. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *TMI*, 2022. 6