Dynamic Policy-Driven Adaptive Multi-Instance Learning for Whole Slide Image Classification

Tingting Zheng[1]   Kui Jiang[1]   Hongxun Yao[1*]

[1] Harbin Institute of Technology

23b903051@stu.hit.edu.cn, {jiangkui, h.yao}@hit.edu.cn

https://vilab.hit.edu.cn/projects/pamil

Abstract

Multi-Instance Learning (MIL) has shown impressive performance for histopathology whole slide image (WSI) analysis using bags or pseudo-bags. It involves instance sampling, feature representation, and decision-making. However, existing MIL-based technologies at least suffer from one or more of the following problems: 1) requiring high storage and intensive pre-processing for numerous instances (sampling); 2) potential over-fitting with limited knowledge to predict bag labels (feature representation); 3) pseudo-bag counts and prior biases affect model robustness and generalizability (decision-making). Inspired by clinical diagnostics, using the past sampling instances can facilitate the final WSI analysis, but it is barely explored in prior technologies. To break free these limitations, we integrate the dynamic instance sampling and reinforcement learning into a unified framework to improve the instance selection and feature aggregation, forming a novel Dynamic Policy Instance Selection (DPIS) scheme for better and more credible decision-making. Specifically, the measurement of feature distance and reward function are employed to boost continuous instance sampling. To alleviate the over-fitting, we explore the latent global relations among instances for more robust and discriminative feature representation while establishing reward and punishment mechanisms to correct biases in pseudo-bags using contrastive learning. These strategies form the final Dynamic Policy-Driven Adaptive Multi-Instance Learning (PAMIL) method for WSI tasks. Extensive experiments reveal that our PAMIL method outperforms the state-of-the-art by 3.8% on CAMELYON16 and 4.4% on TCGA lung cancer datasets.

1. Introduction

Histopathology slides provide detailed information on the morphology and scale of the tumor, being the gold standard for disease diagnosis [5, 33]. However, complex bio-

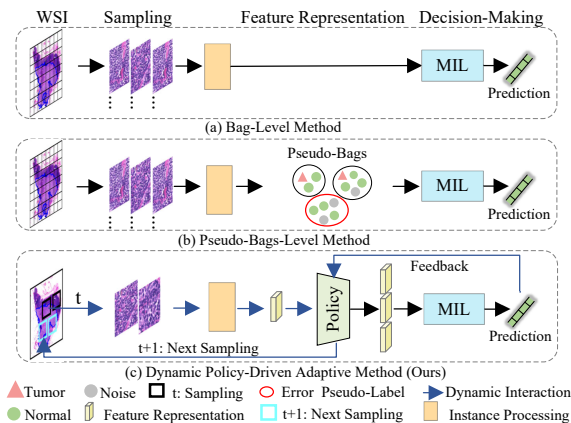---

*Corresponding authors

Figure 1. Different schemes for Multi-Instance Learning (MIL). (a) Each bag is treated as a group. (b) Each bag is divided into pseudo-bags. (c) **The proposed Dynamic Policy-Driven Adaptive Multi-Instance Learning (PAMIL) Method**.

logical structures in slides present a challenge for pathologists to discern intricate variances [6, 22]. Recently, with the development of bioimaging and computation display technologies, digital whole slide images (WSIs) have emerged as effective diagnostic aids [13, 15]. In particular, a substantial progress has been achieved by introducing deep learning approaches for WSI analysis [7, 9, 11, 16, 40]. However, due to irregularities of tumor tissue and gigapixel resolution for WSIs, the collection of high-precision is laborious and time-consuming [5, 33]. Weakly supervised solutions have drawn widespread attention than supervised schemes [18, 29]. In particular, multi-instance learning (MIL) [24] schemes emerge as the times require, providing alternative solutions for WSI analysis. Specifically, each WSI is described as a "bag" containing numerous instances (or "patches") sampled from the WSI [15]. A "bag" is labeled with a positive or negative tip, but its instances are label-free. The core mission of MIL-based methods is to aggregate information from all instances to provide bag-level predictions. According to grouping the instances, these methods can be broadly divided into two categories: bag-level and pseudo-bags-level methods, as illustrated in Fig-

ure 1. Bag-level-based methods pack all instances or embed their features into the model to regress the final prediction [6, 14, 16, 29] (Figure 1(a)). Besides the computational and memory burden, it is non-trivial to identify the most informative features from thousands of patches [26, 40]. In particular, since the mutual relationships among instances are barely explored, the distribution biases between minor tumor regions and the extensive normal tissues exacerbate the risk of model over-fitting [20, 38, 45].

By contrast, pseudo-bags-based methods leverage clustering or random schemes to decompose all instances or embedding features into multiple classes or groups, denoted as pseudo-bags (or sub-bags) [40, 41, 45]. Then the bag labels are assigned to the pseudo-bags as pseudo-labels, as shown in Figure 1(b) [9, 20, 26, 30, 38]. With the guidance of pseudo-labels, a substantial improvement has been achieved by pseudo-bags-based methods [9, 26, 30, 40]. But notably, the pre-processing for cluster and pseudo-label generation is tedious. Besides that, there exist at least two other shortcomings: 1) The tumor instances may be scattered over wide groups, especially for WSIs with small and dispersed lesions, where the informative features are diminished. It consequently increases the risk of false negative reactions. 2) The pseudo-bags derived from random sampling disrupt intra-class instance homogeneity, while the pseudo-bags derived from clustering may overly prioritize contextual similarities irrelevant to the label. Both strategies heighten the risk of misalignment between pseudo-bags features and bag labels, arising prediction errors [30, 38].

To alleviate these issues, some efforts employ attention mechanisms to adaptively focus on the label-relation components [9, 38, 40]. However, capturing the most informative features from thousands of instances with a simple Sigmoid layer is a non-trivial issue. In addition, the feature relationships among continuous sampling groups are barely explored during the diagnostic process, and thus these methods are short of stability and robustness. This inspires us to ask the following question: *Whether more elaborate and simplified sampling schemes and information aggregation mechanisms can be developed to capture the most discriminative features for accurate prediction?*

By reviewing the procedures of existing MIL methods [6, 22, 45], WSI analysis can be roughly decomposed into three sub-tasks: instance sampling, feature representation, and decision-making (as shown in Figure 1). Although some efforts have highlighted that the current decision can benefit from past knowledge and continuous learning [4, 13, 19, 25, 36, 42], the latent priors among instance sampling, feature representation, and decision-making are barely exploited. More specifically, the extracted features of current instances can encourage both sampling and feature representation of the next sampling stage to distill more informative components oriented toward the given label.

However, introducing past information for guidance naturally raises two issues: *1) How to sample the most informative instances from the rest of the sampling bag; 2) How to aggregate historical features for robust decision-making?*

To tackle the aforementioned issues, we improve the existing MIL by exploring the intrinsic relations of instance sampling, feature representation, and decision-making, and advise a dynamic policy-driven adaptive multi-instance learning (PAMIL) framework for WSI classification (Figure 1(c)). To tackle the first question, unlike the continuous random sampling [41], we introduce dynamic instance sampling and the experience-based learning of reinforcement learning (RL) and construct the dynamic policy-driven instance selection (DPIS) scheme. Specifically, DPIS takes the local neighbor relation and distance similarity between the current instance representation and the remaining instances into consideration for guiding the instance selection. Additionally, a reward function aligned with the target and a self-guided punishment mechanism are introduced to encourage the model to focus on label-related and robust instances to obtain the most representative bag predictions.

Besides the sampling optimization, how to aggregate the knowledge of continuous sampling stages to produce more distinguishable features is crucial for generating robust and accurate predictions. Since there exists uncertainty to drive the representative token from input instances, we thus propose a selection fusion feature representation (SFFR) strategy. Specifically, to achieve general and robust tokens, we introduce contrastive loss among the adjacent stage pseudo-bags to guide the generation of "positive" features (related to the given label) in the current stage. Meanwhile, to avoid forgetting historical information, we fully exploit the capability of the class token (CLS) in Transformer, where the previous tokens serve as the initial representation of the current stage to produce a more precise bag representation. Experimental results demonstrate that our proposed PAMIL framework significantly outperforms current state-of-the-art (SOTA) methods.

The main contributions of this paper are as follows:

- We examine the sampling, representation, and decision-making processes in MIL tasks, and investigate their underlying connections to establish a dynamic policy-driven adaptive multi-instance learning framework (PAMIL) for precise bag-label inference.
- We pioneer a dynamic policy-driven instance selection (DPIS) method for sample selection. This is achieved by considering the local neighbor relationship and distance similarity between the current instance representation and the remaining instances.
- We advise a selection fusion feature representation (SFFR) method for more precise bag representation by fully exploiting the historical information of sub-bags.
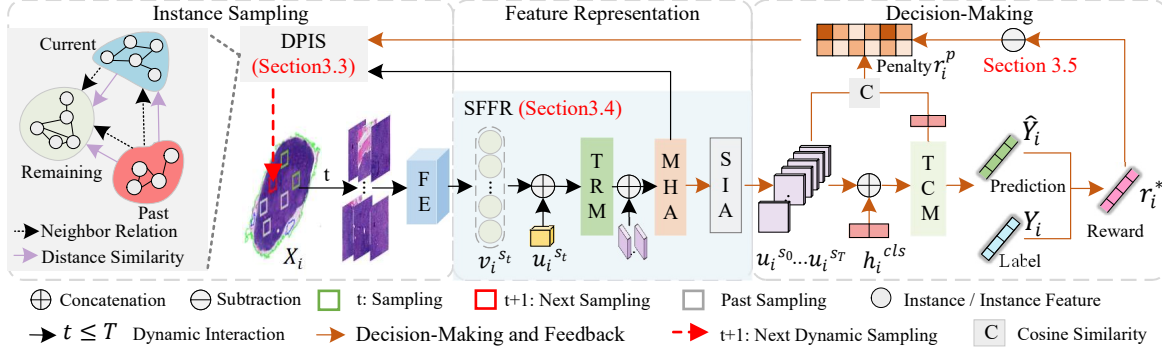
Figure 2. The architecture of our proposed dynamic policy-driven adaptive multi-instance learning (PAMIL) framework. PAMIL consists of a dynamic policy-driven instance selection (DPIS) scheme, a selection fusion feature representation (SFFR) module, and a Transformer classification module (TCM). DPIS learns relationships among past, current, and remaining instances of X_i, and samples instances embedded as $v_i^{s_t}$ using a feature extractor (FE). Then, SFFR takes $v_i^{s_t}$ and an initial token $u_i^{s_t}$ as inputs, refining $u_i^{s_t}$ by utilizing a Transformer module (TRM) and a multi-head attention (MHA) mechanism to fuse $v_i^{s_t}$ and past tokens. Meanwhile, we introduce a Siamese (SIA) structure between $u_i^{s_t}$ and $u_i^{s_{t-1}}$ to enhance the robustness of $u_i^{s_t}$. Finally, TCM uses a class token (CLS) h_i^{cls} to aggregate $\{u_i^{s_t}\}_{t=1}^T$ for inferring the bag label probability $\hat{Y}_i$ of WSI X_i, providing feedback as reward r_i^* and penalty r_i^p to the DPIS.

2. Related Work

2.1. Multi-Instance Learning in WSI Classification

MIL aims to determine bag labels from sampled instance representations. As shown in Figure 1, there are two primary MIL categories based on instance grouping: bag-level and pseudo-bags-level methods. The former applies diverse selection strategies, ranging from the conventional Max-Pooling, MeanPooling, and Top-K [6, 15, 22] to popular attention and Transformer [14, 16, 29, 34], to integrate most salient features for prediction. However, directly predicting the label from thousands of instances derived from gigapixel resolution WSIs often struggles to produce robust and satisfactory results. The latter tends to exploit correlations within instances or embedding features, employing clustering or random grouping to form pseudo-bags or labels to improve inference [26, 38, 40, 41, 45]. Nevertheless, grouping based on subtle correlations may disrupt intragroup feature consistency, leading to unstable predictions, especially for class-imbalanced datasets. To address this issue, some works use a teacher-student model to focus on hard instances [32] or refine pseudo-bags features from classifier predictions [9, 17, 37]. Besides, depending on precise regression results, these methods still fall short in fundamentally ensuring that the sampled and aggregated representations accurately and consistently align with bag labels.

2.2. Reinforcement Learning in Medical Imaging

Reinforcement learning (RL) [31] has drawn increasing attention in computer vision tasks for handling long sequential data via dynamic interactions [2, 39]. In particular, self-driven learning based on past experiences to correct biases for more accurate predictions is in line with clinical diagnosis [15]. Some interesting practices have been ap-

plied in medical image landmark detection [3, 4], segmentation [19, 36], and classification [13, 25, 44]. To optimize the WSI tumor detection, little attention has been paid to explore historical information in MIL methods [25, 43, 44]. Due to the absence of instance labels, the main challenge is to define rewards strategy. Moreover, subtle variations in designing state and action spaces can drastically affect model performance, with some not converging [13]. To tackle these challenges, we introduce a model-independent reward mechanism. In particular, an effective action-to-state mapping is achieved by employing the spatial correlations and feature similarity relations among instances.

3. Proposed Method

3.1. Review for MIL Formulation

For a binary classification task, given the dataset $X = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$, it comprises N WSIs X_i and the corresponding labels $Y_i \in \{0, 1\}$. Taking X_i as an example, it is treated as the ith "bag" involving B_i instances $x_{i,b} \in \mathbb{R}^{W \times H \times 3}$ sampled from it, where H and W denote the height and width and $b \in [1, B_i]$. All instances $x_{i,b}$ are label-free, but part of the "bag" label of $Y_i \in \{0, 1\}$.

$$Y_i = \begin{cases} 0, & \text{if } \sum_{b=1}^{B_i} y_{i,b} = 0, \quad y_{i,b} \in \{0, 1\}, \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

It can be interpreted as only if the labels of all instances are 0 (negative), the final prediction is 0 (negative). Otherwise, it would be labeled as 1 (positive).

A common MIL framework consists of two trainable modules: feature extractor (FE) G_f and classifier G_c. In this work, we use the pre-trained G_f (detailed in Section 4.2) to embed $x_{i,b}$ into a D-dimensional feature vector $v_{i,b} \in \mathbb{R}^{1 \times D}$. The G_c is used to predict the class probability

$\hat{Y}_i$ of WSI X_i, defined as $P(\hat{Y}_i) = G_c(\{v_{i,b}\}_{b=1}^{B_i})$. Due to the variability of B_i, various issues arise in MIL approaches that use the same program for all instances $\{x_{i,b}\}_{b=1}^{B_i}$.

3.2. Overview for PAMIL

Figure 2 outlines the framework of the proposed PAMIL. Our primary goal is to construct adaptive and reliable instance sampling and fusion schemes for WSI analysis by leveraging intrinsic relationships between past and current knowledge.

Drawing inspiration from [42], we introduce a Markov Decision Process (MDP) [13] for PAMIL and enhance it using the RL algorithm. Key components of the MDP comprise: $\{State, Action, Reward\}$. In this study, we define $\{v_{i,b}\}_{b=1}^{B_i}$ as state space $\mathcal{I}_i^{B_t}$ and select randomly $v_i^{s_t} = \{v_{i,j}^{s_t}\}_{j=1}^{M}$ as the initial state at $t = 1, t \in T$, where M denotes the size of $v_i^{s_t}$. The $T = \frac{B_i}{M}$ is sampling step. To efficiently link MDP with the instance sampling, feature representation, and decision-making of MIL, unlike current methods [43] to directly aggregate information from thousands of instances, we devise a dynamic policy-driven instance selection (DPIS) scheme to explore the intrinsic relation among the current instances, previous knowledge and remaining state space $\mathcal{I}_i^{B_{t+1}}$. It encourages the network to select the most informative instances to facilitate the decision-making, detailed in Section 3.3. Besides the instance sampling, we introduce a selection fusion feature representation (SFFR) module to aggregate information of the current features $v_i^{s_t}$ and past tokens $\{u_i^{s_k}\}_{k=1}^{t-1}$ for a stable and specific token $u_i^{s_t}$ representation. $u_i^{s_t}$ is passed to the DPIS and decision module for optimal sampling and robust prediction. Meanwhile, to enable DPIS to sample label-relevant and robust instances, instead of rewards from sparse WSI label support or teacher models [43, 44], we introduce a self-guided feedback mechanism from the Transformer classification module (TCM), detailed in Section 3.5.

3.3. Dynamic Policy Instance Selection Scheme

DPIS aims to select the informative samples from the remaining instances with the guidance of previous knowledge. Therefore, we first introduce an experience-based learning of proximal policy module (PPM) G_p. As shown in Figure 3, $u_i^{s_t}$ is input into a recurrent neural network (RNN) G_p^{RNN} [28] with captured temporal dependencies, followed by a multi-layer perceptron (MLP) G_p^{MLP} to guide the next sampling. This process is expressed as

$$P_i^t(a_i^{s_t}|u_i^{s_t}) = G_p^{MLP}(G_p^{RNN}(u_i^{s_t})), \quad (2)$$

Where $a_i^{s_t}$ is next sampling instances indexes. The primary goal is to improve the sampling process. By considering the connectivity and proximity between the current and the remaining instances, three different schemes are introduced to
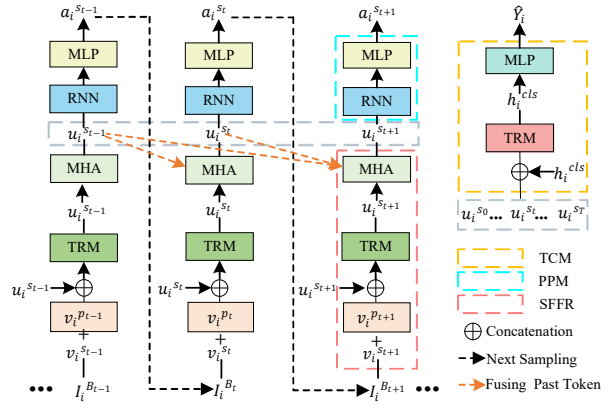

Figure 3. Illustration of the proposed **selection fusion feature representation (SFFR) module**, **proximal policy module (PPM)** and **Transformer classification module (TCM)** in PAMIL.

optimize instance selection, including greedy policy-based max similarity scheme (GMSS), greedy policy-based hybrid similarity scheme (GHSS), and policy-optimized linear interpolation instances scheme (LIIS). **More details for sampling strategies are included in the Supplementary.**

3.4. Selection Fusion Feature Representation

To aggregate robust and decision-friendly representations, we design the SFFR G_s module to comprehensively explore relationships across current and historical information. As shown in Figure 3, we first initialize a vector $v_i^{p_t}$ to learn spatial information in $v_i^{s_t}$. The Transformer module (TRM) G_s^{TRM} with an initial token $u_i^{s_t}$ is used to extract both local and global representations from $v_i^{s_t}$. To alleviate the catastrophic forgetting of historical information while thoroughly exploring the global texture, SFFR is equipped with a multi-head attention (MHA) mechanism G_s^{MHA}. Unlike randomly initializing vectors as tokens, MHA utilizes $u_i^{s_t}$ as a token to combine past tokens $\{u_i^{s_k}\}_{k=1}^{t-1}$ for enriched current representation. In addition, to promote the generalization of $u_i^{s_t}$, SFFR employs a more stable and efficient Siamese (SIA) structure [8] among $u_i^{s_{t-1}}$ and $u_i^{s_t}$. The procedures in SFFR are expressed as

$$u_i^{s_t} = G_s^{MHA}([G_s^{TRM}([u_i^{s_t}; (v_i^{p_t} + v_i^{s_t})]); \{u_i^{s_k}\}_{k=1}^{t-1}]), \quad (3)$$

$$\mathcal{L}_{SIA}^i = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{2}\mathcal{D}(p_i^{s_t}, u_i^{s_{t-1}}) + \frac{1}{2}\mathcal{D}(p_i^{s_{t-1}}, u_i^{s_t})\right], \quad (4)$$

where $p_i^{s_t}$ and $p_i^{s_{t-1}}$ are the MLP output and $\mathcal{D}$ refers to the negative cosine similarity [8].

3.5. Decision-Making and Feedback

Decision-Making. Considering the tumor features may be diluted in long sequences, potentially leading to false negatives, we use a MLP G_s^{MLP} to infer the category $\{\hat{y}_{i,t}\}_{t=1}^{T}$

of $\{u_i^{s_t}\}_{t=1}^T$. In addition, to enhance prediction performance, we construct the Transformer classification module (TCM) G_c^{TCM}, as shown in Figure 3. The TCM further fuses $\{u_i^{s_t}\}_{t=1}^T$ via an initialized class token h_i^{cls} for robust high-level global representation and employs a MLP G_c^{MLP} to predict the label $\hat{Y}_i$. For the WSI task, the class token h_i^{cls} prediction is used to minimize false positives due to noise. The final decision-making $\hat{Y}_i'$ for X_i is depicted as

$$\hat{Y}_i' = \begin{cases} \frac{\hat{y}_{i,\max} + \text{avg}(\hat{y}_{i,1:3}) + \text{avg}(\hat{y}_{i,1:5}) + \hat{Y}_i}{4}, & \text{if } Y_i = 1, \\ \hat{Y}_i, & \text{if } Y_i = 0. \end{cases} \quad (5)$$

where $\hat{y}_{i,\max}$, $\text{avg}(\hat{y}_{i,1:3})$ and $\text{avg}(\hat{y}_{i,1:5})$ denote the top-1 max, and averages of the top 3 and 5 in $\{\hat{y}_{i,t}\}_{t=1}^T$, respectively.

Reward and Punishment. Although utilizing $\hat{Y}_i$ and $\{\hat{y}_{i,t}\}_{t=1}^T$ with label Y_i consistency as rewards is straightforward, limited WSI-label Y_i provides sparse guidance and approximate label prediction of $\hat{y}_{i,t}$ introduces bias. Moreover, while the most label-related instances are useful for decision-making in the inference stage, they are not conducive to training a well-generalized model.

Based on this analysis, we first introduce h_i^{cls} as target and $\{u_i^{s_t}\}_{t=1}^T$ as negative samples. The cosine similarity between h_i^{cls} and $\{u_i^{s_t}\}_{t=1}^T$ acts as the penalty term $r_i^p = \{r_{i,t}^p\}_{t=1}^T$. Meanwhile, we integrate prediction-labeling consistency to design the reward r_i^*. These strategies guide the DPIS to select instances for prediction accuracy and stability. The feedback R_i is described by

$$R_i = \begin{cases} r_i^* - r_i^p, & \text{if } \hat{Y}_i = Y_i, \\ 0 - r_i^p, & \text{otherwise.} \end{cases} \quad (6)$$

3.6. Optimization Method

Instance Sampling Scheme Optimization. It is known that the proximal policy optimization (PPO) [27] algorithm is efficient in high-dimensional spaces and widely used in RL tasks [13, 45]. We employ PPO to optimize sampling strategies combined with Eq. 6, focusing on relevant instances and ensuring robust bag representations.

Feature Representation and Decision-Making Optimization. To ensure the co-evolution of feature representation and decision-making, we design three losses to optimize both the SFFR and TCM (SFTC): the WSI probability $\hat{Y}_i$ cross-entropy loss (WSL) $\mathcal{L}_{\text{WSL}}^i$, the sub-bag token probability $\hat{y}_{i,t}$ cross-entropy loss (STL) $\mathcal{L}_{\text{STL}}^i$, and the $u_i^{s_{t-1}}$ to $u_i^{s_t}$ SIA contrastive loss [8] $\mathcal{L}_{\text{SIA}}^i$ (Eq. 4). The losses are expressed as

$$\mathcal{L}_{\text{WSL}}^i = -[Y_i \log \hat{Y}_i + (1 - Y_i) \log(1 - \hat{Y}_i)], \quad (7)$$

$$\mathcal{L}_{\text{STL}}^i = -\frac{1}{T} \sum_{t=1}^T [Y_i \log \hat{y}_{i,t} + (1 - Y_i) \log(1 - \hat{y}_{i,t})]. \quad (8)$$

The total loss function is given by

$$\mathcal{L}_{\text{SFTC}} = \frac{1}{N} \sum_{i=1}^N [\mathcal{L}_{\text{WSL}}^i + \lambda_{\text{STL}} \cdot \mathcal{L}_{\text{STL}}^i + \lambda_{\text{SIA}} \cdot \mathcal{L}_{\text{SIA}}^i], \quad (9)$$

where λ_{STL} and λ_{SIA} are adjusted from initial to final value via integrating epoch and cosine curve [21]. For more details on the optimization methods, please refer to the **Supplementary**.

4. Experiments

To validate our PAMIL, we conduct extensive experiments on the CAMELYON16 [5] and TCGA Lung Cancer [33] datasets with mainstream MIL methods, involving bag-level methods (MeanPooling [15], MaxPooling [15], ABMIL [14], RNNMIL [6], DSMIL [16], CLAM [22], TransMIL [29], IAT [35], and MHIM [32]) and pseudo-bags or pseudo-labels methods (DTFD [40] and DG-MIL [26]). Subsequently, we evaluate common grouping methods within our SFFR and TCM. Meanwhile, ablation studies are conducted to verify individual effects of basic components on the final performance.

4.1. Datasets and Metrics

CAMELYON16 Dataset [5]. The dataset consists of 270 training WSIs (159 normal, 111 tumor) and 129 testing WSIs for breast cancer lymph nodes. Following TransMIL [29], we apply CLAM [22] to identify tissues on WSIs and obtain non-overlapping 256×256 instances at $20\times$ magnification.

TCGA Lung Cancer Dataset [33]. This database contains two cancer sub-categories: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). It provides 541 LUAD slides from 478 cases and 512 LUSC slides from 478 distinct cases. We adopt the same pre-processing as DSMIL [16] for 1046 WSIs. Each WSI is then segmented into non-overlapping patches of 224×224 at $20\times$ magnification.

Evaluation Metrics. The commonly used metrics, like Area Under Curve (AUC), accuracy, and F1 score (F1) are introduced for evaluation. Precision and recall are also considered, with a threshold of 0.5. We adopt the same settings as [29, 32, 40, 41] for a fair comparison due to discrepancies in dataset splits. The CAMELYON16 official training set is further randomly divided into training and validation sets at $9:1$. The TCGA Lung is randomly split into training, validation, and testing sets with ratios of $65:10:25$. Detailed data processing is in the **Supplementary**.

4.2. Implementation Details

For the CAMELYON16 dataset, following [22, 29, 32, 40], we extract 1024-dimensional feature vectors from each instance via a pre-trained ResNet50 [12] with ImageNet [10]. For the TCGA Lung dataset, we employ the

Table 1. **Quantitative comparison of our results and pseudo-bags-level methods on CAMELYON16 and TCGA Lung datasets.** Results from [6, 14–16, 22, 29] are derived from DTFD [40], with others [26, 32, 35, 40] from their published papers. The numbers in **red**, **blue** and underline indicate the best, second best, and previous best performance, respectively.

Methods	CAMELYON16			TCGA Lung		
	Accuracy	F1	AUC	Accuracy	F1	AUC
MeanPooling [15]	0.626±0.008	0.355±0.007	0.528±0.008	0.833±0.011	0.809±0.012	0.901±0.012
MaxPooling [15]	0.826±0.023	0.754±0.048	0.854±0.030	0.846±0.029	0.833±0.027	0.901±0.033
ABMIL [14]	0.845±0.005	0.780±0.009	0.854±0.005	0.869±0.032	0.866±0.021	0.941±0.028
RNNMIL [6]	0.844±0.021	0.798±0.006	0.875±0.002	0.845±0.024	0.831±0.023	0.894±0.025
DSMIL [16]	0.856±0.010	0.815±0.014	0.899±0.007	0.888±0.013	0.876±0.011	0.939±0.019
CLAM-MB [22]	0.823±0.022	0.774±0.017	0.878±0.013	0.878±0.043	0.874±0.028	0.949±0.019
CLAM-SB [22]	0.837±0.023	0.775±0.016	0.871±0.012	0.875±0.041	0.864±0.043	0.944±0.023
TransMIL [29]	0.858±0.008	0.797±0.017	0.906±0.025	0.883±0.022	0.876±0.021	0.949±0.013
DTFD (AFS) [40]	0.908±0.013	0.882±0.017	0.946±0.004	0.891±0.033	0.883±0.025	0.951±0.022
DTFD (MaxMinS) [40]	0.899±0.010	0.865±0.014	0.941±0.003	0.894±0.033	0.891±0.027	0.961±0.021
DGMIL [26]	0.802± —	—	0.837± —	0.920± —	—	0.970± —
IAT [35]	0.899±0.001	0.874±0.009	0.946±0.006	0.921±0.008	0.849±0.015	0.849±0.016
MHIM (TransMIL) [32]	0.920±0.008	0.901±0.010	0.965±0.004	0.900±0.025	0.897±0.026	0.949±0.021
MHIM (DSMIL) [32]	0.925±0.003	0.908±0.007	0.965±0.006	0.898±0.033	0.897±0.029	0.955±0.017
DPIS-GMSS	**0.963±0.033**	**0.948±0.048**	**0.971±0.044**	0.953±0.002	0.949±0.003	0.988±0.003
DPIS-GHSS	0.942±0.012	0.922±0.017	0.970±0.017	**0.962±0.003**	**0.958±0.004**	**0.990±0.003**
DPIS-LIIS	**0.961±0.006**	**0.947±0.009**	**0.977±0.026**	**0.965±0.007**	**0.961±0.008**	**0.994±0.002**

Table 2. **Quantitative comparison of our results and pseudo-bags schemes on CAMELYON16 and TCGA Lung datasets.** [§] indicates 512 instances per group, selected either randomly or by positional sorting. [♯] denotes k-Means or random grouping of each bag into 10 groups. The results come from the class token h_i^{cls} prediction.

Methods		CAMELYON16					TCGA Lung				
		Accuracy	Precision	Recall	F1	AUC	Accuracy	Precision	Recall	F1	AUC
Position Sorting [§]	512	0.861	0.878	0.735	0.800	0.942	0.947	0.934	0.950	0.942	**0.991**
Random Sampling [§]	512	0.861	0.878	0.735	0.800	0.897	0.947	0.920	0.966	0.943	0.990
K-Means Grouping [♯]	10	0.899	0.950	0.776	0.854	0.940	0.947	0.927	0.958	0.942	0.990
Random Grouping [♯]	10	0.920	0.924	**0.861**	0.891	0.926	0.951	0.914	**0.983**	0.947	0.989
DPIS-GMSS		0.915	**0.975**	0.800	0.876	0.905	**0.954**	0.921	**0.983**	**0.951**	0.987
DPIS-GHSS		**0.923**	0.933	0.857	**0.894**	**0.944**	**0.958**	**0.950**	0.958	**0.954**	0.989
DPIS-LIIS		**0.954**	**0.978**	**0.898**	**0.936**	**0.944**	0.954	**0.965**	0.933	0.949	**0.992**

SimCLR [23] with a ResNet18 [12] encoder, to obtain 512-dimension feature vectors from each patch. AdaMax optimizer [1] with a weight decay of $1e-5$ and the initial learning rate of $1e-4$ are used. With the above settings, we train our PAMIL with 300 epochs with batch size 1 on one NVIDIA 2080Ti GPU.

4.3. Comparison with State-of-the-Art

Quantitative results are presented in Table 1. As expected, our method achieves the best scores across all metrics and datasets, surpassing the SOTA MHIM [32] by 3.8% in terms of accuracy on the CAMELYON16 dataset and over IAT [35] by 4.4% with accuracy on the TCGA lung over. It is noted that the competitors obtain impressive performance on the TCGA lung dataset with high consistency, as the TCGA lung dataset consists of more than 80% tumor areas [29]. Since positive WSIs on the CAMELYON16 dataset occupy only small portions of tumor [16, 29], it is observed that the bag-level methods struggle to capture positive features to drive a credible decision [6, 14, 15, 22].

Exploring the global response with attention-based and Transformer-based methods [16, 29, 35] achieve better performance by producing discriminant representations across numerous samples. However, it is still behind the pseudo-bags [40]. In particular, MHIM [32] improves bag generalization features by masking key instances with well accuracy, yet it falls 3.8% behind our method. The main reason is that these methods barely explore the intrinsic relations between the historical knowledge and current instances during sampling and feature representation. It greatly facilitates robust and generalized performance. In addition, DG-MIL [26] employs K-Means grouping introducing a prior bias, resulting in poor performance. These results further demonstrate the effectiveness of our policy-driven instance selection and fusion feature representation schemes.

4.4. Comparison with Pseudo-Bags Schemes

Quantitative Comparison. To investigate the effects of sampling strategies, Table 2 tabulates the quantitative comparisons against different pseudo-bags schemes, includ-
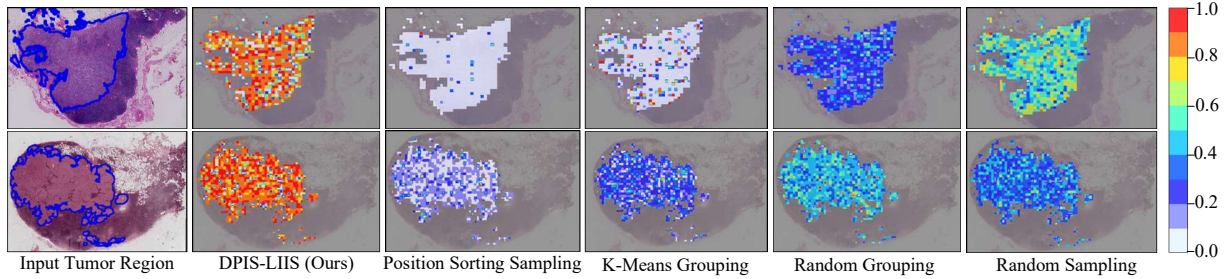
Figure 4. **Visual comparisons with pseudo-bags-level methods on CAMELYON16 dataset.** The blue outline indicates the tumor region. Attention score bar (0-1): indicates model focus on tumor instances, higher values show more attention, not positive probability.

ing the commonly used position sorting, random sampling and group, as well as our proposed greedy policy-based max similarity scheme (GMSS), greedy policy-based hybrid similarity scheme (GHSS) and policy-optimized linear interpolation instances scheme (LIIS). Notably, most of the strategies demonstrate impressive performance on the TCGA lung dataset. Our method still performs favorably on the challenging CAMELYON16 dataset. Specifically, our LIIS scheme improves accuracy by 9.3% and 3.4% against random sampling and random grouping, respectively. Although position sorting and K-Means schemes consider the spatial and feature relationships, they struggle to exploit contextual prior, leading to undesired accuracy, respectively lower than LIIS and GHSS by 9.3%, 5.5%, 6.2%, and 2.4%. We speculate that these visible improvements in scores benefit from the elaborate sampling schemes, which encourage the network to aggregate the knowledge from historical instance representation according to the similarity. It is crucial to guide the instance sampling and robust decision-making.

Qualitative Comparison and Interpretability. Figure 4 visualizes the attention scores from the last attention layer in TCM, demonstrating that our scheme can encourage the model to focus on tumor regions. It is barely explored in pseudo-bags methods. Furthermore, when grouping based on prior context and position, biases in the data lead the model to focus on instances that are not label-related. This evidence not only supports previous analyses but also suggests that our scheme facilitates more accurate learning of bag representations that are closely associated with labels.

4.5. Adaptability for MIL Method

Due to the considerable versatility and scalability of our proposed DPIS and SFFR (unified into the DPSF module), we equip four representative MIL frameworks with the DPSF module and evaluate them on the CAMELYON16 dataset. The quantitative results, tabulated in Table 3, demonstrate significant improvements over existing results. These findings validate the flexibility of DPIS and confirm that SFFR is proficient in generating discriminative representations to boost prediction. Intu-

Table 3. **Quantitative results of DPIS and SFFR (DPSF) scheme for four MIL methods on CAMELYON16 dataset.** The **bold** indicates an improvement over previous results. The results are based on the class token h_i^{cls} prediction.

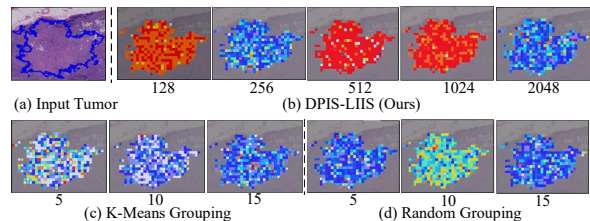| Methods	Accuracy	F1	AUC
MaxPooling [15]	0.826	0.754	0.854
DPSF-MaxPooling	0.930 (**+10.4%**)	0.901(**+14.7%**)	0.938(**+8.4%**)
MeanPooling [15]	0.626	0.355	0.528
DPSF-MeanPooling	0.876(**+25.0%**)	0.830(**+47.5%**)	0.916(**+38.0%**)
DSMIL [16]	0.856	0.815	0.899
DPSF-DSMIL	0.907(**+5.1%**)	0.882(**+6.7%**)	0.935(**+4.6%**)
TransMIL [29]	0.858	0.797	0.906
DPSF-TransMIL	0.884 (**+2.6%**)	0.845 (**+4.8%**)	0.926 (**+2.0 %**)

Figure 5. **Visual effects of pseudo-bag sizes and grouping schemes on CAMELYON16 dataset.**
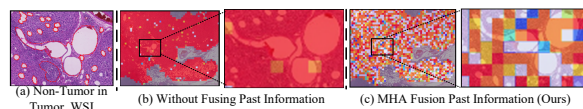

Figure 6. **Visual effects of selection fusion feature representation on CAMELYON16 dataset.** (a) Red contours indicate non-tumor cells in the tumor WSI. (b) and (c) Black boxes indicate the attention of different fusion mechanisms to non-tumor instances.

itively, DPSF-MaxPooling selects the most salient tokens from $\{u_i^{s_t}\}_{t=1}^T$ for prediction and obtains optimal results compared to DPSF-DSMIL and DPSF-TransMIL. We attribute this to SFFR providing the token with highly label-consistent features. DPSF-MeanPooling shows a 38% improvement in AUC over standard MeanPooling. However, due to the uncertain representation from input instances, DPSF-MeanPooling generates bag representations by averaging $\{u_i^{s_t}\}_{t=1}^T$ inferior attention fusion.

Table 4. **Ablation study on the CAMELYON16 dataset**. The baseline uses randomly selected 512 instances to form a pseudo-bag for training SFFR and TCM under the loss defined in Eq 7.

(a) **Effects of pseudo-bag size and grouping scheme**. "Mix" and "Max" indicate pseudo-bag counts from the smallest and largest WSIs. M denotes the instance count sampled by DPIS at time t, while "KM" and "RG" represent K-Means or random grouping of each bag into 5, 10, or 15 groups. We set $M = 512$ in our DPIS.

Methods		Accuracy	Precision	Recall	F1	Mix	Max
M	128	0.923	0.954	0.837	0.891	11	437
	256	0.923	**0.976**	0.816	0.889	6	219
	512	**0.954**	**0.978**	**0.898**	**0.936**	3	110
	1024	**0.930**	0.935	**0.878**	**0.905**	2	55
	2048	0.907	0.894	0.857	0.875	1	28
KM	5	0.907	0.974	0.776	0.864	5	5
	10	0.899	0.950	0.776	0.854	10	10
	15	0.915	0.913	0.857	0.884	15	15
RG	5	0.884	0.925	0.755	0.831	5	5
	10	0.920	0.924	0.861	0.891	10	10
	15	0.907	0.894	0.857	0.875	15	15
Baseline		0.861	0.878	0.735	0.800	3	110

(b) **Effects of selection fusion feature representation.**

Attention	MHA	Accuracy	Precision	Recall	F1
✗	✗	0.938	0.956	0.878	0.915
✓	✗	**0.946**	**0.977**	**0.878**	**0.925**
✗	✓	**0.954**	**0.978**	**0.898**	**0.936**

(c) **Effects of optimizing strategy.**

$\mathcal{L}_{\text{WSL}}$	$\mathcal{L}_{\text{SIA}}$	$\mathcal{L}_{\text{STL}}$	Accuracy	Precision	Recall	F1
✓	✗	✗	0.915	0.896	0.878	0.887
✓	✓	✗	**0.930**	**0.935**	**0.878**	**0.905**
✓	✗	✓	0.915	0.913	0.857	0.884
✓	✓	✓	**0.954**	**0.978**	**0.898**	**0.936**

(d) **Effects of reward and punishment**.

Model	r_i^P	r_i^*	Accuracy	Precision	Recall	F1
$w/o\ R_i$	✗	✗	0.930	0.917	**0.898**	0.907
$w\ r_i^*$	✗	✓	0.915	0.932	0.837	0.882
$w\ r_i^P$	✓	✗	**0.938**	**0.956**	0.878	**0.915**
$w\ R_i$	✓	✓	**0.954**	**0.978**	**0.898**	**0.936**

4.6. Ablation Study

To validate the PAMIL and analyze the contributions of individual components, comprehensive studies are conducted using the LIIS scheme, specifically sampling and results from the CAMELYON16 dataset and h_i^{cls} prediction.

Effects of Pseudo-bag Size and Grouping Scheme. Figure 5 and Table 4(a) display visual and numerical comparisons that reveal a small pseudo-bag lacks the detail needed for accurate tumor-normal feature discrimination, resulting in higher false-negative rates. However, a large pseudo-bag may obscure sparse tumor instance correlations, leading to more false positives. Comparatively, fixed grouping schemes exhibit lower accuracy and less attention to tumor instances than our DPIS strategy. This affirms the shortcomings of MIL methods stated in Section 1 and emphasizes the need for dynamic instance sampling in WSI analysis.

Effects of Selection Fusion Feature Representation. To demonstrate that incorporating features between past and current sampling instances facilitates robust and specific representation, we constructed three fusion experiments for G_s^{MHA} in the SFFR. The visual and quantitative results are shown in Figure 6 and Table 4(b), showing that incorporating historical tokens enhances model focus on label-related features. The MHA outperforms softmax attention, further demonstrating that a fully fused feature representation yields a more precise bag representation.

Effects of Optimizing Strategy. Table 4(c) demonstrates that our sampling strategy improves accuracy by 5.4% ($\mathcal{L}_{\text{WSL}}$) compared to the baseline, using the same loss Eq 7. This indicates the advantage of distance-related sampling for better bag representation. Additionally, the contrast loss in adjacent pseudo-bags further enhances features, resulting in improved inference accuracy. However, it is important to note that highly sensitive pseudo-labels can introduce biases. To address this, a hybrid loss function produces optimal results by enhancing the differentiation between tumor and normal features in imbalanced WSI datasets.

Effects of Reward and Punishment. The reward and punishment mechanisms are crucial in the DPIS scheme. Integrating penalty and reward terms, as shown in Table 4(d), significantly enhances model accuracy by promoting attention to robust and common features. Instead of using r_i^*, penalizing based on the similarity values h_i^{cls} and $u_i^{s_t}$ can effectively reduce the false-positive by correcting biases in pseudo-bags and diversifying bag representation. More in-depth analyses are included in the **Supplementary**.

5. Conclusion

In this study, we propose a new paradigm, termed the dynamic policy-driven adaptive multi-instance learning framework (PAMIL) for WSI. The main focus is on investigating the interactions between instance sampling, feature representation, and decision-making. To enhance the accuracy of inference, we utilize past predictions and sampled instances to guide the selection of optimal features in the next step. To achieve this, we introduce a DPIS scheme that effectively incorporates current and previous rich information, along with reward-penalty mechanisms to optimize sampling. Additionally, the SFFR method is employed to filter pseudo-bags and improve bag representation for more robust predictions. Extensive experiments have been conducted, demonstrating the impressive performance of our proposed PAMIL and validating the effectiveness of exploring interrelationships. However, it is worth noting that the complex structural information of WSI is not fully exploited by the distance guide strategies. A potential improvement would involve refining the representation of the state token, which could be beneficial for sampling.

References

[1] Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412, 2014. 6

[2] M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022. 3

[3] Walid Abdullah Al and Il Dong Yun. Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images. *IEEE Trans. Medical Imaging*, 39(4):1245–1255, 2020. 3

[4] Amir Alansary, Ozan Oktay, Yuanwei Li, Loïc Le Folgoc, Benjamin Hou, Ghislain Vaillant, Konstantinos Kamnitsas, Athanasios Vlontzos, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. Evaluating reinforcement learning agents for anatomical landmark detection. *Medical Image Anal.*, 53:156–164, 2019. 2, 3

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, 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, 5

[6] 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. 1, 2, 3, 5, 6

[7] Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *CVPR*, pages 15661–15670, 2023. 1

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 4, 5

[9] Yuan-Chih Chen and Chun-Shien Lu. Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In *CVPR*, pages 23936–23945, 2023. 1, 2, 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. Ieee, 2009. 5

[11] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *CVPR*, pages 3852–3861, 2020. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 6

[13] Mingzhe Hu, Jiahan Zhang, Luke Matkovic, Tian Liu, and Xiaofeng Yang. Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions. *CoRR*, abs/2206.14302, 2022. 1, 2, 3, 4, 5

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2127–2136. PMLR, 2018. 2, 3, 5, 6

[15] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier, 2020. 1, 3, 5, 6, 7

[16] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, pages 14318–14328, 2021. 1, 2, 3, 5, 6, 7

[17] Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *CVPR*, pages 7454–7463, 2023. 3

[18] Yi Li and Wei Ping. Cancer metastasis detection with neural conditional random field. In *MIDL*, 2018. 1

[19] Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *CVPR*, pages 9394–9402, 2020. 2, 3

[20] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *CVPR*. IEEE, 2023. 2

[21] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*. OpenReview.net, 2017. 5

[22] 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. 1, 2, 3, 5, 6

[23] Faisal Mahmood, Richard J. Chen, and Nicholas J. Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Trans. Medical Imaging*, 37(12):2572–2581, 2018. 6

[24] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 1

[25] Talha Qaiser and Nasir M. Rajpoot. Learning where to see: A novel attention model for automated immunohistochemical scoring. *IEEE Trans. Medical Imaging*, 38(11):2620–2631, 2019. 2, 3

[26] Linhao Qu, Xiaoyuan Luo, Shaolei Liu, Manning Wang, and Zhijian Song. Dgmil: Distribution guided multiple instance learning for whole slide image classification. In *MICCAI*, pages 24–34. Springer, 2022. 2, 3, 5, 6

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[28] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 4

[29] 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. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 1, 2, 3, 5, 6, 7

[30] Zhuchen Shao, Yifeng Wang, Yang Chen, Hao Bian, Shaohui Liu, Haoqian Wang, and Yongbing Zhang. Lnpl-mil: Learning from noisy pseudo labels for promoting multiple instance learning in whole slide image. In *CVPR*, pages 21495–21505, 2023. 2

[31] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 3

[32] 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. 3, 5, 6

[33] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 1, 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[35] Conghao Xiong, Hao Chen, Joseph Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2301.08125*, 2023. 5, 6

[36] Dong Yang, Holger Roth, Ziyue Xu, Fausto Milletari, Ling Zhang, and Daguang Xu. Searching learning strategy with reinforcement learning for 3d medical image segmentation. In *MICCAI*, pages 3–11. Springer, 2019. 2, 3

[37] Jiawei Yang, Hanbo Chen, Yu Zhao, Fan Yang, Yao Zhang, Lei He, and Jianhua Yao. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In *MICCAI*, pages 35–45. Springer, 2022. 3

[38] Rui Yang, Pei Liu, and Luping Ji. Protodiv: Prototype-guided division of consistent pseudo-bags for whole-slide image classification. *arXiv preprint arXiv:2304.06652*, 2023. 2, 3

[39] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, Jae Sung Park, Ximing Lu, Rowan Zellers, Prithviraj Ammanabrolu, Ronan Le Bras, Gunhee Kim, et al. Fusing pretrained language models with multimodal prompts through reinforcement learning. In *CVPR*, pages 10845–10856, 2023. 3

[40] 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. 1, 2, 3, 5, 6

[41] 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 preprint arXiv:2306.05029*, 2023. 2, 3, 5

[42] Zhiyue Zhang, Hongyuan Mei, and Yanxun Xu. Continuous-time decision transformer for healthcare applications. In *AISTATS*, pages 6245–6262. PMLR, 2023. 2, 4

[43] Boxuan Zhao, Jun Zhang, Deheng Ye, Jian Cao, Xiao Han, Qiang Fu, and Wei Yang. Rlogist: fast observation strategy on whole-slide images with deep reinforcement learning. In *AAAI*, volume 37, pages 3570–3578, 2023. 3, 4

[44] 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. *Comput. Medical Imaging Graph.*, 108:102275, 2023. 3, 4

[45] Zhonghang Zhu, Lequan Yu, Wei Wu, Rongshan Yu, Defu Zhang, and Liansheng Wang. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *IEEE Trans. Medical Imaging*, 2022. 2, 3, 5